Preface

The 8th International Conference on Computer Science, Engineering and Applications (CCSEA 2018) was held in Melbourne, Australia, during February 17~18, 2018. The 4th International Conference on Networks & Communications (NCOM 2018), The 4th International Conference on Artificial Intelligence and Applications (AIFU 2018), The 6th International Conference on Data Mining & Knowledge Management Process (DKMP 2018), The 7th International Conference on Embedded Systems and Applications (EMSA 2018), The 4th International Conference on Signal and Image Processing (SIPRO 2018) and The 7th International Conference on Software Engineering and Applications (SEA 2018) was collocated with The 8th International Conference on Computer Science, Engineering and Applications (CCSEA 2018). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The CCSEA-2018, NCOM-2018, AIFU-2018, DKMP-2018, EMSA-2018, SIPRO-2018, SEA-2018 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.


We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Natarajan Meghanathan
David C. Wyld
Organization

General Chair

David C. Wyld          Southeastern Louisiana University, USA
Jan Zizka               Mendel University in Brno, Czech Republic

Program Committee Members

Abe Zeid                Northeastern University, USA
Ahmad Khasawneh         Hashemite University, Jordan
Ahmed Yahya             Al-Azhar University, Egypt
Alaa Hamami             Princess Sumaya University for Technology, Jordan
Alfredo Cuzzocrea       University of Trieste, Italy
Ali AYDIN M.            Istanbul University, Turkey
Alimohammad Shahri      NESCOT, United Kingdom
Annamalai Annamalai     Prairie View A&M University, USA
Ayad Ghany Ismaeel      Al-Kitab University College, IRAQ.
Aysegul OZDEMIR TOPALOGLU Trakya University, Turkey
B.N.Gupta               Institute of Management & Science, India
Bapu Bandgar            JSPM's Jayawant Technical Campus, India
Basar Oztaysi           Istanbul Technical University, Turkey
Bedir Tekinerdogan      Wageningen University, Netherlands
Bing Zhou               Sam Houston State University, USA
Chanda C K              IIEST, India
Chin-Chih Chang         Chung Hua University,Taiwan
Chin-Wang Tao           National Ilan University,Taiwan
Christophe NICOLLE      University of Bourgogne Franche, France
Da Yan                  The University of Alabama at Birmingham, USA
Dabin Ding              University of Central Missouri, United States
Davoud Mogouei          Monash University, Australia
Debabrata Datta         St. Xavier's College, India
Debi Prasanna Acharjya  VIT University, India
Di Pietro               Roberto-HBKU-CSE, Qatar
Diego Reforgiato Recupero University of Cagliari, Italy
Donggil Song            Sam Houston State University, USA
Doo Heon Song           Yong-in SongDam College, Korea
Ekbal Rashid            Aurora's Technical and Research Institute, India
Fabio Silva             Federal University of Pernambuco, Brazil
Fadi Wedyan             Hashemite University, Jordan
Farhad pourfarzi        Ardabil University of Medical Sciences, Iran
Fatih Korkmaz           Çankiri Karatekin University, Turkiye
Figen Balo              Firat University, Turkey
Guoqing Xiao            Hunan University, China
Haeng-Kon Kim           Catholic University of Daegu, Korea
Haibat Jadhav           Flora Institute of Technology, India
HASSINI Noureddine
Hiromi Ban
Houda KHROUF
Irena Patasiene
Ishfaq Ahmad
Israa Sh. Tawfic
Ivo Pierozzi Junior
Iyad Alazzam
Jafar Mansouri
Jamaiah Yahaya
Jamal El Abbadi
Jasmine Seng K. P.
Jatindra Kumar Deka
Jeong-Hyun Park
Jia Zhu
Jinsul Kim
Lim Tien Sze
Madjid Tavana
Madya Dr. Yuhani binti Yusof
Mamoun Alazab
Mehran Garmehi
Min Kyung An
Mohd Nordin Abdul Rahman
Mokhairi Makhtar
Nasser Tairan
Nizar
Oleksii Tyshchenko
Orhan Dagdeviren
Ovidiu Valentin Drugan
Philippe Thomas
Rahil Hosseini
Ramgopal Kashyap
Rana Rahim
Razieh malekhoseini
Roberto Paiano
Samia Nefti-Meziani
Saraee Mo
Sherif Rashad
Shifei Ding
Stefania Tomasiello
Tony Tsang
TSE GUAN
Veton Kepuska
Vittoriano Muttillo
Xiaoxiong Zhong
Xin-Wei Yao
Yagoubi Belabbas
Yao-Nan Lien
Zihao Zhou

University of Mostaganem, Algeria
Nagaoka University of Technology, Japan.
Atos Innovation Lab, France
Kaunas University of Technology, Lithuania
The University of Texas at Arlington, U.S.A
University of Gaziantep, Turkey.
Embrapa Agricultural Informatics, Brazil
Yarmouk University, Jordan
Ferdowsi University of Mashhad, Iran
National University of Malaysia, Malaysia
Mohammadia V University Rabat, Morocco
Charles Sturt University, Australia
Indian Institute of Technology Guwahati, India
South China Normal University, China
Chonnam National University, Korea.
Multimedia University, Malaysia
La Salle University, USA
Universiti Utara Malaysia, Malaysia
Macquarie University, Australia
University of Bojnord, Iran
Sam Houston State University, USA
University of Sultan, Malaysia
Universiti Sultan Zainal Abidin, Malaysia.
King Khalid University, Saudi Arabia
University of Carthage, Tunisia
University of Ostrava,Ukraine
Ege University, Turkey
DNV GL, Norway
Université de Lorraine, France
Islamic Azad University, Iran
Sagar Institute of Science and Technology, India
Lebanese University, Lebanon
Islamic Azad University, Iran
University of Salento, Italy
University of Salford, UK
University of Salford-Manchester, UK
Florida Polytechnic University, USA
China University of mining and technology, China
University of Salerno, Italy
Honfg Kong College of Technology, China
Universiti Malaysia Kelantan, Malaysia
Florida Institute of Technology, USA
University of L'Aquila, Italy
Tsinghua University, China
Zhejiang University of Technology, China
University of Oran, Algeria
Asia University, Taiwan
University of California, USA
Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Networks & Communications Community (NCC)

Information Technology Management Community (ITMC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)
## TABLE OF CONTENTS

### 8th International Conference on Computer Science, Engineering and Applications (CCSEA 2018)

<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load Balancing Management Using Fuzzy Logic to Improve the Report</td>
<td>Sanghyeok Lim and Taeho Cho</td>
<td>01-13</td>
</tr>
<tr>
<td>On The Duality Feature of P-Class Problems and NP Complete Problems</td>
<td>Wenhong Tian</td>
<td>29-34</td>
</tr>
</tbody>
</table>

### 4th International Conference on Networks & Communications (NCOM 2018)

<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perception of Home-Sharing Industry : A Study of Comparing AIRBNB and Couchsurfing</td>
<td>JinTao Huang, KangNyeon Yi and JiHye Han</td>
<td>35-45</td>
</tr>
<tr>
<td>Social Live-Streaming : Twitch.TV and Uses and Gratification Theory Social Network Analysis</td>
<td>James Dux</td>
<td>47-61</td>
</tr>
<tr>
<td>Examining Risk Communication of Samsung by Analysing Customer Reaction on Social Media : From Galaxy Note 7 Explosion to Galaxy S8 Release</td>
<td>Jaewon Lim and Jang Hyun Kim</td>
<td>63-71</td>
</tr>
<tr>
<td>Improved Handover Algorithm for Proxy MIPv6 Based on AAA Server</td>
<td>Hewei Yu and Meiling Zhou</td>
<td>165-173</td>
</tr>
</tbody>
</table>

### 4th International Conference on Artificial Intelligence and Applications (AIFU 2018)

<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptualizing AI Risk</td>
<td>Martin Ciupa and Keith Abney</td>
<td>73-81</td>
</tr>
</tbody>
</table>
CNN Features are also Great at Unsupervised Classification
Joris Guérin, Olivier Gibaru, Stéphane Thiery, and Eric Nyiri

6th International Conference on Data Mining & Knowledge Management Process (DKMP 2018)

Comparison of Bankruptcy Prediction Models with Public Records and Firmographics
Lili Zhang, Jennifer Priestley and Xuelei Ni

Predicting Players' Performance in One Day International Cricket Matches Using Machine Learning
Kalpdrum Passi and Niravkumar Pandey

7th International Conference on Embedded Systems and Applications (EMSA 2018)

Implementing UHF RFID Reader on Smartphone Platform for IoT Sensing
Douglas Lautner, Xiayu Hua, Scott DeBates and Shangping Ren

4th International Conference on Signal and Image Processing (SIPRO 2018)

Image Quality Assessment - A Survey of Recent Approaches
Noor Al Madeed, Zainab Awan and Somaya Al Madeed

7th International Conference on Software Engineering and Applications (SEA 2018)

The Web-Based Education Journey: A Constant Lifeline
Vidhu Mitha

Analysis of Mobile Payment Influencing Factors
Sitalakshmi Venkatraman

GRC-MS: A Genetic Rule-Based Classifier Model for Analysis of Mass Spectra Data
Sara Al-Osimi and Ghada Badr
Data Science Methodology for Cybersecurity Projects ....................................... 205 - 218
Farhad Foroughi and Peter Luksch

Understanding People Title Properties to Improve Information Extraction Process ................................................................. 219 - 228
Saleem Abuleil and Khalid Alsamara

Strategy of The Remove and Easy TBT in GCC 6 Countries ............................. 229 - 234
Yong-Jae Kim

Performance Analysis of Symmetric Key Ciphers in Linear and Grid Based Sensor Networks .......................................................... 235 - 250
Kaushal Shah and Devesh C. Jinwala

A Proposed HSV-Based Pseudo-Coloring Scheme for Enhancing Medical Images ................................................................. 251 - 262
Noura A. Semary
LOAD BALANCING MANAGEMENT USING FUZZY LOGIC TO IMPROVE THE REPORT TRANSFER SUCCESS RATE

Sanghyeok Lim¹ and Taeho Cho²

¹College of Information and Communication Engineering, Sungkyunkwan University, Republic of Korea
²College of Software, Sungkyunkwan University, Republic of Korea

ABSTRACT

A wireless sensor network (WSN) consists of multiple sensor nodes and base stations (BS) that collect information over widely deployed sensor nodes. Sensor nodes have limited energy source and low computing power. Due to these features, there is a disadvantage that user's individual node management is difficult and they are easily captured by attackers. Therefore, efficient energy allocation of nodes is important and network security protocol is needed. The Probabilistic Voting Filtering System (PVFS) is a system that prevents false vote injection attack and false report attack injected from attackers. The reason for the existence of this protocol is for energy management of nodes through defense against those attacks and in order to efficiently manage the network based on PVFS, load balancing of nodes should be performed. In the proposed scheme, fuzzy logic is applied to each cluster head node (CH) to perform load balancing by determine whether to perform a role as a verification node and an event forwarding node. The experiment shows that the event detection rate and the report delivery success rate are improved in proposed scheme compare with original PVFS.

KEYWORDS


1. INTRODUCTION

WSNs are used for data collection and event detection in various fields such as home networks, military systems, and forest fire monitoring [1] and are composed of many sensor nodes and a base station (BS). When an event occurs, the sensor node detects the event and makes report of that event and send it over multiple hops of the sensor nodes to the BS[2]. However, sensor nodes are vulnerable to attack because of the disadvantages of limited computation, limited energy, random distribution in an open environment that operates independently, and difficulties in individual management [3], [4].Attackers exploit these vulnerabilities to attack WSNs by injecting reports containing false information or injecting false Message Authentication Codes (MACs). Figures 1 shows a schematic of these attacks. These attacks reduce the energy of the node, shorten the lifetime of the network, and prevent detection and reporting of normal events. Several protocols have been developed to prevent false report attacks[5],[6] and Li and Wu proposed a probabilistic voting-based filtering scheme (PVFS) [7]to prevent both false report and false vote injection attacks.
In PVFS, all nodes constitute a network that exploits cluster-based organization. When a cluster head (CH) recognizes an event, it generates a report of that event. It then sends this report to the member nodes. Next, the member nodes judge the authenticity of the report and generate their own MACs, alternatively referred to as votes in PVFS.

Nodes that act as verification nodes in PVFS are more energy intensive than other member nodes, and if these nodes die, report generation and delivery of the region becomes impossible. In this paper, each CH’s load balancing is performed by fuzzy logic based node role determination to increase event detection success rate and report transmission success rate with PVFS. Fuzzy logic is based on the fuzzy set thinking concept introduced by Professor L. A. Zadeh of the University of California at Berkeley in 1965 to quantitatively express the ambiguity of natural language and the like [8]. In the proposed method, fuzzy logic is used to help determine the role balancing of nodes by quantifying ambiguous properties such as how big the amount of nodes’ energy is, how important nodes are.
2. RELATED WORK

2.1. PVFS

PVFS uses a true threshold value (Tt) and a false threshold value (Tf) to detect and filter false reports and false vote injected reports with validation nodes. PVFS has three phases; a key distribution phase, a report generation phase, and a verification phase.

Figure 3. Key distribution process

Figure 3 shows the key allocation step in which the BS divides the key pool into N partitions and delivers them to each CH. Each partition contains L keys. The CH uses one of the keys in the partition as its own key and distributes the remaining L-1 keys to its member nodes. A key is allocated to the member nodes according to the partition of the key pool. With this process, every node gets a single key from a global key pool.

Figure 4. Report generation and verification node selection process

Figure 4 shows the report generation and validation node selection process. In the initial network configuration, nodes are divided into cluster units and the CHs responsible for report generation are selected in each cluster. When an event occurs, the cluster closest to the event area wins the chance to generate a report of that event. CH generates a report and broadcasts it to member
nodes of its cluster. The member nodes confirm this, and if the report is judged to be a normal report, the MAC created by its own key is transmitted to the CH. CH extracts a predetermined number of MACs received from the member nodes and adds them to the report. The node selected as the verification node stores the keys of the member nodes of the event occurrence cluster one by one. In the report verification process, the verifying nodes compare their own key indexes with that of in the report. If they have same key index, they extract the MAC through their key. If the MAC generated by the same key is different, the vote is regarded as fake and Tf is increased. In the filtering process, if the false count reaches the threshold value, the report is judged to be false and is immediately dropped. If the true count value reaches the threshold value, the report is considered legitimate and is sent to the BS without further validation. The verification node selection process is shown in Figure 4-b. The verification nodes among the CHs on the path to the BS are probabilistically selected to verify the report. The probability ‘p’ uses the distance $d_0$ which is the hopcount from the BS to the event cluster and the distance $d_i$ between the BS and CH$i$. The closer the CHs are to the event cluster, the higher the probability of verification.

### 2.2. Fuzzy logic system

Fuzzy logic is based on the fuzzy set thinking concept introduced by Professor L. A. Zadeh of the University of California at Berkeley in 1965 to quantitatively express the ambiguity of natural language and the like. [9], [10], [11]. The concept of a fuzzy set is a set out of binary logic that each object belongs to or belongs to a certain group and represents the degree to which each object belongs to the group as a membership function. Also, the fuzzy measure indicates the ambiguity of the fact that the ambiguous element $a$ in the general set $A$ belongs to the subset $P$ of $A$, so that the relation between $P$ and $A$ is a mathematically continuous attribute. The configuration of the Fuzzy Logic Controller is shown in figure 5.

![Figure 5. Fuzzy logic controller](image)

The fuzzy controller consists of a fuzzy rule base, a fuzzy inference engine, and a defuzzifier. Fuzzification Interface is responsible for Scale Mapping task that changes the value range of the input variable to the universe of discourse. Fuzzy Rule Base is a set of linguistic control rules written in IF-THEN form. Fuzzy Inference Machine is decision-Making Logic that adopts rules from Fuzzy Input and Fuzzy Rule and in defuzzication Interface, scale mapping work is performed to convert into Universe of Discourse which is matched with the range of output variable value.

### 3. PROPOSED METHOD

#### 3.1. Problem statement

In an environment where a WSN is installed, there are many areas where the user cannot distribute nodes directly and the nodes are scattered randomly. Therefore, there is a high possibility that a good routing path based on energy efficiency and characteristics of PVFS is not generated every time. One of the main purposes of the WSN is to gather information about events that occur in vast areas where it is difficult for a user to directly reach and respond appropriately. If the report of the detected event is not successfully transmitted to the BS, the main purpose of
the WSN is to lose, or if the energy of the node is depleted so that event detection in the corresponding region is not possible. Also, if the residual energy of the sensor nodes is high but the events to be delivered are not properly transmitted, the sensor nodes distributed in those areas become useless. Therefore, it is important to improve the event detection rate and report delivery success rate even if the total energy consumption of the WSN increases by modifying the node to be available for those multiple purposes. Since PVFS shows probabilistic security, there is a tendency to show extremely high and low performance in random routing environments. Therefore, it is important to use nodes efficiently in the WSN by adjusting the role of nodes appropriately and raising the event detection rate. Figure 6 shows the part of the field where the WSN is configured. As shown in the figure, CHs that detect the event have limited communication distances, so they need to cooperate with other nodes to deliver the report to the BS through hop movement.

![Figure 6. Calculating tail number](image)

Therefore, each node has different number of downstream nodes depending on its location. Tail refers to the number of CH nodes that use the node to forward detected events. If the CH is positioned at the end of the field, the tail value of the corresponding node is zero. A node with many tails can consume a large amount of energy for sending, receiving, and verifying the report, which plays an important role in transferring report of the events occurring in other regions. If such node dies, the event detection node must forward the report through the new path except that node, and the event detected in the area where the dead node was receiving is no longer detectable because the CH is selected only once in the initial node dispatch phase.

### 3.2. System overview

Figure 7 shows an example of the network on a field using the proposed method. Every CH has a list of the IDs of the nodes that are closer to the BS within its communication range. This list is used for the formation of new routing except the nodes selected by the tagging node among the list members. The sensor nodes have enough memory to store this brief information.
Nodes with depleted energy are dropped from the list and the tagged nodes remain in the list but are excluded from the routing priority. If all the nodes in the list are tagged, the node closest to the BS is forcibly routed again. The node tagged by the fuzzy logic based decision system on the original route requests the previous route node to set up new routing, and the tagged node performs only event detection and report generation during the remaining period. The fuzzy function has three inputs: the energy of the node, the geographical importance of the node, and the frequency of occurrence of the event in the region. The calculation of the fuzzy function is performed with a constant period when it falls below a certain energy because frequent fuzzy calculations shorten the lifetime of the node.
In the actual WSN field, there is very little chance that events and attacks will occur uniformly locally. In particular, in the case of an attack, it is difficult for the attacker to travel around the entire field, and attack occurs through a few nodes that have been hijacked. Therefore, load balancing considering the local attack rate is necessary. This is also why the frequency of occurrences of fuzzy input values in the proposed method is getting into. Figure 8 shows how these attacks and normal events can occur in an actual field. The proposed scheme extend the lifetime of the network through load balancing of nodes considering these characteristics, and improve report transmission success rate and event detection rate.

3.3. Fuzzy based fitness evaluation

The proposed method exploits a fuzzy rule based system for load balancing. Part of the appeal of fuzzy rule based systems is that they can be used for approximate reasoning. Which is particularly important when there is uncertainty in reasoning processes in addition to imprecision in data. The fuzzy input includes the amount of residual energy of the node, the importance of the position of the node, and finally the frequency of local event occurrence. The fuzzy output for these three inputs helps determine whether to use the CH as an event delivery and verification node.

3.4. Fuzzy membership function

The functions for the three inputs of the fuzzy logic are shown in the figure 9. The following is the reason for selecting the values used in the fuzzy input. The fuzzy functions contain the importance, energy amount, and frequency of occurrence of event and attacks respectively. Optimization of functions can be performed using genetic algorithms [12], [13], [14]. In the proposed method, several functions are randomly simulated and a function that most satisfying the if-then rule is used.

![Figure 9. Fuzzy membership functions](image-url)
1. Importance (a): Importance is a measure of whether a given CH plays an important role in the network. It is a value considering the probability that the CH becomes the verification node of the downstream nodes and the tail number of the corresponding CH. A node with a high probability of being a verification node is expected to have a high energy consumption and its importance is increased. In general, in PVFS, the closer to BS, the more tail increases and the probability of becoming a verification node becomes lower. Therefore, the distance with BS does not lower the importance of corresponding nodes every time and it’s hard to measure it just considering distance. Therefore, the proposed method uses the measurement function to measure the importance. The evaluation function of i-th node $E_{tag}^i$ is expressed as the sum of $\alpha$, $\beta$ and $\gamma$, which are the amount of energy used during the role of the forwarding node, the amount of energy consumed during the role of the event detection node and the energy used by the verification node to report transmission after event detection, respectively. $\alpha$ is given by $\alpha = e_t \times \frac{N_{evt}}{N_n}$, where $e_t$ is the transmitting cost, $N_n$ is the number of nodes in the field and $N_{evt}$ is the number of events. $\beta$ is given by $\beta = T \times (e_t + e_r) \times \frac{N_{evt}}{N_n}$, where $e_r$ denotes the receiving cost and $T$ is the i-th CH’s tail node. $\gamma$ is more complicated, and can be expressed as the sum of probabilities that the i-th node is a verification node of each tail node, all times the probability of having the same key when it is a verification node. The sum of probabilities that a node is a verification node of the tail nodes is given by $e_{cal} \times \sum_{j=1}^{^{\text{HopCount}_{N_i}}} P_j^i$, where $P_j^i = \frac{1}{\text{HopCount}_{N_j}} \times \sum_{j=1}^{\text{HopCount}_{N_j}}$. When the node performs verifying, the probability that the key of the corresponding node overlaps with the key of the report is $\frac{s}{L}$, where $s$ is the number of votes in the report and $L$ denotes the number of member nodes from each tail nodes. Therefore, the evaluation function is as follows:

$$E_{tag}^i = e_t \times \frac{N_{evt}}{N_n} + T \times (e_t + e_r) \times \left(\frac{N_{evt}}{N_n}\right) + \frac{S}{L}e_{cal} \times \frac{1}{\text{HopCount}_{N_i}} \times \sum_{j=1}^{\text{HopCount}_{N_j}}$$

The output of the function is integerized into the fuzzy input.

2. Energy (b): The energy of the node is the most important element of the fuzzy inputs. When nodes with sufficient energy are tagged, the routing path is unnecessarily increased, which has the adverse effect of increasing transmission and reception costs. Also, the probability of encountering the hijacked node during the hop movement increases. On the contrary, if the CH is tagged in a situation where the amount of energy is too small, there is not enough residual energy to perform an event detection function, so that a node is killed during the generation of a report or waiting for a report, Report passing rates are also lower.

3. Event occurrence frequency (c): The occurrence frequency of events indicates the ratio of events occurring in the area where the corresponding node is located among all the fields. If the event frequency is high, the probability of event detection and report generation increases, and the probability of being tagged increases accordingly. On the other hand, if the frequency of occurrence of events in the area is low, it is desirable from the viewpoint of the whole network to delay the tagging timing and induce the node to perform the role of the verification and transmission node for a longer time. The frequency of occurrence is difficult to measure in real time, and the amount of energy consumed in real-time measurement increases, which is rather counterproductive.
The following is if-then rule of fuzzy system. It contains those 3 inputs that previously explained. Each input values have 3, 4 and 3 levels, respectively and it’s the output of the rule is determined by the expert.

<table>
<thead>
<tr>
<th>No.</th>
<th>Importance</th>
<th>Energy</th>
<th>Event occurrence</th>
<th>OUTPUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LOW</td>
<td>VERYLOW</td>
<td>LOW</td>
<td>ON</td>
</tr>
<tr>
<td>13</td>
<td>MID</td>
<td>LOW</td>
<td>LOW</td>
<td>OFF</td>
</tr>
<tr>
<td>15</td>
<td>MID</td>
<td>LOW</td>
<td>HIGH</td>
<td>ON</td>
</tr>
<tr>
<td>18</td>
<td>HIGH</td>
<td>LOW</td>
<td>HIGH</td>
<td>ON</td>
</tr>
<tr>
<td>31</td>
<td>MID</td>
<td>HIGH</td>
<td>LOW</td>
<td>OFF</td>
</tr>
<tr>
<td>36</td>
<td>HIGH</td>
<td>HIGH</td>
<td>HIGH</td>
<td>OFF</td>
</tr>
</tbody>
</table>

Table 1 shows some of the 36 if-then rules used in the proposed fuzzy system. The 3 input values are Importance, Energy, and event occurrence, and the output contains whether or not to tag the node as a report-generating-only node. For example, if the importance is LOW, the energy is VERYLOW, and the event occurrence is LOW, the node is selected as a node for event generation only. If the importance is MID, energy is LOW and event occurrence is LOW, It is decided not to tagged yet. The input can’t be arbitrarily adjusted, and it must have a total of 36 values. The output of the rule reflects the case with the best results through experiments with several combinations. Rule with inconsistent output results in system performance degradation, so the exorbitant output was excluded from the experiment.

4. EXPERIMENTAL RESULT

4.1. Experimental environment

<table>
<thead>
<tr>
<th>Item</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensor field size(m)</td>
<td>1000×1000</td>
</tr>
<tr>
<td>Number of sensor nodes</td>
<td>2000</td>
</tr>
<tr>
<td>Number of cluster head nodes</td>
<td>200</td>
</tr>
<tr>
<td>S</td>
<td>5</td>
</tr>
<tr>
<td>L</td>
<td>10</td>
</tr>
<tr>
<td>Packet size(byte)</td>
<td>24</td>
</tr>
<tr>
<td>Transmission range(m)</td>
<td>150</td>
</tr>
</tbody>
</table>

Experiments were conducted assuming that the attack rates were 0 to 90%. The reason a 100% attack rate is excluded from experiments is that this experiment is an experiment that evaluates the failure rate of a normal report. The transmission and reception costs are set to 16.25μj and 12.5μj, respectively, and the calculated cost of voting was set to 15μj [15]. The threshold values Tt and Tf used in the experiments were in accordance with the experimental environment of the original PVFS[7]. The reason for this is to increase the reliability of the proposed method by making the environment of original version of PVFS equal to the proposed method.

4.2. Assumptions

The CH’s location and key distribution phase does not change during the event. If the initial energy of the node is too high, it will be difficult to calculate the event detection failure and report transmission failure accurately because the node’s energy depletion does not occur during the
experiment. Therefore, the initial energy of the node is set to an appropriate amount. The BS has all the keys distributed to the node and has the computing power to verify all false reports, false votes.

4.3. Experimental result

Figure 10. Number of transfer Failure

Figure 10 shows the number of report failures according to the attack rate at 600 events when the local occurrence frequency is 20, 40, 30, and 10%. The experiment was conducted with the exception of the transmission success rate for false report attacks because the network users are not mean to receive false reports derived from the attacker. It can be seen that the number of report transmission failures of the existing PVFS and the proposed technique decreases as the attack rate increases because when an attack ratio increases, the number of normal event’s report sending itself is decreases. And as the attack rate increases, it can be seen that the difference between the number of report transmission failures of the proposed technique and PVFS is also reduced.

Figure 11. Number of surviving Nodes
Figure 11 shows the number of surviving nodes according to the attack rate of the existing PVFS and the proposed scheme. The number of surviving nodes was measured at every cycle after the experiment, and there is some part with little difference from original version of PVFS (70%, 90%). The reason for this result is that the proposed scheme has more number of surviving nodes than the existing PVFS during the experiment, resulting in additional energy consumption due to event reception, report generation, and report transmission. These results show that the number of failures of the proposed method and the report delivery failure of the existing PVFS are significantly different from those of the Figure 10, but the number of surviving nodes is not so different.

Figure 12 shows the performance difference between the proposed method and the existing PVFS according to the change of local incidence a, b, c, and d of Figure 12 represent the frequency of events occurring in the regions of Figures 8-a, 8-b, 8-c, and 8-d, respectively and each variable means a = 25,25,25,25(%), b = 30,40,20,10(%), c = 20,20,50,10(%), d = 0,50,0, 50(%). As can be seen from the graph, load balancing with fuzzy logic increases the lifetime of the entire network and increases the success rate of the normal report on the event. A large number of surviving nodes means that event detection for the region can be longer and more. The residual energy of all the nodes can be measured somewhat lower in the proposed method because the report about the event that occurred in the dead node is generated and transmitted, the event detection is not performed, and report transmission more.

3. CONCLUSIONS

In the proposed method, we experimentally confirmed that a WSN's event detection success rate, report transmission success rate, and overall network lifetime are increased by organically changing the roles of the nodes through fuzzy logic based decision making system that evaluate the node status. What the user wants through the WSN is successful reporting of normal events. Therefore, the experiment of the proposed technique excluded how successful the transmission of false reports was. The WSN user periodically replaces the node battery placed in the entire field or picks up the nodes placed in the field when the local network is not needed. Therefore, the total sum of the residual energy amount of the entire nodes placed in the field does not influence the performance of the network. Although it is relatively low energy, it should be the criterion for determining the performance of the network whether or not the node can generate a report on the
live event and send it to BS. The proposed method is a method of raising the report success rate and the number of surviving nodes simultaneously. For further study, the load balancing using the membership function of the fuzzy logic which is optimized through the genetic algorithm will be implemented. The routing environment changes every moment and it is impossible for humans to optimize the membership function at every moment, so the genetic algorithm should be applied to automatically adapt the changing routing environment to obtain the optimal output.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. NRF-2015R1D1A1A01059484)

REFERENCES


AUTHORS

Sanghyeok Lim

received a B.S. degree in Digital Information Engineering from Hanguk University of Foreign Studies in 2017, and is now working toward an M.S. degree in the Department of Electrical and Computer Engineering at Sungkyunkwan University.

Taeho Cho

received a Ph.D. degree in Electrical and Computer Engineering from the University of Arizona, USA, in 1993, and B.S. and M.S. degrees in Electrical and Computer Engineering from Sungkyunkwan University, Republic of Korea, and the University of Alabama, USA, respectively. He is currently a Professor in the College of Information and Communication Engineering, Sungkyunkwan University, Korea.
PERFORMANCE EVALUATION OF MOBILE IP ON MOBILE AD HOC NETWORKS USING NS2

Kingsley K. Ofosu¹, Jamal-Deen Abdulai (PhD)² and Ferdinand A. Katsriku (PhD)³

Department of Computer Science, University of Ghana, Legon-Accra, Ghana

ABSTRACT

Mobile computing devices equipped with transceivers form Mobile Ad Hoc Networks (MANET) when two or more of these devices find themselves within transmission range. MANETs are stand-alone (no existing infrastructure needed), autonomous networks that utilise multi-hop communication to reach nodes out of transmitter range. Unlike infrastructure networks e.g. the Internet with fixed topology, MANETs are dynamic. Despite the heterogeneous nature of these two networks, integrating MANETs with the Internet extends the network coverage and adds to the application domain of MANETs. One of the many ways of combining MANETs with the Internet involves using Mobile Internet Protocol (Mobile IP) and a MANET protocol to route packets between the Internet and the MANET via Gateway agents. In this paper, we evaluate the performance of Mobile IP on MANET in Network Simulator 2 (NS2). We have implemented Mobile IP on Ad hoc On-demand Distance Vector (AODV), Ad hoc On-demand Multiple Distance Vector (AOMDV) and Destination-Sequence Distance Vector (DSDV) routing protocols and compared performances based on Throughput, End-to-End Delay (E2ED), Packet Delivery Ratio (PDR) and Normalized Packet Ratio (NPR). The simulation results suggest that on-demand routing within the MANET better serves Mobile IP on MANETs.

KEYWORDS

Mobile Ad Hoc Network, Mobile IP, AODV, AOMDV & DSDV

1. INTRODUCTION

Communication has shifted from a predominantly wired setup towards an entirely wireless setup or a merger. Computer networks aid in faster and reliable communications over long distances. The Internet, a network of networks, has become a vital utility in our lives that enable us to communicate around the globe. Mobility as a feature in communication has gained the acceptance of end-users, and so, it is not surprising that Mobile Ad hoc Networks (MANETs) have attracted much attention from researchers. A MANET is an autonomous, infrastructure-less, self-forming and self-repairing data network of mobile devices that support multi-hop communication. MANET could be used to provide Internet connectivity beyond the reach of fixed or cellular infrastructure [1].

The Mobile Internet Protocol (MIP), which was designed to allow mobile nodes to move inside the fixed Internet without losing connectivity, has been experimented in MANET to provide Internet connectivity. We have implemented three variants of Mobile IP on MANET:
Mobile IP on Ad Hoc On-demand Distance Vector (AODV)

Mobile IP on Ad Hoc On-demand Multiple Distance Vector (AOMDV)

Mobile IP on Destination-Sequenced Distance Vector (DSDV)

and evaluated their performances using Network Simulator 2 (NS2).

2. LITERATURE REVIEW

Not much work has been published concerning providing Internet connectivity for MANETs. A number of publications have suggested corporation between Mobile IP [2] and an ad hoc routing protocol to provide Internet access for MANETs. In “Ad Hoc Networking with Mobile IP” [3], a solution was presented whereby a proactive MANET routing protocol was used with Mobile IP. This solution was not compatible with on-demand routing in MANET. In “MIPMANET – Mobile IP for Mobile Ad Hoc Networks” [4], an on-demand MANET routing protocol (AODV) is used alongside Mobile IP with foreign agent care-of-addresses to connect a MANET to the Internet. Likewise MIPMANET, the Internet draft “Global Connectivity for IPv4 Mobile Ad hoc Networks” (Global4) [5] presented a solution whereby AODV interacts with Mobile IP however, the foreign agent discovery mechanism is incorporated into AODV.

Other publications have used different gateway discovery methods within the MANET routing protocol instead of Mobile IP. The publication titled “Wireless Multihop Internet Access: Gateway Discovery, Routing and Addressing” [6] discusses how MANET nodes can discover gateways to the Internet and the issue of routing and addressing in heterogenous environments. The master’s thesis “A Study of Internet Connectivity for Mobile Ad Hoc Networks in NS2” [7] implemented the Internet draft “Global Connectivity for IPv6 Mobile Ad hoc Networks” [8] also known as Global6 in NS2. The AODV routing protocol was modified to include three different gateway detection mechanisms that were tested using NS2 simulations.

Regarding papers that focused more on performance analysis, [9] and [10] conducted separate performance evaluation of Mobile IP on proactive and reactive MANET routing respectively. We will return to these papers as we discuss related works to thus thesis in section 2.3, after we have discussed Mobile IP and Mobile Ad hoc Networks in Sections 2.1 and 2.2 respectively.

2.1. Mobile IP

Mobile IP works by allowing the Mobile Node (MN) to maintain two IP addresses; the home address and the Care-of-Address (CoA) [2]. The home address is static and binds MN to a known network called the Home Network (HN). This address is used by transport and application layer protocols to maintain connectivity with MN. On the other hand, the CoA is a dynamic address that MN obtains as it moves outside HN into an unknown network termed as a Foreign Network (FN). The CoA reflects the physical location of MN outside of HN.

When MN moves from the HN into FN, it sends the CoA obtained to the Home Agent (HA) on its HN via a similar agent on FN known as Foreign Agent (FA). This message prompts HA to receive packets destined for MN and arriving at HN on behalf of MN and then tunnels the packets to the CoA. MN repeats this procedure each time MN obtains a new CoA. There are three building blocks of Mobile IP: agent discovery, registration and datagram delivery.
2.1.1. Agent Discovery

This process is an extension of the router advertisement procedure, as specified in the Internet Control Message Protocol (ICMP) router discovery messages in RFC 1256 [11]. Specifically, the only difference is that an agent advertisement has, also, one or more CoAs made available by FA.

[2] explains that an agent advertisement does the following:

- Allows the detection of mobility agents;
- Lists the available CoA;
- Informs MN of the services offered by FA e.g. alternative encapsulation techniques;
- Aids MN to determine the network address and the status of the link to the Internet;
- Helps MN to determine whether the agent is a Home Agent, Foreign Agent or both, and therefore whether it is inside HN or FN.

Agents broadcast advertisements periodically, once a second or once every few seconds. However, an MN may solicit for an agent advertisement regardless of this schedule.

2.1.2. Registration

When MN has a new CoA, it must inform HA to render its services to MN. This procedure is known as registration [1]. MN sends a registration request containing:

- Current CoA
- How long MN intends to use the CoA i.e. Time-to-Live (TTL)
- Parameters and flags that specify how HA should forward packets
- Special services that MN requests of HA.

When HA receives the registration request, it then authenticates MN and then decides whether to accept or reject depending on the outcome of the authentication process. HA sends a reply to MN via FA. If HA accepts the request, it maintains in its cache the home address, CoA and the TTL that MN specified in the request. This trio is known as a binding information, and so registration request is sometimes referred to as a binding update. Although the FA remains passive in this process, it maintains a visitors list of each MN that successfully registers with HA through it.

2.1.3. Datagram Delivery

Figure 1. Mobile IP architecture.
As illustrated in Figure 1 above, the Corresponding Node (CN) transmits packets destined for MN towards HN through standard IP routing. When the packets reach HN, the HA accepts these packets as though it is MN. HA then encapsulates the packet with the CoA as the new destination and forwards the packets to MN. This process is known as tunnelling. Two tunnelling protocols often used are IP within IP [12] and minimal IP within IP [13]. HA continues to receive packets on behalf of MN until TTL runs out. If MN wishes to continue this service, it must register again.

2.2. Mobile Ad Hoc Networks

A MANET is an autonomous, infrastructure-less, self-forming and self-repairing network of mobile devices that support multi-hop communication. The mode of transmission is wireless and has a very dynamic topology due to node movements. MANET nodes are usually small, and battery powered. MANETs are deployed in disaster recovery [14], military applications [15], sensor networks [16], etc.

2.2.1. Manet Routing Protocols

Further Initially, MANET routing protocols were proactive. They store and update periodically, information on existing paths to all possible destinations for data delivery. Proactive protocols are quicker in choosing routes for data delivery, however, they do not easily converge [17]. E.g., Destination-Sequenced Distance Vector (DSDV) [18].

Reactive or On-demand protocols, unlike the proactive protocols do not anticipate routes between nodes. Nodes discover routes only when there is data to transmit. These protocols only maintain a node neighbour list necessary for route solicitation. E.g. Ad hoc On-demand Distance Vector (AODV) [19].

**Destination-Sequenced Distance Vector (DSDV)**

DSDV is based on a distance-vector algorithm and guarantees a loop-free route unlike earlier distance vector based routing protocols by introducing sequence numbers [18]. Nodes maintain a routing table of all destinations with information on the next-hop to the destination, the number of hops and destination sequence number.

Periodically, a node shares its route table information with neighbouring nodes known as route updates. Attached to each update is a sequence number. An odd sequence number indicates an unreachable destination whereas an even sequence number indicates a reachable destination. The greater the sequence number, the fresher the update.

It is possible to have different updates to the same destination with the same sequence number. In such scenarios, nodes prefer the route with a lesser number of hops to the destination.
Ad Hoc On-Demand Distance Vector (AODV)

AODV [19] is a reactive protocol also based on the distance-vector algorithm. In AODV, nodes only seek for route information only when they have data to transmit. Loop-free routes are guaranteed with the use of sequence numbers. A node maintains route information only when transmitting data or it finds itself along an active path.

Each time a node has data to share, it solicits for a route to the destination by broadcasting a route request (RREQ) message. A node which is the destination itself or another node that happens to know a route to the destination may unicast a route reply (RREP) to the source or to the intermediate node from which it received the forwarded RREQ and maintains a pointer to that node. This process is known as a forward route setup.

Otherwise, the intermediate node forwards the RREQ and maintains a pointer to the node from which it received the RREQ. This process is known as a reverse route setup. Data transmission may begin as soon as the source receives the RREP via the forward route setup to the destination. Nodes maintain the forward route until there is no more data to transmit or the path becomes broken due to an unreachable intermediate node or destination in which case route solicitation restarts.

Ad Hoc On-Demand Multiple Distance Vector (AOMDV)

AOMDV [20] is a reactive protocol that improves on AODV. In AOMDV, nodes maintain multiple forward route setups during a single route solicitation procedure. These multiple routes are link disjoint which means that no two routes share a common intermediate node. This procedure reduces the probability that all routes fail at once. A primary route for data transmission is then selected, and nodes maintain the rest as alternate routes. When the primary route fails, nodes select an alternate route for data delivery, unlike AODV that repeats route solicitation whenever there is route failure.

Nodes repeat route solicitation when there is a new transmission session or when all alternate routes and the primary path fails.

2.3. Related Works

In this section, we discuss some closely related works to this thesis as we build up the motivation for this study.

In [3], a proposal for connecting MANET to the Internet using Mobile IP is presented in which a modified Routing Information Protocol (RIP) like DSDV was used in routing packets within the MANET. In this proposal, a single routing table is used and shared by Mobile IP and the MANET routing protocol to reduce the management tasks involved in maintaining separate routing tables for Mobile IP and the MANET protocol. Therefore, a route manager is introduced to coordinate route table management between Mobile IP and the modified RIP on the shared routing table. With this, neither Mobile IP or the modified RIP could modify the routing table directly. Route manipulation requests are sent to the route manager which then acts on behalf of the protocols. Although this proposal was successful in providing Internet access to the MANET, it was more proactive and did not support reactive MANET protocols since it relied on the periodic control messages of the MANET routing protocol to propagate agent advertisements.

To fill the gap, [4] presented a master thesis that detailed the use of Mobile IP on a reactive MANET routing protocol. The authors implemented Mobile IP on AODV in NS2 and then
performed some simulations to validate their study. In this proposal, the authors used Mobile IP with Foreign Agent care-of-addresses to reach Mobile Nodes from the Internet. Packets were exchanged between Home Agent and the Mobile Node via reverse tunnelling to minimize the requirements on AODV. A novel Internetworking unit was introduced between Mobile IP and AODV to ensure that no modifications be made to Mobile IP outside the MANET. The emphasis of the thesis was not on performance evaluation and so the simulations were not quite extensive. Also, the Internet draft “Global Connectivity for IPv4 Mobile Ad Hoc Networks” [5] sometimes referred to as Global4, presented a solution which involves the use of Mobile IP with foreign agent care-of-addresses and AODV just as was done in MIPMANET. However, foreign agent discovery was made a part of AODV while mobile node registrations with the foreign agent via Mobile IP is maintained.

In the paper titled “Wireless Multihop Internet Access: Gateway Discovery, Routing and Addressing” [6], an alternative approach to providing Internet Access for MANETs is describes. In this approach, Mobile IP is eliminated. The use of specific routers that serve as gateways resolve the heterogeneity between the fixed Internet and the MANET. Although the solution looks promising, it was based on IPv6 networks which has not gained popularity over IPv4 networks.

Again, the master’s thesis “A Study of Internet Connectivity for Mobile Ad Hoc Networks in NS2” [7] implemented the Internet draft “Global Connectivity for IPv6 Mobile Ad hoc Networks” [8] also known as Global6 in NS2. Global6 is an enhancement on Global4 however, Mobile IP services are not used. The AODV routing protocol was modified to include three different gateway detection mechanisms that were tested using NS2 simulations. The gateway detection mechanisms were classified as proactive, reactive and hybrid. The results from simulations showed no significant difference in packet delivery ratio between the three gateway discovery mechanisms. However, the proactive and hybrid mechanisms performed slightly better than the reactive mechanism regarding average end-to-end delay.

With regards to measuring performance of Mobile IP on MANET, [9] evaluated the performance of Mobile IP on a proactive MANET protocol, DSDV. They created simulation scenarios involving a Home Agent, three Routers, four Foreign Agents, a Corresponding Node and a varying number of Mobile Nodes at varying speeds. Number of nodes were varied to test the robustness of the solution as traffic increases within the simulation space. This was done to rightly model the real Internet world, as we are witnesses to the increasing number of mobile devices on the Internet. They evaluated performance of Mobile IP based on received packets, lost packets, throughput and End-to-End Delay (E2ED).

Also, [10] performed simulations in NS2 to evaluate the performance of Mobile IP on AODV. They used a single Home Agent and one Foreign Agent on a 670 ×670 rectangular field with a few Mobile Nodes. Performance metrics used were throughput, delay and packet overhead. Although the methodology was convincing, the simulation setup was poorly described.

In our thesis, we take these works further as we present on the same simulation platform, scenarios to evaluate the performance of Mobile IP on both reactive and proactive MANET routing on IPv4 networks. Our choice of solution involving corporation between Mobile IP and a MANET routing protocol over methods involving IPv6 is mainly because IPv4 as of now is still widely used over IPv6.

We evaluate the performance of Mobile IP on MANET by performing simulations in Network Simulator 2 (NS2). Likewise in [9], we vary number of nodes in our simulation scenarios to account for increasing traffic within the network.
3. METHODOLOGY

As mentioned earlier, we modelled a Mobile Ad Hoc Network using Network Simulator 2 software. Due to the overwhelming dynamism and cost involved in running a live experiment, we chose a computer simulation model specifically using NS2 which has a widely accepted error margin within the networking research community.

3.1. Simulation Scenario

We used a square flat surface of dimension 670m×670m with a simulation time of 200sec as seen in Figure 3. We modelled the fixed Internet using a Home Agent and four Foreign Agents all with a transmission range of 100m. We placed a gap of 5m between any two adjacent agents. The gap was to ensure that Mobile Nodes decide between Foreign Agents quickly. We simulate up to 175 Mobile Nodes. For each number of Mobile Nodes selected, we performed at least 10 different simulation runs and averaged our results.

Table 1. Simulation Parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movement model</td>
<td>Random waypoint</td>
</tr>
<tr>
<td>Traffic type</td>
<td>CBR</td>
</tr>
<tr>
<td>MAC layer</td>
<td>802.11</td>
</tr>
<tr>
<td>Transmitter range</td>
<td>~100m</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>5Mbps</td>
</tr>
<tr>
<td>Simulation time</td>
<td>200sec</td>
</tr>
<tr>
<td>Simulation field</td>
<td>670m×670m</td>
</tr>
<tr>
<td>Number of foreign agents</td>
<td>4</td>
</tr>
<tr>
<td>Packet rate</td>
<td>1Mbps</td>
</tr>
<tr>
<td>Pause time</td>
<td>0</td>
</tr>
<tr>
<td>Maximum speed</td>
<td>5m/s</td>
</tr>
</tbody>
</table>

Figure 3. Wired cum wireless scenario.
4. RESULTS AND DISCUSSION

Figure 4. Throughput in Mb/s.

Figure 4 above shows the network throughput as recorded in our simulations. Throughput is a measure of the actual data that can be sent across a channel per unit time. Throughput is usually lesser than channel bandwidth because, channel bandwidth is just a theoretical estimation of how much data could be sent across a channel per unit time. In measuring the network throughput, we excluded control packets and focused on the CBR data packets only. This is sometimes referred to as Goodput. Our results show that, throughput decreases as number of nodes increase. AOMDV had the best throughput performance whereas DSDV had the worst, as number of nodes increased. The observed difference in performance is because, AOMDV sends fewer control messages to nodes than AODV and DSDV and so the channel is less occupied and data packets can be transmitted with ease.

Figure 5. Average end-to-end delay.

Average End-to-End Delay (E2ED) is a measure of how fast a unit size of data can be transmitted across a network. E2ED is a sum of propagation delay, queueing delay, transmission delay and processing delay at intermediate nodes. E2ED is influenced by node mobility and transmission distance. Frequent mobility and longer distances result in longer E2ED. In our simulations, the
same mobility files were used in all three implementations to nullify the effects of mobility and distances when comparing E2ED performances. In all 3 implementations, E2ED increased as nodes increased. AODV and AOMDV showed similar trends as number of nodes increased. Since they are both On-demand routing protocols, path discovery affects E2ED. AOMDV and AODV use the same route discovery mechanism and deliver packets in a similar fashion therefore, within the same mobility scenario they are expected to have the same E2ED. The extra delay recorded by AOMDV is because it spends additional time during each route discovery process to discover multiple routes. DSDV had the worst E2ED performance with a sharp rise in E2ED trend as number of nodes increased. The observed difference is because, intermediate nodes spend longer processing delays as they check their route tables for next hops to forward data, whereas AOMDV and AODV use forward route setups resulting in less processing delays at intermediate nodes.

Packet Delivery Ratio (PDR) is the ratio of delivered packets to the total number of packets generated. In figure 6 above, we expressed this ratio as a percentage. PDR is inversely correlated to Packet Loss Ratio. A higher packet delivery ratio means fewer packet loss. AOMDV proved to be the most reliable in terms of packet delivery. As mentioned earlier, AOMDV discovers multiple disjoint paths between source and destination. These disjoint paths are less probable to fail all at once hence, a higher reliability in delivering packets from source to destination.
Normalized Packet Ratio (NPR) is the ratio of control packets to actual data packets. NPR can be interpreted as the number of control messages needed by the MANET routing protocol to send a single data packet from source to destination. A lesser NPR is most preferred in networks as control messages are not useful to end users. In figure 7, it is observed that, all three protocols have similar trend and values from 25 to 100 nodes. The significant difference in trend begins after 100 nodes. AOMDV had the best NPR performance whereas AODV recorded the worst performance. The observed pattern is attributed to the fact that AODV nodes repeat the route discovery procedure each time a link is broken unlike AOMDV that resorts to a secondary path earlier discovered during the immediate past route discovery phase. It is only when all the multiple paths fail that AOMDV nodes repeat the route discovery procedure.

In figure 8 and figure 9 below, we zoom in on figure 7 to show a better picture of the trends between 25 and 100 nodes.

![Figure 8. Normalized packet ratio (A).](image)

![Figure 9. Normalized packet ratio (B).](image)
5. CONCLUSIONS

We have presented a performance evaluation of Mobile IP on MANET. We have studied for the first time the performance of Mobile IP on MANET using Ad Hoc On-demand Multiple Distance Vector (AOMDV) for routing packets inside the MANET. We have also studied the performance of Mobile IP using Destination-sequenced Distance Vector (DSDV) and Ad Hoc On-demand Distance Vector (AODV) for routing packets inside the MANET.

We modelled Mobile IP on MANET using a wired cum wireless scenario in NS2. Our scenario included one Home Agent, a Corresponding Node and four Foreign Agents. Mobile nodes moved randomly between these sub networks and registered with an agent whenever they entered a new sub network. Communication between Mobile Nodes and Corresponding Node continued even as Mobile Nodes moved across sub networks.

Results obtained from simulation suggested that On-demand routing improved the performance of Mobile IP on MANET regarding the average end-to-end delay, throughput, packet delivery ratio and normalized packet ratio. Specifically, Mobile IP on AOMDV outperformed AODV and DSDV.

ACKNOWLEDGEMENTS

We are grateful to God for the gift of life and good health throughout this study. We thank family, friends and colleagues for the support and encouragements thus far.

REFERENCES


AUTHORS

Kingsley K. Ofosu is currently an MPhil student at the Department of Computer Science, University of Ghana. He received his BSc. degree in Psychology and Computer Science (Combined) from University of Ghana in 2013. His research interests include mobile networks, cloud computing and intelligent systems.

Jamal-Deen Abdulai received his BSc. degree in Computer Science in 2002 from the Kwame Nkrumah University of Science and Technology (KNUST). In 2006, he was awarded an MPhil and later in 2009 a PhD in Computer Science by the Department of Computing Science at the University of Glasgow, UK.

Dr. Abdulai’s is currently a lecturer at the Department of Computer Science, University of Ghana, investigating how probabilistic methods can be used to optimize the performance of both wired and wireless networks. His research interests include Performance modelling and evaluation of Mobile Wireless Ad hoc and Sensor Networks, Network Security and Management, Embedded Systems, Parallel and Distributed Systems, Artificial Intelligence and its application in information security.
Ferdinand A. Katsriku is the current head of the department of computer science, University of Ghana. In 1989, he received an M. Eng. in computer systems engineering with distinction from Kharkov Polytechnic Institute. He was awarded an MSc in Laser Engineering and Pulsed Power Technology in August 1992 and later awarded postgraduate certificate in education from King’s College London in 1996. In 2000, he received a PhD in Information Engineering from City University, London. His current research interests are wireless sensor networks, cognitive radio and water quality monitoring. He is also a technical Reviewer of EPSRC grants and for major publishers including John Wiley and Elsevier.
ON THE DUALITY FEATURE OF P-CLASS PROBLEMS AND NP COMPLETE PROBLEMS

Wenhong Tian

1University of Electronic Science and Technology of China, Chengdu, China.
2Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Science.

ABSTRACT

In term of computational complexity, P-class (abbreviated as P) problems are polynomial-time solvable by deterministic Turing machine while NP complete (abbreviated as NPC) problems are polynomial-time solvable by nondeterministic Turing machine. P and NPC problems are regularly treated in different classes. Determining whether or not it is possible to solve NPC problems quickly, is one of the principal unsolved problems in computer science today. In this paper, a new perspective is provided: both P problems and NPC problems have the duality feature in terms of computational complexity of asymptotic efficiency of algorithms.

KEYWORDS

P Problems; NP problems; NP Complete Problems; the P versus NP Problem; The Duality Feature.

1. INTRODUCTION

In 1971, Cook [1] firstly established a theorem that a class of problems can be P-reducible (polynomial time reducible) to each other and each of them can be P-reducible to Boolean Satisfiability (SAT) problem, this class of problems is called NP (nondeterministic polynomial time) problems. Karp [2] applied Cook’s 1971 theorem that the SAT problem is NP Complete (also called the Cook-Levin theorem) to show that there is a polynomial time reduction from the SAT problem to each of 21 combinatorial problems, thereby showing that they are all NP complete (NPC). This was one of the first demonstrations that many natural computational problems occurring throughout computer science are computationally intractable, and it drove interest in the study of NP-completeness and the P versus NP problem [3]. The P versus NP problem, determining whether or not it is possible to solve NP problems quickly, is one of the principal unsolved problems in computer science today and listed as one of seven millennium problems [3,4], challenging tens of thousands of researchers.

Simply speaking, P problems mean that the class of problems can be solved exactly in polynomial time while NPC problems stands for a class of problems which might not be solvable in polynomial time. NPC problems has far-reaching consequences to other problems in mathematics, biology, philosophy and cryptography. More specifically, in Big O-notation (refer...
to Definition 1 and 2) of computational complexity in asymptotic efficiency of algorithms, P problems can be solved in polynomial time of $O(n^k)$ for some constant $k$ where $n$ is the size of input to the problem, while NPC problems may have computational complexity of $O(2^n)$ including both exponential time and sub-exponential time, where $c$ is a positive constant larger than zero.

Karp [2] ever claimed that if any of NPC problems have efficient polynomial time algorithms, then they all do. It is for this reason that research into the P versus NP problem centers on NPC problems, i.e., looking for efficient polynomial time algorithms for NPC problems. Through decades’ efforts by many researchers, it is still an open question. There are quite many results but none of them is commonly accepted yet by the research community.

The current author also classified NPC problems based on their natures and other methods [5] and found that NPC problems are not equivalent in computational complexity.

In the following, we show a new perspective: both P problems and NPC problems have duality features of each other in terms of computational complexity of asymptotic efficiency of algorithms, especially by considering the representation of the input.

2. PROBLEM FORMULATIONS

The following definitions are based on the computational complexity of different problems in the worst case.

**Definition 1:** The asymptotic efficiency of algorithms [6,7]: concerns with how the running time of an algorithm increases with the size of the input in the limit, as the size of the input increases without bound.

**Definition 2:** The O-notation of computational complexity of an algorithm: asymptotically bounds a function from above within a constant factor [6,7]. For a given function $g(n)$, we denote by $O(g(n))$:

$O(g(n)) = \{f(n) : \text{there exist positive constants } c \text{ and } n_0 \text{ such that } 0 \leq f(n) \leq cg(n) \text{ for } n \geq n_0\}$.

Fig.1 shows the intuition behind O-notation. For all values $n$ to the right of $n_0$, the value of function $f(n)$ is on or below $g(n)$.

![Figure 1: $f(n) = O(g(n))$](image)

Figure 1.$f(n) = O(g(n))$ [6]
Since O-notation describes an upper bound, when we use it to bound the worst-case running time of an algorithm, we have a bound on the running time of the algorithm on every input. For example, the doubly nested loop structure of the insertion sort algorithm has an \( O(n^2) \) upper bound; equivalently, we mean that the worst-case running time is \( O(n^2) \).

**Definition 3:** P-class problems in term of O-notation of computational complexity: in the worst case, their exact algorithms have computational complexity of \( O(n^k) \) for some constant \( k \) where \( n \) is the size of input to the problem.

**Definition 4:** NPC problems: in the worst case, their exact solutions may have computational complexity of \( O(2^n) \) \textbf{in O-notation}, where \( n \) is the size of input to the problem and \( c \) is a positive constant.

Table 1 provides a summary of variables used in this paper. Table 2 shows computational complexities of exact solutions to some NPC problems (abstracted from [8]).

**Table 1. The summary of variables**

<table>
<thead>
<tr>
<th>Variables</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k, c, n_0 )</td>
<td>a positive constant</td>
</tr>
<tr>
<td>( n, e )</td>
<td>The size of inputs to a problem</td>
</tr>
<tr>
<td>( s, m )</td>
<td>The number of bits (binary)</td>
</tr>
<tr>
<td>( B_0 )</td>
<td>( \min(2^{2n}, n^k, 2^m, 2^s) )</td>
</tr>
<tr>
<td>( W, C, b )</td>
<td>a positive number</td>
</tr>
</tbody>
</table>

**Table 2. The complexities of exact solutions to some NPC problems from [8]**

<table>
<thead>
<tr>
<th>Problem</th>
<th>Complexity</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIS (maximum independent set)</td>
<td>1.1996*</td>
<td>Packing</td>
</tr>
<tr>
<td>Set covering problem (SCP) *</td>
<td>1.1996*</td>
<td>Covering</td>
</tr>
<tr>
<td>3CP (3-coloring of planar graph)</td>
<td>1.3446*</td>
<td>Partitioning</td>
</tr>
<tr>
<td>SSP (subset sum problem) [7]</td>
<td>( nW )</td>
<td>Numerical</td>
</tr>
<tr>
<td>3SAT</td>
<td>1.4802*</td>
<td>Satisfaction</td>
</tr>
<tr>
<td>TSP</td>
<td>( n^2n )</td>
<td>Sequencing</td>
</tr>
</tbody>
</table>

*: derived from MIS since SCP and MIS are complementary [7].

We also consider that the representation of the input to the problem by \( m \) bits (in binary) and the space complexity of a problem by \( s \) bits (in binary) in memory.

**Fact 1:** A regular system (the computer hardware, called the system in the following) has capability of handling \( B_0 \) numbers in term of computational and space complexity, i.e., handling the input efficiently within reasonable time (may be within a few hours or minutes depending on the applications) without overflowing the system, where \( B_0 \) may be related to the memory size or whichever (CPU, memory, disk et.) is the bottleneck of the system.
Based on the definitions above, the capability $B_0$ can be represented by

$$B_0 = \min(2^{cn}, n^k, 2^m, 2^n)$$

(1)

3. RESULTS

**Lemma 1.** Dynamic programming algorithm for Subset Sum problem (SSP) is pseudo-polynomial time.

**Proof.** removed

**Lemma 2.** Any NPC problem can be reduced to SSP.

**Theorem 1:** The instances of SSP can have computational complexity of NPC problems and P-class problems, this is called the duality feature in this paper.

**Proof:** removed

**Fact 2:** The instances of other NPC problems also have the duality feature.

**Proof:** removed.

In Table 3, records of optimum solutions to TSP problem are provided. One can see that the size of the problem is increasing as year goes. Notice that the instance with 1904711 nodes is still not yet solved exactly but just has a good lower bound [14]. Similar results are also observed for other NPC problems. For real-life problems with small or moderate number of variables, they can be solved to exactly very easily (LKH [15] and Concorde [16] are considered as two of the best such solvers which can now find exact solutions for medium size TSP Problems). However, if the problem size increases to very large, finding efficient solutions to NPC problems just become intractable.

**Theorem 2:** The P-class Problems can have the duality feature in terms of computational complexity.

**Proof:** removed.

**Remarks:** Actually SSP can be treated as polynomial time solvable problem (P-class problem) when both $n$ and $W$ is not very large while it is of exponential complexity when both $n$ and $W$ are very large.

**Observation 1:** Even one can find efficient solution (polynomial time algorithm) to one or more NPC problems, some P problems may become exponential time solvable in computational complexity when the representation of the input to them become very large.

4. DISCUSSIONS AND CONCLUSION

NPC problems have different natures, they can be classified into six basic genres [2, 7], i.e., Satisfaction, Packing, Covering, Partitioning, Sequencing, Numerical computing.

Originally, PRIMES and Graph Isomorphism were hard to determine to be in NPC or not in Cook’s paper [1]. In 2004, PRIMES is found in P-class and accepted by the research community [10]. And in 2015, Graph isomorphism is reported to have Quasipolynomial time solution [11,
12], though the results are still under verification. These show some new perspectives and trends on NPC problems.

As another perspective, it is recently proved mathematically that memcomputing machines (a novel non-Turing paradigm) have the same computational power of nondeterministic Turing machines [13]. Therefore, they can solve NPC problems in polynomial time with resources that only grow polynomially with the input size.

NPC problems and the P versus NP problem challenge many researchers to tackle them through decades of efforts. In this paper, a new perspective is provided: both P problems and NPC problems have the duality feature in terms of computational complexity of asymptotic efficiency of algorithms.

Table 3. Records of optimum solutions to TSP problems [14] where $n$ is the number of nodes in TSP. All TSP problems in the table are solved to optimum except for the last one

<table>
<thead>
<tr>
<th>n</th>
<th>Year (solved)</th>
<th>Node type</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>1954</td>
<td>USA cities</td>
</tr>
<tr>
<td>64</td>
<td>1971</td>
<td>random nodes</td>
</tr>
<tr>
<td>80</td>
<td>1975</td>
<td>random nodes</td>
</tr>
<tr>
<td>120</td>
<td>1977</td>
<td>Germany cities</td>
</tr>
<tr>
<td>318</td>
<td>1987</td>
<td>cities</td>
</tr>
<tr>
<td>532</td>
<td>1987</td>
<td>USA cities</td>
</tr>
<tr>
<td>666</td>
<td>1987</td>
<td>World cities</td>
</tr>
<tr>
<td>1002</td>
<td>1987</td>
<td>cities</td>
</tr>
<tr>
<td>2392</td>
<td>1987</td>
<td>cities</td>
</tr>
<tr>
<td>3038</td>
<td>1992</td>
<td>cities</td>
</tr>
<tr>
<td>13509</td>
<td>1998</td>
<td>USA cities</td>
</tr>
<tr>
<td>15112</td>
<td>2001</td>
<td>cities</td>
</tr>
<tr>
<td>24978</td>
<td>2004</td>
<td>Sweden cities</td>
</tr>
<tr>
<td>85900</td>
<td>2006</td>
<td>cities</td>
</tr>
<tr>
<td>100000</td>
<td>2009</td>
<td>Japan</td>
</tr>
<tr>
<td>1904711</td>
<td>2010*</td>
<td>World TSP Challenge</td>
</tr>
</tbody>
</table>

ACKNOWLEDGMENTS

The abstract version of this manuscript is posted at my research gate.

REFERENCES


PERCEPTION OF HOME-SHARING INDUSTRY: A STUDY OF COMPARING AIRBNB AND COUCHSURFING

JinTao Huang, KangNyeon Yi and JiHye Han

Department of Interaction Science, SungKyunKwan University, Seoul, South Korea

ABSTRACT

The aim of this paper is to examine people’s perceptions of home sharing industry through comparing Airbnb and Couchsurfing, using python to crawling twitter’s data. We conducted semantic network analysis by using UCINET, which is embedded with NETDRAW, for calculating betweenness centrality and visualizing semantic network based on multidimensional scaling(MDS). And we also used LIWC to analysis public sentimental perceptions. The present study’s results show that the characteristics discovered about Airbnb and Couchsurfing in cyberspace have possible future directions in view of word usage frequency, centrality and semantic networks. Moreover, the results show that in sentimental aspects, there are different public trends of emotions on Airbnb and Couchsurfing. Through those results, researchers provide information to understand which sectors should entrepreneurs put more efforts and money.

KEYWORDS

Sharing Economy, Home-sharing, Airbnb, Couchsurfing, Semantic Network Analysis

1. INTRODUCTION

Sharing is a phenomenon as old as humankind, while collaborative consumption and the “sharing economy” are phenomena born of the Internet age (Russell, 2014). The sharing economy is a socio-economic system that coordinates “the peer-to-peer-based activity of obtaining, giving, or sharing the access to goods and services” through “community-based online services” (Hamari et al., 2015). This economy system is a new cooperative production and consumption mode that enables the sharing of various goods and services by virtue of internet. These types of economies have recently become one of the important forms of achieving sustainable development by use of scientific and technological innovations. And especially, we focus on space sharing, which is one of the sharing economy. Space sharing includes releasing the information and prices of idle properties through various platforms on the Internet for tourists, people who are taking business trips, and other people with a need for, or those who are looking for housing, to select (Xinyuea et al., 2016). The home sharing economy already has a foothold in the vacation rental market, which accounted for 9% of the traditional hotel market in 2014. This market is now becoming more organized and gaining more widespread use as trust in these services increases (Saussier, 2015).
However, not many sharing economy platforms survive and only a few are economically successful. The success of these platforms in developing from start-ups to sustainable businesses depends on their ability first to mobilize initial user networks and then to maintain users' active participation and ensure growth of the user base sufficient for sustainable revenue generation (Constantiou et al., 2016). We investigate the case of Airbnb and Couchsurfing, which is the mainstream of home sharing. Airbnb and Couchsurfing are sharing platforms network in the home sharing industry. Our study is about examine public sentiments on these home-sharing companies. We could find public positive and negative emotion about Airbnb and Couchsurfing in Twitter from 2010 to 2017. We discover the different characteristics and identify the historical Changes of tweets. Using UCINET, the authors analysis the betweenness centrality and visualizing semantic network based on multidimensional scaling(MDS). Our study provides public perceptions of home sharing industry through comparing Airbnb and Couchsurfing.

2. LITERATURE REVIEW

2.1. Airbnb

Airbnb is an online community for vacation rentals that allows travelers and accommodation providers to list, find and rent lodging. Airbnb was founded in 2008 in San Francisco. With the increasing demands in travel accommodation, the company discovered the great potential of this market; thus, Airbnb was then created to provide low-cost and unique travel experience to capture the market demand. Airbnb has focused on the economic impact they can bring internationally and endeavors to align it with the goal of maintaining the current development in the neighborhoods in which they operate. On average, Airbnb generated more than $300 million in economic activity in several major cities, such as New York, Paris, Berlin, and London (Chan et al., 2016).

2.2. Couchsurfing

Couchsurfing is a non-profit organization that ‘seeks to internationally network people and places, create educational exchanges, raise collective consciousness, spread tolerance and facilitate cultural understanding’ (Couchsurfing, 2006). The project was originally launched in 2003 (Molz & Jennie Germann, 2013). This community works on the premise that a user has potential friends around the world. These friends are hosts who are willing to let a stranger sleep over at their home rather than staying at a hotel. People can find hosts and share their experiences via the website, which is made up of a community of 10 million people in 200,000 cities (Global Investor, 2015).

2.3. Applicable theory

The authors found applicable theories for applying the results.

*Cognitive dissonance theory*

“Dissonance is aroused whenever a person engages in an unpleasant activity to obtain some desirable outcome. From the cognition that the activity is unpleasant, it follows that one would not engage in the activity; the cognition that the activity is unpleasant is dissonant with engaging
in the activity. Dissonance should be greater, the greater the unpleasant effort required to obtain the outcome.” (Harmon-Jones & Mills, 1999)

Consider someone who stays an expensive Airbnb house but discovers that it is not comfortable to stay long. Dissonance exists between their beliefs that they have paid a good Airbnb house and that a good house should be comfortable. Dissonance could be eliminated by deciding that it does not matter since the house is mainly used for short trips (reducing the importance of the dissonant belief) or focusing on the house’s strengths such as safety, appearance (thereby adding more reasonable beliefs). Thus people can rationalize their conduct by thinking so.

2.4. Mere exposure effect

The mere exposure effect is a psychological phenomenon by which people tend to develop a preference for things or people that are more familiar to them than others. Repeated exposure increases familiarity. This effect is therefore also known as the familiarity effect. Therefore misattribution of positive feelings (due to fluency) to the ad or brand comes into effect. This means that with mere exposure the more often you see an information the better you can process it and the more fluent it is. Human beings like easy processing rather than difficult cognitive processes so they stick with familiar things. This may be the explanation for their positive feelings towards an ad or brand (Falkenbach et al., 2013). When people evaluate Airbnb and Couchsurfing, there are many factors why people evaluate that way. Among them, mere exposure effect theory can be a factor. We can see Airbnb’s commodity advertisement a lot more than Couchsurfing. Through the advertisement, people are exposed several times with Airbnb, and it affects raising people’s positive evaluation about Airbnb because of the familiarity that improved by exposure.

2.5. Research questions

RQ 1. Semantic Networks. In the datasets from tweets “Airbnb” and “Couchsurfing”, what were the different characteristics discovered?
RQ 1-1. In view of word usage frequency, centrality and semantic networks, what were the characteristics found in tweets about “Airbnb” and “Couchsurfing” respectively?
RQ 1-2. As time went by, what were the remarkable differences discovered from tweets about “Airbnb” and “Couchsurfing” respectively?
RQ 2. What are the theoretical and practical implications of the study’s findings?

3. METHODS

3.1. Data gathering

The datasets were retrieved from Twitter by using Python. We gathered past tweets from Twitter through a python library. Twitter is an online news and social networking service which users post and interact with messages, and it is frequently used by the general public. In addition, to investigate how the home-sharing industry has developed since related enterprises showed up, we used two keywords, “Airbnb” and “Couchsurfing”, which are two representative companies in the home-sharing industry. The time periods of datasets were divided into four groups respectively: before 2015(2010-2014), 2015, 2016 and 2017.
3.2. Semantic network analysis

This study employed semantic network analysis (SNA), which is a numerical and graphical measurement method for a variety of text data. Semantic network analysis is a semi-automated text analysis for determining the most frequently used words and the differences among texts in terms of word frequency and clusters (Lee, Kim, & Rosen, 2009; Doerfel & Barnett, 1999). Python was employed for collecting data from Twitter. ‘WORDij’ (Danowski, 2010) was also employed to convert text data from Twitter, into co-occurrence matrices. ‘WORDij’ is word-pair based software for semantic structure retrieval without any prior categories (Danowski, 2010). Moreover, according to Danowski(2010), word pairs are made based on word proximity which is calculated by a “window that slides through the text, counting all word pairs inside as it moves from word in the full text”. For calculating betweenness centrality and visualizing semantic network based on multidimensional scaling (MDS), we used UCINET6 for calculating centrality from co-occurrence matrix. Besides, we also used NETDRAW, which is embedded from UCINET6, for drawing semantic network based on co-occurrence matrix and centrality.

4. RESULTS


Figure 1 is about semantic network of Airbnb from Twitter. ‘uber’, ‘sharing’, and, ‘economy’ have high frequency and betweenness centrality. There are many names of cities such as ‘Amsterdam’, ‘San Francisco’, ‘New York’ in the network. There are words related to a case of hidden camera issue such as ‘german’, ‘hidden’, ‘camera’, ‘suing’. In 2015, a german woman argued that she was unaware of the camera, which allegedly captured both naked footage and private conversations. In addition, there are four clusters: the first cluster (A) is about ‘Airbnb’-relevant issues; the second cluster (B) stands for user experience on mobile app and web site; the third cluster (C) is about trip-relevant startups; the fourth cluster (D) describes car-sharing industry (Uber).
Table 1. Frequency and centrality of ‘Airbnb’ data from 2010 to 2017.

<table>
<thead>
<tr>
<th></th>
<th>2010-2015</th>
<th>Frequency</th>
<th>Centrality</th>
<th>2016</th>
<th>Frequency</th>
<th>Centrality</th>
<th>2017</th>
<th>Frequency</th>
<th>Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Airbnb</td>
<td>Airbnb</td>
<td></td>
<td></td>
<td>Airbnb</td>
<td></td>
<td></td>
<td>Airbnb</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Airbnb</td>
<td>Airbnb</td>
<td></td>
<td></td>
<td>Airbnb</td>
<td></td>
<td></td>
<td>Airbnb</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Uber</td>
<td>Uber</td>
<td></td>
<td></td>
<td>Travel</td>
<td></td>
<td></td>
<td>App</td>
<td>First</td>
</tr>
<tr>
<td>3</td>
<td>Travel</td>
<td>Mashable</td>
<td></td>
<td></td>
<td>Ceo</td>
<td></td>
<td></td>
<td>Travel</td>
<td>Hosts</td>
</tr>
<tr>
<td>4</td>
<td>Sharing</td>
<td>Birdnb</td>
<td></td>
<td></td>
<td>Twitter</td>
<td></td>
<td></td>
<td>Ceo</td>
<td>First</td>
</tr>
<tr>
<td>5</td>
<td>Business</td>
<td>Money</td>
<td></td>
<td></td>
<td>Uber</td>
<td></td>
<td></td>
<td>Twitter</td>
<td>Hi</td>
</tr>
<tr>
<td>6</td>
<td>Economy</td>
<td>Sharing</td>
<td></td>
<td></td>
<td>Bitcoin</td>
<td></td>
<td></td>
<td>Uber</td>
<td>Dm</td>
</tr>
<tr>
<td>7</td>
<td>Stay</td>
<td>Economy</td>
<td></td>
<td></td>
<td>Needs</td>
<td></td>
<td></td>
<td>Bitcoin</td>
<td>Study</td>
</tr>
<tr>
<td>8</td>
<td>Hosts</td>
<td>Tech</td>
<td></td>
<td></td>
<td>Bookings</td>
<td></td>
<td></td>
<td>Most</td>
<td>Reject</td>
</tr>
<tr>
<td>9</td>
<td>Love</td>
<td>Order</td>
<td></td>
<td></td>
<td>Grow</td>
<td></td>
<td></td>
<td>Users</td>
<td>Guests</td>
</tr>
<tr>
<td>10</td>
<td>Company</td>
<td>Trip</td>
<td></td>
<td></td>
<td>Full-fledged</td>
<td></td>
<td></td>
<td>Cheksy</td>
<td>Email</td>
</tr>
<tr>
<td>11</td>
<td>Amsterdam</td>
<td>Business</td>
<td></td>
<td></td>
<td>Hub</td>
<td></td>
<td></td>
<td>Gift</td>
<td>Help</td>
</tr>
<tr>
<td>12</td>
<td>Hotel</td>
<td>Amsterdam</td>
<td></td>
<td></td>
<td>Hosts</td>
<td></td>
<td></td>
<td>Business</td>
<td>Thanks</td>
</tr>
<tr>
<td>13</td>
<td>Startup</td>
<td>Hosts</td>
<td></td>
<td></td>
<td>App</td>
<td></td>
<td></td>
<td>Housing</td>
<td>Stay</td>
</tr>
<tr>
<td>14</td>
<td>Apartment</td>
<td>Share</td>
<td></td>
<td></td>
<td>Rent</td>
<td></td>
<td></td>
<td>Customers</td>
<td>Business</td>
</tr>
<tr>
<td>15</td>
<td>Place</td>
<td>Site</td>
<td></td>
<td></td>
<td>Stay</td>
<td></td>
<td></td>
<td>Start</td>
<td>Rent</td>
</tr>
<tr>
<td>16</td>
<td>San</td>
<td>Mobile</td>
<td></td>
<td></td>
<td>Hotel</td>
<td></td>
<td></td>
<td>Hosts</td>
<td>Finds</td>
</tr>
<tr>
<td>17</td>
<td>Klm</td>
<td>Hotel</td>
<td></td>
<td></td>
<td>Business</td>
<td></td>
<td></td>
<td>Tenants</td>
<td>Disabilities</td>
</tr>
<tr>
<td>18</td>
<td>Tourist</td>
<td>Host</td>
<td></td>
<td></td>
<td>First</td>
<td></td>
<td></td>
<td>Hub</td>
<td>Trip</td>
</tr>
<tr>
<td>19</td>
<td>Nyc</td>
<td>App</td>
<td></td>
<td></td>
<td>Updates</td>
<td></td>
<td></td>
<td>Grow</td>
<td>People</td>
</tr>
<tr>
<td>20</td>
<td>Rent</td>
<td>People</td>
<td></td>
<td></td>
<td>Front</td>
<td></td>
<td></td>
<td>Full-fledged</td>
<td>Love</td>
</tr>
</tbody>
</table>

Table 2. Frequency and centrality of ‘Couchsurfing’ data from 2010 to 2017.

<table>
<thead>
<tr>
<th></th>
<th>2010-2015</th>
<th>Frequency</th>
<th>Centrality</th>
<th>2016</th>
<th>Frequency</th>
<th>Centrality</th>
<th>2017</th>
<th>Frequency</th>
<th>Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Couchsurfing</td>
<td>Couchsurfing</td>
<td></td>
<td></td>
<td>Couchsurfing</td>
<td></td>
<td></td>
<td>Couchsurfing</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Travel</td>
<td>Data</td>
<td></td>
<td></td>
<td>Travel</td>
<td></td>
<td></td>
<td>People</td>
<td>Travel</td>
</tr>
<tr>
<td>2</td>
<td>Site</td>
<td>Tour</td>
<td>Moscow</td>
<td>Lot</td>
<td>Host</td>
<td>Airbnb</td>
<td></td>
<td>Host</td>
<td>Airbnb</td>
</tr>
<tr>
<td>3</td>
<td>People</td>
<td>Business</td>
<td></td>
<td></td>
<td>Host</td>
<td>Better</td>
<td>Traveling</td>
<td>City</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Airbnb</td>
<td>Check</td>
<td>Airbnb</td>
<td>Italian</td>
<td>Experience</td>
<td>Host</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Moscow</td>
<td>App</td>
<td>World</td>
<td>Travel</td>
<td>Airbnb</td>
<td>Experience</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Check</td>
<td>Her</td>
<td>People</td>
<td>Italy</td>
<td>First</td>
<td>Traveler</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Great</td>
<td>Shes</td>
<td>Italian</td>
<td>Couches</td>
<td>People</td>
<td>First</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Host</td>
<td>Building</td>
<td>Site</td>
<td>Policeman</td>
<td>Ttot</td>
<td>Ttot</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Website</td>
<td>Ever</td>
<td>First</td>
<td>Rape</td>
<td>Friends</td>
<td>Added</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Love</td>
<td>Hook-up</td>
<td>Couch</td>
<td>Site</td>
<td>Moscow</td>
<td>Ways</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Friends</td>
<td>Travel</td>
<td>Photo</td>
<td>Host</td>
<td>Great</td>
<td>People</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Time</td>
<td>Use</td>
<td>Great</td>
<td>World</td>
<td>Check</td>
<td>Best</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Thanks</td>
<td>Sharing</td>
<td>Friends</td>
<td>Moscow</td>
<td>Best</td>
<td>Why</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Chrome</td>
<td>Management</td>
<td>Journey</td>
<td>Airbnb</td>
<td>Free</td>
<td>Youtube</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>App</td>
<td>Brand</td>
<td>Time</td>
<td>First</td>
<td>World</td>
<td>Casey</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Couch</td>
<td>Fail</td>
<td>Round</td>
<td>Website</td>
<td>Time</td>
<td>Free</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Good</td>
<td>Poor</td>
<td>Experience</td>
<td>Women</td>
<td>Meet</td>
<td>Fenton</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Android</td>
<td>Kill</td>
<td>Photographers</td>
<td>Accused</td>
<td>Couch</td>
<td>Friends</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Cs</td>
<td>Airbnb</td>
<td>Sleep</td>
<td>News</td>
<td>Love</td>
<td>Fiona</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Figure 1. Semantic Network from Twitter on 'Airbnb' from 2010 to 2015

Figure 2. Semantic Network from Twitter on 'Airbnb' in 2016
There are words relevant to a suit case about contract on lease of apartment such as ‘tenants’, ‘fights’ in Figure 2. There are words about blockchain technology such as ‘bitcoin’. There are three clusters: the first cluster (A) depicts travel-relevant services and startups; the second cluster (B) stands for user experience; the third cluster (C) is about car-sharing industry (Uber, Careem, Mercedes-benz).

There are words about discrimination on disabilities from academia such as ‘disabilities’, ‘researchers’, ‘study’ in Figure 3. There are three clusters: the first cluster (A) describes house-relevant comments (interior design, architecture, decoration); the second cluster (B) stands for ‘Airbnb’-relevant issues including enter chinese market; the third cluster (C) is about travel-relevant reports and perception.

![Semantic Network from Twitter on ‘Airbnb’ in 2017](image)

There are words related to sex culture such as ‘hook-up’, ‘secret’, ‘sex’ in Figure 4. There are names of cities such as ‘Moscow’, ‘San Francisco’, ‘Amsterdam’. There are three clusters: the first cluster (A) is about violence-relevant news; the second cluster (B) describes user experience (os, android, site, apps, web, etc.); the third cluster (C) stands for Competitors and other startups.

The word ‘ttot’ is abbreviation of ‘Travel Talk on Twitter’. There are words related to a murder case in Nepal such as ‘Nepal’, ‘American’, ‘murdered’ in Figure 5. There are three clusters: the first cluster (A) depicts travel-relevant issues; the second cluster (B) is about violence-relevant news; the third cluster (C) stands for competitors and other startups.

There are many words relevant to travel experience such as ‘experience’, ‘thailand’, ‘sudan’, ‘cambodia’, ‘iran’, ‘toronto’, ‘quebec’, ‘san francisco’, ‘europe’, ‘russia’ in Figure 6. There are three clusters: the first cluster (A) describes travel experiences; the second cluster (B) is about violence-relevant news; the third cluster (C) stands for competitors-relevant issues.
Figure 4. Semantic Network from Twitter on ‘Couchsurfing’ from 2010 to 2015

Figure 5. Semantic Network from Twitter on ‘Couchsurfing’ in 2016
5. CONCLUSIONS

By analysing the semantic networks of text data from Twitter and the words which have high frequency and centrality in the networks from four different time periods, the current study seeks to figure out how home sharing economy change over time from early stage to 2017. In most semantic networks in this study, some kind of flash and horrible event including murder or sexual assault cases has been mentioned frequently. As a result, their frequency and centrality are high. In the semantic networks from keyword ‘airbnb’, ‘uber’ appears frequently, because ‘uber’ is another popular sharing economic service for travelling. In the networks from search keyword ‘couchsurfing’, ‘airbnb’ has often been mentioned, because ‘airbnb’ service hosts can provide their houses to travelers for ‘couchsurfing’. Since 2016 the word ‘Youtube’ has been mentioned frequently than ever in the networks on ‘couchsurfing’. Perhaps a lot of couchsurfing hosts make Youtube main advertisement tools. Additionally, there are more violence-relevant issues about ‘couchsurfing’ than ‘airbnb’.

Time span of data from twitter in this study is not that large enough to cover the development periods of Airbnb and couchsurfing. In addition, Twitter is not only one representative social media website so that the dataset of universality had limitations. For further development, time span of data should expand to around 10 years ago. The variety of data sources should be used by not only twitter but also facebook, reddit and other popular social media websites. Deep Learning method based on Natural Language could develop Semantic Network Analysis including clustering and categorizing data set. (e.g., Deep Learning Open Source including Google Tensor Flow, etc). Innovation theories including ‘Diffusion of innovation theory’ could be employed in the further study.
REFERENCES


AUTHORS

JinTao Huang
Master of interaction science, SungKyunKwan University

KangNyeon Yi
Ph.D of interaction science, SungKyunKwan University

JiHye Han
Master of interaction science, SungKyunKwan University
SOCIAL LIVE-STREAMING: TWITCH.TV 
AND USES AND GRATIFICATION THEORY
SOCIAL NETWORK ANALYSIS

James Dux

Department of Interaction Science, Sungkyunkwan University, 
Seoul, South Korea

ABSTRACT

The recent explosive growth of live-streaming, particularly video game streaming on Twitch.tv, has led to the development of young communities in a diverse and frequently evolving virtual social space. This study is a Social Network Analysis that aims to uncover how communities are interlocked based on attributes such as Community membership and stream style/type. Additionally, the network geography is related to the Uses and Gratification Theory (UG) that Twitch has been proven to fulfill according to genre and stream type (Sjoblom, Torhonen, Hamari, & Macey, 2017). By collecting data in a user-based fashion and using tools like Gephi, this project presents the Twitch community in a way that reveals users’ interest seeking behavior trends. Implications from this study are a resource for: understanding Twitch and live-streaming communities at large, streamers looking for an insight into their audience’s interests, and developers seeking the best way to provide and display beneficial content.

KEYWORDS

Social Network Analysis, Uses and Gratification, Twitch.tv, Social Live Streaming, Social Communities

1. INTRODUCTION

Today’s internet speeds allow for a new type of media to flourish: live streaming video. Twitch.tv (colloquially known as just ‘Twitch’) is a prime example of this medium that creates a virtual ‘Third Place’ where users watch live-streamed videos while interacting with each other simultaneously (Brundl, Matt, & Hess, 2017). ‘Streamers’ are the performers who broadcast the video while engaging with their live audience ‘followers’ to varying degrees. Twitch focuses on the streaming of video games such as ‘PLAYERUNKNOWN’S BATTLEGROUNDS’ (PUBG) and ‘League of Legends’ (LoL). Why is live-streaming an important subject? Twitch.tv was only founded in 2011, but was bought by Amazon in 2014 for nearly 1 billion US dollars. Twitch.tv boasts 100+ million unique monthly visitors, and about 10 million daily active users (Smith, 2017). Clearly this new type of virtual ‘living room experience’ has struck a chord with audiences, and deserves further and deeper research.

After thorough literature review of related topics and papers, it is clear that Twitch presents a unique opportunity to learn more about this new digital form of interaction. Previous research has
proven that live-stream users enjoy and benefit from the social interaction provided by Twitch and similar sites. In particular, the “Co-Experience” provided by this medium has been shown to have a strong influence on the Active and Passive Enjoyment of live-streaming, marking the social aspect as a significant use motivation (Brundl et al., 2017).

Twitch is a social live-streaming site heavily driven by its community (and sub-communities) of streamers and the users who watch, follow, and subscribe to them. The availability of this connection information and the nature of this kind of organization lends itself very well to social network analysis. Social network analysis is “a visual and a mathematical analysis of human relationships”, which gives researchers the opportunity to dissect how users in a network are connected by peers (Krebs, n.d.). Not only do social networks reveal connections, but can also be used to discover groupings based on node (individuals’) attributes like common interests, communities, and other values. Analysis of these social networks provides benefits to the overall organization and members themselves. For Twitch streamers this might mean insight in where to “strengthen relationships within and across groups”; for Twitch.tv it could mean the ability to provide better ‘related content’ or to identify key nodes worthy of increased sponsorship (Serrat, 2009). It’s clear that social network analysis of Twitch is a worthwhile endeavor that can provide real value for the community and site.

2. LITERATURE REVIEW

2.1. Twitch

Live streaming platforms such as Twitch allow streamers to project live video from their home to viewers over the internet. This video can be sourced directly from their PC for the intent of displaying video game game-play, program use, or any other entertaining/instructional visuals. Twitch in particular is primarily a video game streaming site. Streaming video directly from a webcam is also available. It’s very common for streamers to overlay their webcam video inside of their game-play video, so audiences can see their face as the play and talk (see Figure 1). In fact, this method has become expected; viewers enjoy having the opportunity to share in the emotions of the streamer through their facial expressions (Hamilton, Garretson, & Kerne, 2014).

The stream is only half of the Twitch/live streaming equation; the user chat is the other half. The user chat is a simple chat room in which viewers of the stream can communicate with each other (and the streamer) with text and emoticons. The chat is a deeply significant part of the Twitch experience, providing users a way to chat with each other about what’s happening on the stream as events unfold. In a 2015 survey of Taiwanese live stream users, 58% of users said they used the chat, and chat room design influenced the intent to use the live stream service (Ho & Yang, 2015). The chat is an integral part of the Twitch experience, morphing what could be a passive video watching experience into a modern participatory, social occasion.

Twitch exists only because of its community. Streamers form the base; they are the ones who produce the content (live streams), displayed on their ‘channel’. Viewers support them by tuning in and watching their streams. Revenue for the site comes from ad space and ‘subscriptions’. Subscribing to a streamer allows a user to voluntarily pay (certain amount on a monthly basis) a streamer they enjoy. In return, the subscribing user can benefit from things like ‘subscriber only mode’, when only subscribers are allowed to use the chat, and having unique emoticons available to them (Hamilton et al., 2014). The streamer backbone of Twitch becomes apparent in a 2016
survey which showed that 45.8% of respondents chose ‘Specific twitchers’ as a reason to watch Twitch, followed by ‘Entertainment based on owned game’ at 29.1% (Gandolfi, 2016).

Figure 1. Twitch streamer ‘wintergaming’ plays Starcraft 2 with a webcam overlay of his face. Users participate in the chat on the right of the screen.

2.2. UGT and Twitch

Knowing what functions Twitch offers is the foundation to understanding why people use it. The interest of games appears to be the initial draw, as respondents to a 2014 survey most frequently cited their desire “to learn something about a particular game” as the reason they started watching streams (Hamilton et al., 2014). Thanks to this shared interest, Twitch forms what it known as a ‘Third Place’ for its users, an informal digital meeting place where people happily congregate to enjoy a common interest and socialize with each other (Hamilton et al., 2014). Holbert, Kwak, & Shah (2001) also state “chatrooms and multiplayer games increasingly provide the illusion of face-to-face social interaction and belonging”, further reinforcing Twitch as ‘Third Place’.

To delve deeper into the reasons for using Twitch, researchers have commonly turned toward the ‘Uses and Gratification” (UG) theory. UG is a prominent theory when dealing with personal motivations behind using media ranging from passive television watching to participatory social media use, making UG an appropriate measurement tool for Twitch as well (Sjoblom, Torhonen, Hamari, & Macey, 2017). In UG theory, users are considered to be a conscious ‘fulfillment seeker’. That is, users knowingly and self-reflectively seek out media that will satisfy their personal motivations according to certain guidelines as demonstrated in Table 1 (Sjoblom & Hamari, 2016).

Sjoblom and Hamari (2016) compared the UG need types with Twitch users’ Hours Watched, Streamers Watched, Streamers Followed, and Subscription; their results showed a positive correlation for all need types excepting Personal Integration with ‘Hours Watched’ and ‘Streamers Watched’. Personal Integrative was only positively correlated to ‘Streamers Followed’, and the only need that was negatively correlated to ‘Hours Watched’ and ‘Streamers Watched’, demonstrating a unique characteristic (Sjoblom & Hamari, 2016).
Table 1. The five aspects of UG that were tested by Sjoblom & Hamari, 2016

<table>
<thead>
<tr>
<th>Need type</th>
<th>Description</th>
<th>Media examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive</td>
<td>Acquiring information, knowledge, comprehension</td>
<td>Television (news), video (&quot;How to install Ceramic Tile&quot;), movies (documentaries or films based on history e.g., The Other Boleyn Girl)</td>
</tr>
<tr>
<td>Affective</td>
<td>Emotional, pleasant, or aesthetic experience movies, television (strums, soap operas)</td>
<td></td>
</tr>
<tr>
<td>Personal</td>
<td>Enhancing credibility, confidence, and status Video (&quot;Speaking With Conviction&quot;)</td>
<td></td>
</tr>
<tr>
<td>Social</td>
<td>Enhancing connections with family, friends, Internet (email, chat rooms, Listserv, BM) and so forth</td>
<td></td>
</tr>
<tr>
<td>Tension release</td>
<td>Escape and diversion</td>
<td>Television, movies, video, radio, Internet</td>
</tr>
</tbody>
</table>

Overall, Twitch displays the ability to satisfy UG motivations with “social interaction, learning and entertainment” as “important aspects of spectatorship” (Sjoblom & Hamari, 2016). Twitch UG satisfaction was further investigated by Sjoblom et al. (2017) where the unique characteristics of game genre and stream type were related to UG needs, revealing that the unique characteristics of game genres and stream types watched were correlated with different motivations. Even more significant was the fact that stream type had more impact on affecting user gratifications than did game genre (Sjoblom et al., 2017). This reinforces the notion of Twitch being a “Third Place”, allowing users to informally seek use gratification by dividing themselves into groups/streams based on interest in not just game genre, but also in the stream type or ‘atmosphere’ that has been developed by the streamer (Hamilton et al., 2014).

It is clear that all UG needs contribute in some way to the way users watch streams, and that these UG needs also affect users’ choice of stream (Sjoblom & Hamari, 2016; Sjoblom et al., 2017). Beyond how stream type and content (such as genre) affect the UG needs for Twitch, there is the effect of stream size and possibly other factors playing into users’ stream choices.

In addition to the above-mentioned stream attributes, the size of the stream also plays a role in the interactivity and social usability of Twitch. Hamilton et al. (2014) make specific note of the effect of stream size: “as streams scale up, information overload renders chat unreadable…… participants become frustrated with the difficulty of interacting in these streams”, possibly leading to users seeking smaller, less populated streams. This is directly backed by Sjoblom et al.’s (2017) finding that Social Integrative and Personal Integrative UG needs are a part of Twitch users’ motivation. The interaction between streamers and their audience is affected by the size of the stream. Streamers generally reported that they could interact effectively with 100-150 viewers by reading comments on stream, taking game play suggestions, and so on; beyond that number interaction was often unmanageable and the social aspects, even between users, was diminished (Hamilton et al. 2014).

A curious finding about the relation between stream type and game genre emerges when looking at the size of communities as a whole. Churchill and Xu (2016) found in a social network analysis of streamers that the Casual Streaming culture is the largest, followed by the Speed-Running culture, and finally the Competitive culture. The claim is that this ranking comes from the technical difficulty of games, Casual being the easiest to understand and Competitive being the most difficult and technically intensive (Churchill & Xu, 2016). However, Sjoblom et al. (2017) show in their UG study that when factoring game genre and stream/culture type together, Competitive stream types offer more, and more significant UG need satisfaction than Speed-running stream types, indicating that perhaps the Competitive community should be larger
(Sjoblom et al., 2017). These two types provide similar levels of motivation when only considering stream type. Taking this anomaly into account, extra consideration should be given to what streams users choose and the multitude of factors at play.

During their social network analysis of Twitch, Churchill and Xu (2016) visualized communities by specific games and community/culture types (Competitive, Casual, and Speedrun). Strong connections between game franchises were found, along with fairly well integrated cultures. During this study, however, only large, popular streamers who followed other streamers were considered. The inclusion of smaller streams of varying types might have an impact in subsequent studies. This is because, as mentioned by Hamilton et al. (2014), the size of streams/streamers impact the quality of socialization and therefore possibly users’ habits.

Twitch itself provides a platform for every game (e.g. League of Legends, Warframe), every genre (e.g. Action, Real-time Strategy), and every stream/culture type (e.g. Speed Running, Casual). Users then filter this multitude of streams based on their personal interest, style, and motivations such as entertainment based on owned games, evaluation of non-owned games, or the ability to comment on the game (Gandolfi, 2016).

3. Method

This study aims to construct an exploratory social network analysis of Twitch, expanding on previous efforts with a focus on how communities are possibly affected by UG afforded by different types/cultures of streams, as well as game genre and stream size. Recently introduced by Twitch is browsing by ‘Communities’ that includes sections like ‘VarietyStreaming’, ‘Positivity’, and ‘eSports’, allowing another level of social network comparison. There is also the availability of coding streams based on their culture/type. This represents a move away from the traditional classification of streams by game or genre, and presents the opportunity for a fresh look at community structure.

3.1. Research Questions

RQ1) How does Twitch’s social network look considering stream size, but before partitioning based on stream attributes?

RQ2) How does Twitch’s social network look when based on streams’ Community Membership Type?

RQ3) How does Twitch’s social network look when based on stream Type (e.g. Competitive, Casual, Speedrunning, Social, or Learn)?

RQ4) How does Twitch’s social network look when based on game genre?

RQ5) How does Twitch’s social network look when based on a modularity algorithm?

RQ6) What can detailed views of single nodes reveal about the network?

RQ7) Which streams are key in terms of degree centrality, and how does this compare to their follower size and other traits?
3.2. Collection/Sampling

A total of 175 streams were selected. First, 75 streams were selected from Twitch’s Communities. A specific community was selected and automatically sorted by live-viewers. A stream must have had more than one live-viewer and more than 100 followers to have been selected. Streams were selected systematically by amount of live viewers; approximately 10% of the Community’s streams were selected each time. Additionally, 100 streams were selected from ‘TwitchSwitch’, a website that generates a link to a random Twitch stream, with at least one live-viewer.

Streams were required to have at least 100 followers to be more likely to guarantee links between nodes. Streams of all sizes were desired for diversity. Smaller streams are valued for their representation of increased streamer-follower interaction, and larger streams for generating edges as popular media (Hamilton et al. 2014).

3.3. Nodes

Each node represents a stream. Each node has certain attributes to help group them. Existing attributes are:

- Number of Followers (stream size)
- Game Played

Other attributes had to be assigned by the researcher. These are:

- Type of Stream (casual – streamers play a game, competitive – streamers play competitively against other players [in leagues or for ranks], speed running – players race to complete a game as fast as possible, social – streamers are doing some other activity, learn – streamers note on their stream that they try to teach a game)

- Genre (the genre of the game being played, ie First Person Shooter (FPS), Multiplayer Online Battle Area (MOBA), Massively Multiplayer Online (MMO), In-Real-Life (irl), Sports, Horror, Strategy, Card, Sandbox, Adventure, Fighting, Role Playing Game (RPG), etc)

- Community Membership Type (General – belongs only to social communities, Competitive – belongs only to competitive/game communities, Mixed – belongs to at least one social and at least one competitive/game community, None – belongs to no communities. Streamers choose which communities they want to join. Streams can belong to up to three communities at a time. Due to the infinite amount of communities available, streamers were coded into one of the four classifications.)

3.4. Edges

Edges are determined by actual users. That is, followers that streams share in common. The weight of the edge is determined by the amount of followers in common.

3.5. Programs/Resources

Twitch.tv, TwitchTools.com, TwitchSwitch.tv, Gephi, Python, Microsoft Excel
3.6. Process

Once node information from Twitch was recorded into an Excel sheet, the stream’s ID was given to TwitchTools.com to retrieve a list of its followers in CSV format. A Python script was used to compare stream follower files. The followers were sorted by ‘Notification’ and then ‘Follow Date’, only the first 5000 followers were compared. ‘Notification’ means the user receives a notification when that stream comes online, ‘Follow Date’ is the date when the user started to follow the streamer. These were chosen to best represent habitual stream watchers. The amount of followers in common between streamers was used to determine the weight of edges between streams. This node and edge information was fed into and analyzed with Gephi.

When analyzing, edges with a weight of less than two were filtered out. This is to slightly increase the meaningfulness of the graph, as there were quite a few edges with a weight of only one. This caused eight of the nodes to be excluded from the final graph, leaving 167 nodes. The layout algorithm used was Force Atlas 2, and edges are undirected. Edge weight plays a role in determining the shape of the network.

4. Results

RQ1) How does Twitch’s social network look considering stream size, but before partitioning based on stream attributes?

![Figure 2. Nodes are presented simply, without any groupings. The size of the nodes is relative to the amount of followers they have.](image)

Overall, the social graph is quite compact with few outliers. Nearly every node has a connection to more than one other node. Analysis provides an Average Degree of 38.299, an Average Path Length of 1.87, Diameter of 4, and Density of .221. This data and Figure 2 show a quite tightly knit community despite the number of nodes. Notable is that some of the largest nodes are actually near the outside of the graph, when generally the largest nodes would be expected to be more central. In fact, many of the central nodes small and medium sized. This shows the significant factor of interactivity that smaller streams can deliver above larger streams (Hamilton
et al. 2014). Due to the data collection method, it can be hypothesized that many people visit and follow these larger streams, but they do not visit them as habitually (no ‘Notification’) as smaller streams.

RQ2) How does Twitch’s social network look when based on streams’ Community Membership Type?

Introduced as a beta feature in February 2017, Twitch’s ‘Communities’ are “self-organizing, public groups centered around specific activities”; these communities act as a way for users to more easily discover streams that are more pertinent to their interests (Karsevar, 2017). Figure 3 reveals that ‘General’ community streams are strewn throughout the graph, likewise with ‘None’. General (more social focused) streams form the ‘core’ of the network. This comes as no surprise seeing as relaxed and socially open streams such as these represent several of the Uses and Gratification traits like Social and Personal Integration. ‘Competitive’ community streams are a bit scattered, but show a couple of clear groupings, indicating users’ interest in certain gaming mechanics, and then branching off to social streams as well. Interestingly, ‘Mixed’ community streams are a bit scattered as well, but are often found clumped next to the two Competitive groupings. It’s possible these mixed streams form a type of bridge as followers express their intents of satisfying different UG needs while moving between more social and competitive communities.

![Figure 3](image)

Figure 3. Nodes (streams) are partitioned by Community Membership Type, and sized depending on the number of overall followers.

RQ3) How does Twitch’s social network look when based on stream Type (eg Competitive, Casual, Speedrunning, Social, or Learn)?
This classification for streams is based on previous research by Sjoblom et al. (2017). Streams were coded by the researcher based on how the streamer described their stream, or lacking a self-description, a stream type was assigned depending on the style of game-play or culture. For instance, streams which did not display a particular game were assigned ‘Social’, and streams that mentioned competitive play were assigned ‘Competitive’.

Immediately obvious in Figure 4 is that groupings are much clearer by stream Type than they are by Community Membership (Figure 3). This may suggest a couple of things. First, Community Membership is more complicated than the coding allowed for. Second, Community Membership (which is user-decided) is a more esoteric [social] marker than an appropriate stream identifier. More research is needed.

As compared to Figure 3 (General community being moderately related to ‘Casual’ type streams), Figure 4 again shows that Casual streams to make up the bulk and core of the network. Casual type streams had the highest degrees of centrality, acting as a bond between more focused cultures. ‘Speedrunning’, ‘Competitive’, and ‘Social Streams’ were quite tight-knit. This suggests that users congregate to a specific culture/interest, and then branch out to satisfy other needs like Tension use gratifications (or vice versa). No community really has a monopoly on large or small streams, signifying that users still can attain Social or Personal Integrative gratification across cultures and stream styles. Speedrunning does however have the largest concentration of small streams, possibly adding credence to Churchill and Xu’s (2016) study that stated the small attraction of highly technical streams compared to less technical general streams.

RQ4) How does Twitch’s social network look when based on game genre?

Figure 5 shows the social network based on game genres, listing the top ten. Heavily prevalent is the First Person Shooter (FPS) genre, likely due in part to the current popularity of PUBG. FPS games like PUBG and Counter-Strike are scattered across the Competitive and Casual regions.
(Figure 4) likely because of their ability to satisfy multiple UG stream needs; they can be highly technical (Cognitive UG), provide in-game socialization (Social/Affective UG), and can be cinematic (Tension UG). Adventure games make up almost the entire Speedrunning type streams due to their single-player nature, but are still found elsewhere throughout the network. Multiplayer Online Battle Arena (MOBA) games lend themselves well to the Competitive streams thanks to their top-down views of small scale battles, but are still found throughout the network. In-Real-Life (IRL) is relegated to Social streams, as Social streams are defined as doing something other than a game. The IRL designation is given unless the streamer lists a specific activity (like music or art), so general IRL activities (chatting) greatly trump specific non-game activities at this time.

In general, game genres don’t have too many boundaries when it comes to Community Membership or even stream Type. Streamers are simply using the games as a tool to interact with their community, a common bonding type. Sometimes a specific tool (Adventure or MOBA) helps reach a certain type of follower, but the social heart of Twitch allows genres to span the entire network.

RQ5) How does Twitch’s social network look when based on a modularity algorithm?

Running the social network through a modularity algorithm produces results close to consistent with assigned variables like stream type. In this algorithm, the network was divided into 6 sections with no knowledge about the nodes’ attributes stream communities, types, genres, etc. Figure 6 is the result of automatically generated ‘communities’ or sections of the network. Without assigning meaning to the nodes, parts of the network still match together. These ‘organic’ communities provide the insight of the mixing and dividing of users’ interests. Again, competitive and speedrunning cultures stand out clearly. As communities from Figure 3 were mixed at the top, they are here as well. At the bottom, there is the collusion of General/Mixed communities and casual/social type streams.
RQ6) What can detailed views of single nodes reveal about the network?

By focusing on specific nodes, ‘case studies’ about the network and its nodes are possible. The stream with the most followers, ‘dreamhackcs’, is a ‘Competitive’ type stream, and belongs to the ‘Competitive’ community. It’s connected to nearly all of the other Competitive streams as kind of a key figure for that stream type (Figure 6). Despite its size and importance to Competitive streams, it’s not nearly as central as some of the other Casual type streams, or even compared to the other Competitive streams. This is likely due to the fact that ‘dreamhackcs’ is a special use stream that live-broadcasts large competitive events (Dreamhack Tournament) and reruns, but is not a typical stream run by a single streamer.
‘reabs’ is a Speedrun type node (Figure 7). Despite being close to the outskirts of the network, it is still connected to almost every other Speedrun node, showing a tight-knit community. ‘reabs’ is also connected to the majority of Casual nodes, despite being a purely Competitive Community Member. Like Figure 3, this might suggest that Community Membership might not have too big of an impact in what users stay within. Instead, users may allow their UG needs to lead the way and search out a stream like ‘reabs’ for certain gratifications, and the Casual streams for other gratifications.

RQ7) Which streams are key in terms of degree centrality, and how does this compare to their follower size and other traits?

Initially, the most noticeable trait of the centrality heat map is that the largest nodes/streams do not represent the most connective nodes (Figure 9). When compared to Figure 4, the most central nodes are Casual. As Twitch satisfies UG needs, these nodes seem to provide a great even ground for viewers, a kind of comfortable starting point that attends to many needs like Cognitive gameplay and Affective streamer interaction. Also notable is the fact that highly central nodes are often medium sized. Too small, and a stream doesn’t provide enough interaction. Too large, and a stream overwhelms the user with interaction.

Competitive nodes are often also quite central. The base reason for Twitch is to act as a platform for streaming game content, and it’s clear users still move between social/Casual streams and more serious Competitive streams. Speedrunning is quite mixed in terms of centrality. Speedrunning can be quite a specific interest, and this is represented in the centrality map. It’s likely the furthest streams have a specific focus, drawing a dedicated crowd. And yet, as in Figure 7, this crowd still mingles with very central/Casual streams.
5. CONCLUSION / DISCUSSION

Compared to traditional media, social live-streaming sites like Twitch are relatively new. Twitch itself is often in a state of change as trends and streamers come and go, updates are applied, technologies become more accessible, and the internet culture fluctuates. Even Twitch’s Communities section had been added within one year of this study. This study was performed to impart a beginning understanding of the nature of Twitch’s social network onto the audience.

This study found that Community Membership didn’t segregate nodes into sections as well as stream types. This suggests that Communities may be kind of vague, being that they are streamer-defined (ie streamers can create/join any community), or, the method used to define them wasn’t selective enough. There are an infinite amount of communities, which may make it difficult for streams to define themselves, or for users to find an appropriate stream that includes all/most of the streams that would satisfy their UG needs. Well defined communities, however, could be extremely beneficial.

Therefore, this researcher would like to suggest an improved way of classifying streams. Communities (especially large ones) as they are, often don’t correlate quite well with stream type or the game played. A way of breaking down and defining streams may help users find the streams they desire. Since Communities are already streamer-defined, it’s not impossible that Twitch could provide a streamer or user-driven filter system. Streamers could add certain filters to their stream regarding the usual peak viewer count, usual game and game genre, type of stream, and so on. Likewise, users could also be given the power to help define streams/filters.

The method of connecting streams, by followers-in-common and by Notification, was successful. It helped avoid large streams dominating the network by size alone. Streams of all sizes were allowed shine in the network based on habitual users as they do on Twitch itself. Smaller and mid-sized streams often had high centrality (especially Casual types) due to their increased interactivity (smaller streams better suited for socializing). Larger streams like ‘Dreamhackcs’ (Competitive type) helped form the hubs of their respective stream types. Frequent connections between all social circles were found, even from very focused cultures such as Speedrunning.
Twitch is indeed an open community, offering a diverse range of topic and interests, yet linked together by the social experience provided. The existence of specific communities lets users swim around within the overall network to find a stream to satisfy whatever their needs might be.

Due to the ever-evolving nature of Twitch and social sites, this research has some limitations. For instance, streams can have membership in multiple communities. For this study, this node attribute had to be codified based on their Community Memberships, not on their exact communities due to the infinite possibilities. Future research may want to consider a way in which to include this variable. Additionally, stream type and game played can change. Again, these variables were assigned based on the stream at the time of data collections. Classification is unending for continually changing social sites and evolving media. It is also very difficult to dictate the perfect genre for a game, as a game may have both RPG and Adventure elements, so the researcher must code the genre as what appears to be the primary genre. Following studies may benefit from incorporating a system where the genre of games in a stream can be better defined and compared. As this was exploratory research, future research on a specific network trend, or expansion of a specific community, genre, or stream type might be interesting.

ACKNOWLEDGEMENTS

Professors Janghyun Kim and Jongshik Lee, family, friends, 친구들, 맥주, and Rufus

REFERENCES


EXAMINING RISK COMMUNICATION OF SAMSUNG BY ANALYSING CUSTOMER REACTION ON SOCIAL MEDIA: FROM GALAXY NOTE 7 EXPLOSION TO GALAXY S8 RELEASE

Jaewon Lim and Jang Hyun Kim

Department of Interaction Science, Sungkyunkwan University, Seoul, Korea

ABSTRACT

Crisis of product defect causes recalls, which not only causes financial loss but also devastates brand reputation. This study explores customers’ response to the twice recall of the Galaxy Note 7 in Korea. The authors examined the semantic networks of consumers comments from the Samsung official social media ‘Samsung Newsroom (blog and Facebook)’ in five time periods (i.e., release, first recall, recommendation to discontinue use, second recall, and compensation program). Customers in each period showed different emotional states which were positive in the first recall but negative in the second recall. Therefore, timely recall and factual communication is needed to the product-harm crisis management, but root cause investigation should be taken to avoid further recalls.

KEYWORDS

Galaxy Note7, Product harm crisis, Recall, Semantic network analysis

1. INTRODUCTION

The crisis of the Galaxy note 7 recall occurred within a month of its release. Galaxy note 7 was unveiled on August 19, 2016. However, several Galaxy note 7 explosion cases were reported and the epidemic was spreading. Samsung diagnosed the cause of the explosion is a failure of the batteries and recalled the Note 7. Although they provided replacement phones, those began to explode as well. In the end, Samsung issued the second recall and recalled the entire line of Note 7 phones.

[1]Recalling a product can make the involved firms’ brand image worse.[2]Therefore, they should try to reduce negative impacts, one of which is to convey their official position to the public through media.

[3]Since the mid-2000s, Social media platforms increasingly tend to act as both a place for the firms and customers. Focal firms’ risk communication message can be delivered instantly and broadly. With social media, customers can voice their opinion. Samsung posted a statement in
their official social media channel, ‘Samsung Newsroom(blog and Facebook)’ during a series of explosion event. At the same time, the customers commented on each statement to voice their opinion.

As far as the authors investigated, Samsung is perhaps the only company to recall twice and has the largest recall in terms of volume. However, as customers become more active and social network service develops, product recalls may happen more frequently in the future. By assessing Samsung’s crisis management to the Galaxy note 7, this study aims to understand the successful product crisis management. Using the semantic network analysis, the authors analyzed the customers’ comments written under the Samsung official announcements in their social channel. Therefore, we proposed the following research questions:

RQ1: How customers react to Samsung’s risk communication on social media(blog and Facebook)?

RQ1-1. What are semantic traits of the blog and Facebook comments?

RQ1-2. What are common and unique words for each period?

RQ1-3. How did emotional words change over time?

2. METHOD

Figure 1 shows the overall research process.

2.1 Select data collection period

The authors estimated the change of the word by period and analyze the change of customers’ awareness about the issue. The data collection period was divided into six periods. The first data collection period is from August 11, 2016 to August 17. It covers six days from the product launch. The second data collection period is September 2, 2016, when Samsung voluntarily announced the recall. The third data collection period is from September 10, 2016 to September 20, 2016. During this period, Samsung advised their users to discontinue the use of the Galaxy Note7 and announced instructions for commencing the exchange of products. The fourth data
collection period is October 11, 2016, when Samsung officially announced the discontinuation of Galaxy Note7. The fifth data collection period is October 13, 2016 to October 27. It was time for Samsung to announce compensation program, called ‘Galaxy upgrade program’, including exchange and refund. Last period is April 19, 2017, when Galaxy s8 released. Galaxy s8 is the first and new Samsung galaxy series product after the Galaxy Note 7 explosion.

2.2 Data Collection

Customer comments were gathered from the Samsung’s official social media channel ‘Samsung Newsroom (blog & Facebook)’. We scraped all the comments manually. A total of 4,016 comments were collected during the data collection period. All of the comments are in Korean as the authors examined only the newsroom for Koreans.

Only Galaxy s8 customer comments were gathered from six news articles posted in the Naver.

2.3 Data preprocessing

Text preprocessing was carried out through Textom, an online text analysis service which is a Korea-based big data processing solution site.

The authors uploaded the collected text data to Textom and preprocessed. In this process, the main words were identified through the frequency of extracted words. Basically, it provides listing of nouns, adjectives, and verb phrases that remove special letters and postposition through morphological analysis. Words with similar meanings are grouped together into represented word. In this way, only the meaningful words were selected from the word list formed by refining and the co-occurrence frequency matrix data was produced.

2.4 Semantic Network Analysis

[5]Semantic network analysis (SNA) was conducted, which examines the relationships among words such as co-occurrence, frequency, and distance. As SNA adequately shows the significant words regarding the customers opinion and emotions about Samsung’s product crisis management, it is opted to employ for the present study.

[6]To visualize the relationship between words, we used computational text analysis tools NETDRAW from UCINET6.

Among a variety of centrality measures, betweenness centrality was selected for this study.

3. RESULTS

To visualize semantic traits of words, the authors categorized words into four groups. The words which are related to facts of issue were categorized into group ‘Fact’ with black nodes. Also, emotional words were considered for grouping. Positive words were specified ‘Positive feeling’ with blue nodes and negative words were grouped into ‘Negative feeling’ with red nodes. Finally, words about cost were categorized into ‘Cost’ with yellow nodes.
3.1 Six days from the product launch- August 11, 2016 to August 17, 2016

For the first period, 73 comments were collected. The words with the highest frequency next to ‘Samsung’ are article, opening, Galaxy Note7, good, consultation, head office, change and others.

Figure 2. Semantic network of comments on Samsung Newsroom from August 11, 2016 to 17

In Figure 2, words are related to ‘Current Galaxy users (S7, Note5)’ and ‘General expectation for Note7’. For example, words explaining current Galaxy users are product, service center, repair, and Galaxy S7. Such words as design, software, hardware, pre-order, lens and better were mentioned for general expectation for Note7. In this period, most words are categorized into group ‘Fact’ and ‘Positive feelings’.

3.2. Recall announcement - September 2, 2016

For the second period, 462 comments were collected. The words with the highest frequency next to ‘Samsung’ are product, exchange, decision, battery, refund, pre-order, recall, new product, problem, use, customer, corporation, and others.

In Figure 3, words are related to two topics, which are ‘Explosion issue and disappointed feeling’ and ‘Positive opinion for Samsung decision.’ For example, words about the first topic are explosion, tiresome, disappointment, problem, defect, bomb, return, pity, and error. For the second topic, good decision, trust, thanks, and cheer up were mentioned.

At this time, lots of words are categorized into group ‘Fact’ and ‘Positive feelings’ like the first period semantic network. Also, words with ‘Negative feelings’ were salient from this point.
3.3. From recommendation for stop using to exchanging replacement - September 10, 2016 to September 20, 2016

For the third period, 515 comments were collected. The most frequent words next to ‘Samsung’ are exchange, customer, battery, product, problem, wait, Galaxy Note7, authorized dealer, explosion, reservation, recall, purchase and other words.

In Figure 4, words described ‘Explosion issue and product problems’ and ‘Complaint for exchanging and disappointment.’ For example, the words related to explosion issue and product problems are explosion, fault, dead battery, defect, and safety. For the complaint for exchanging and disappointment, the words such as it's annoying, inconvenience, refusal and apology were mentioned.

During this period, while blue nodes with ‘Positive feeling’ decreased sharply compare to previous two periods, red nodes with ‘Negative feeling’ increased.

3.4. Second recall - October 11, 2016

For the fourth period, 1,002 comments were collected. Word ‘Exchange’ is the most frequently mentioned at this time. Refund, Samsung, Galaxy Note7, customer, purchase, use, think, product, compensation, change, wait, and cancellation charge are the highest frequency words with ‘Exchange’
Figure 4. Semantic network of comments on Samsung Newsroom from September 10, 2016 to September 20, 2016

Figure 5. Semantic network of comments on Samsung Newsroom on October 11, 2016
In Figure 5, words are demonstrating two topics, which are ‘Loss, compensation and disappointment’ and ‘Countermeasures for Galaxy Note 7.’ For example, the words regarding the first topic are cancellation, charge, damage, halt production, absurd, and compensation. For the second topic, Galaxy Note 8, Galaxy S8, iPhone, new product, and rental phone were mentioned.

At this time, lots of words are categorized into group ‘Cost’ and ‘Negative feeling’ unlike other three periods.

3.5. Compensation program notice - October 13, 2016 to October 27, 2016

For the last period, 2,022 comments were collected. The most frequent words next to ‘Samsung’ are exchange, customer, refund, compensation, opinion, Galaxy Note 7, use, change, product, program, purchase, return, Galaxy Note 8, and others.

![Figure 6. Semantic network of comments on Samsung Newsroom from October 13, 2016 to October 27, 2016](image)

In Figure 6, most words in categories ‘Fact’, ‘Negative feeling’, and ‘Cost’ are mentioned similarly. Black, red and yellow nodes are spotted sporadically. For example, words like money, installment, (smartphone) case, 30 dollars, and coupon are showing yellow nodes ‘Cost’. Dissatisfaction, trash, fool, wait, fail, discomfort, phone replacement, and pushover present red nodes ‘Negative feeling’. This network has only one blue node for ‘Positive feeling’ with word ‘good’.

3.6. Galaxy S8 release

We can see the unique word ‘Sakura Edition’. Despite having to be cautious after the battery explosion of the Galaxy Note 7, the phenomenon that the display screen is reddish immediately after the launch of the Galaxy S8 was controversial. It got a humiliating nickname of 'Sakura Edition', and consumers blamed it as 'Trash Galaxy' and kicked the tongue. In addition, it is the
first new product after the explosion, so consumers' worries about 'Safety' have increased. The words 'Note 7', 'Battery' and 'Explosion' related to Note 7 were also mentioned.

Although there are few numbers, positive words also appeared. A few consumers were satisfied and admitted with the design of the Galaxy S8 and cheered on Samsung('Fighting'

Figure 7.Semantic network of comments on Galaxy S8

4. DISCUSSION

As a result of examining the crisis response of Samsung in five periods as above, it can be seen that Samsung’s product crisis management on the Galaxy Note 7 was not so effective as desired. It was a good decision to admit the product defect and to make a swift recall. As we can see the words in the second period, people felt “trust” in Samsung’s recall. However, due to the rapid exchange of products without properly identifying the cause, replacements also exploded. This led to second recalls and customers had to do “tiresome” exchanges even twice. Customers even need to seek a smartphone to replace the Galaxy Note 7 and to use ‘iPhone’ or Galaxy S7 edge which is older than Galaxy Note 7.

When Samsung announced the first recall, it would have been possible to reduce the damage if they had enough time to properly identify the cause. It is a smart move to admit product defects in detail and voluntarily recall products as early as possible if the cause of the defect is the problem of the product itself. Without uncovering the root cause, rashness can result in the two times of recalls like Galaxy note 7. Not only timely but accurate and factual communication is needed.

Second, results show that the importance of an appropriate compensation program. After Samsung announced that it has halted sales and production of the Galaxy Note 7 worldwide, Samsung is offering a new compensation program. However, it is not much different from the existing ‘Galaxy program’, which was packaged as if it were a special reward program for Galaxy note 7 customers. So when the authors analyzed the comments, the word ‘trickery’ appeared frequently. It should be noted that anger that is worse than a product defect came from the users
feeling that the company may have deceived its customers.

Third, the performance of the device in launching a new product after the crisis must be emphasized. For the new products released after the crisis, consumers expressed their negative opinions due to the previous explosion (the ‘red display screen’ phenomenon of the Galaxy S8 also contributed to the negative comments), but at the same time it showed a positive response to the ‘performance’ and ‘design’ of the S8. Therefore, when it comes to product launches in the future, it will be more important to improve the ‘performance’ and ‘safety’ of products.

5. CONCLUSION AND LIMITATION

[7] Samsung sits at 49th place in 2017 Reputation Quotient Ratings, which ranks one of American’s 100 most visible companies. This is comparable to the 7th rank in 2016 and all three consecutive years when Samsung secured a spot in the top 10 from 2014 to 2016.

The limitation of this study is in that it explored comments from Korean Facebook and blog, although Galaxy note7 recall occurred worldwide.

REFERENCES


AUTHORS

Jaewon Lim is studying Interaction and Science in Sungkyunkwan University. Main research interests lie in social and semantic network analyses and media and communication.
INTENTIONAL BLANK
CONCEPTUALIZING AI RISK

Martin Ciupa\textsuperscript{1} and Keith Abney\textsuperscript{2}

\textsuperscript{1}CTO calvIO Inc., Webster, New York, USA
\textsuperscript{2}Senior Lecturer, CalPoly, California, USA

\textbf{ABSTRACT}

AI advances represent a great technological opportunity, but also possible perils. This paper undertakes an ethical and systematic evaluation of those risks in a pragmatic analytical form of questions, which we term ‘Conceptual AI Risk analysis’. We then look at a topical case example in an actual industrial setting and apply that methodology in outline. The case involves Deep Learning Black-Boxes and their risk issues in an environment that requires compliance with legal rules and industry best practices. We examine a technological means to attempt to solve the Black-box problem for this case, referred to as “Really Useful Machine Learning” (RUMLSM). DARPA has identified such cases as being the “Third Wave of AI.” Conclusions to its efficacy are drawn.

\textbf{KEYWORDS}

AI Risk, Deep Neural Network, Black-box Problem, Really Useful Machine Learning, RUMLSM, DARPA Third Wave of AI.

\section{1. INTRODUCTION}

A common worry about AI is that it poses an unacceptable risk to humanity (or individual humans) in some way. An extensive literature has begun to emerge about various aspects of Artificial Intelligence (AI) risk, much of it focused on existential risk from Artificial Generic Intelligence (AGI). But AI poses other risks, from how driverless cars solve the ‘trolley problem’, to whether autonomous military robots attack only legitimate targets, to trust in the safety of AI/Robotics in industrial and commercial settings. More generally, the discussion of risks from AI has paid insufficient attention to the nature of risk itself, as well as how decisions about the acceptability of the risks of AI compare to worries about convergent technologies. For example, global debates about autonomous weapons have focused predominantly on robotics, but AI also can be weaponized. For instance, in robotics serious concern exists over a possible lack of “meaningful human control” \cite{1}. Missing is a similar concern for autonomous AI-controlled cyber attacks that would lack the very same control \cite{2}. The Vice Chairman of the Joint Chiefs of Staff understands, saying, “In the [Defense] Department, we build machines and we test them until they break. You can’t do that with an artificial intelligence, deep learning piece of software. We’re going to have to figure out how to get the software to tell us what it’s learned” \cite{3}. Significant aspects of AI risk can thus be characterized as a “black-box” problem.
So, how best to understand the risks of AI, judge them (un)acceptable, and then apply our insights on risk to determine what policies to pursue?

2. DEFINING RISK, AND HOW TO THINK ABOUT IT

So, AI poses many different types of risk – but what exactly is risk? Andrew Maynard [4] suggests that we start with the idea of “value.” If innovation is defined as creating value that someone is willing to pay for, then he suggests risk as a threat to value, and not just in the ways value is usually thought of when assessing risk, such as health, the environment or financial gain/loss, but also well-being, environmental sustainability, deeply held beliefs, even a sense of cultural or personal identity. Risk is thus a potential, but not certain, harm.

Extending a schema based on previous work [5], the major factors in determining ‘acceptable risk’ in AI will include (but are not limited to):

2.1 Acceptable-Risk Factor: Consent

Consent: Is the risk voluntarily endured, or not? For instance, secondhand smoke is generally more objectionable than firsthand, because the passive smoker did not consent to the risk, even if the objective risk is smaller. Will those who are at risk from AI reasonably give consent? When would it be appropriate to deploy or use AI without the meaningful consent of those affected? Would non-voluntariness (in which the affected party is unaware of the risk/cannot consent) be morally different from involuntariness (in which the affected party is aware of the risk and does not consent)? [6]

2.2 Acceptable-Risk Factor: Informed Consent

Even if AIs only have a ‘slave morality’ in which they always follow orders [7], and citizens consent to their use (through, say, political means), that still leaves unanswered whether the risk (of malfunction, unintended consequences, or other error) to unintended parties is morally permissible. After all, even if widespread consent is in some sense possible, it is completely unrealistic to believe that all humans affected by AI could give informed consent to their use. So, does the morality of consent require adequate knowledge of what is being consented to?

Informed consent: Are those who undergo the risk voluntarily fully aware of the true nature of the risk? Or would such knowledge undermine their efficacy in fulfilling their (risky) roles? Or are there other reasons for preferring ignorance? Thus, will all those at risk from AI know that they are at risk? If not, do those who know have an obligation to inform others of the risks? What about foreseeable but unknown risks—how should they (the ‘known unknowns’) be handled? Could informing people that they are at risk ever be unethical, even akin to terrorism?

2.3 Acceptable-Risk Factor: The Affected Population

Even if consent or informed consent do not appear to be morally required with respect to some AI, we may continue to focus on the affected population as another factor in determining acceptable risk:
Affected population: Who is at risk—is it merely groups that are particularly susceptible or innocent, or those who broadly understand that their role is risky, even if they do not know the particulars of the risk? For example, in military operations civilians and other noncombatants are usually seen as not morally required to endure the same sorts of risks as military personnel, even (or especially) when the risk is involuntary or non-voluntary.

2.4 Acceptable-Risk Factor: Step risk versus State risk

A state risk is the risk of being in a certain state, and the total amount of risk to the system is a direct function of the time spent in the state. Thus, state risk is time-dependent; total risk depends (usually linearly) on the time spent in the state. So, for us living on the surface of the Earth, the risk of death by asteroid strike is a state risk (it increases the longer we’re here).

Step risk, on the other hand, is a discrete risk of taking the next step in some series or undergoing some transition; once the transition is complete, the risk vanishes. In general, step risk is not time-dependent, so the amount of time spent on step matters little (or not at all). [8] Crossing a minefield is usually a step risk – the risk is the same whether you cross it in 1 minute or 10 minutes. The development of AGI poses a clear step risk; but, if there is a ‘fast takeoff,’ the state risk may be negligible.

2.5 Acceptable-Risk Factors: Seriousness and Probability

We thereby come to the two most basic facets of risk assessment, seriousness, and probability: how bad would the harm be, and how likely is it to happen?

Seriousness: A risk of death or serious physical (or psychological) harm is understandably seen differently than the risk of a scratch or a temporary power failure or slight monetary costs. But the attempt to make serious risks nonexistent may turn out to be prohibitively expensive. What (if any) serious risks from AIs are acceptable—and to whom: users, nonusers, the environment, or the AI itself?

Probability: This is sometimes conflated with seriousness but is intellectually quite distinct. The seriousness of the risk of a 10-km asteroid hitting Earth is quite high (possible human extinction), but the probability is reassuringly low (though not zero, as perhaps the dinosaurs discovered). What is the probability of harm from AIs? How much certainty can we have in estimating this probability? What probability of serious harm is acceptable? What probability of moderate harm is acceptable? What probability of mild harm is acceptable?

2.6 Acceptable-Risk Factors: Who Determines Acceptable Risk?

In various other social contexts, all of the following have been defended as proper methods for determining that a risk is unacceptable [9]:

Good faith subjective standard: It is up to each individual as to whether an unacceptable risk exists. That would involve questions such as the following: Can the designers or users of AI be trusted to make wise choices about (un)acceptable risk? The idiosyncrasies of human risk aversion may make this standard impossible to defend, as well as the problem of involuntary/non-voluntary risk borne by nonusers.
The reasonable-person standard: An unacceptable risk is simply what a fair, informed member of a relevant community believes to be an unacceptable risk. Can we substitute a professional code or some other basis for what a ‘reasonable person’ would think for the difficult-to-foresee vagaries of conditions in the rapidly emerging AI field, and the subjective judgment of its practitioners and users? Or what kind of judgment would we expect an autonomous AI to have—would we trust it to accurately determine and act upon the assessed risk? If not, then can AI never be deployed without teleoperators—like military robots, should we always demand a human in the loop? But even a ‘kill switch’ that enabled autonomous operation until a human doing remote surveillance determined something had gone wrong would still leave unsolved the first-generation problem.

Objective standard: An unacceptable risk requires evidence and/or expert testimony as to the reality of (and unacceptability of) the risk. But there remains the first-generation problem: how do we understand that something is an unacceptable risk unless some first generation has already endured and suffered from it? How else could we obtain convincing objective evidence?

2.7 Acceptable-Risk Factors: The Wild Card: Existential Risk?

Plausibly, a requirement for extensive, variegated, realistic, and exhaustive pre-deployment testing of AIs in virtual environments before they are used in actual human interactions could render many AI risks acceptable under the previous criteria. But one AI risk may remain unacceptable even with the most rigorous pre-deployment testing. An existential risk refers to a risk that, should it come to pass, would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential. Existential disasters would end human civilization for all time to come. For utilitarians, existential risks are terribly important: doing what we can to mitigate even a small chance that humanity comes to an end may well be worth almost any cost. And for deontologists, the idea that ‘one always has a moral obligation never to allow the extinction of all creatures capable of moral obligation’ is at least a plausible prima facie (and perhaps absolute) duty; such a survival principle appears required for any viable ethics [10]. If there is even a tiny risk that developing AGI would pose an existential risk, this ‘Extinction Principle’ may well imply that we have a duty to stop it.

3. SPECIFIC CASE STUDY, POSSIBLE SOLUTION, AND RISK ANALYSIS

3.1 Specific Case Study

As noted above, a well-known problem of AI risk is applying Deep Learning Neural Networks, specifically the black-box problem. This occurs when the Neural Network has a large set of training data; the consequential connectivity arrangements are such that for a given input, an output is provided to a reinforced goal of the system – typically prediction, decision or identification. These bottom-up systems can learn well, especially in tight domains, and in them can surpass human performance – but, without any means of validation, or clear compliance to regulations. This renders the AI potentially untrustworthy, a serious deployment risk that may be deemed unacceptable. A human can usually provide a top-down rationalization of their behavior, responding to “why did you do that” questions. But a Deep Learning system cannot easily answer such queries. It is not rule-based and cannot easily track its “reasoning.” [11]
We propose a possible solution. Our case study is a Deep Learning system applied to the Path Planning of a Robot Arm, in which vials of biohazardous materials are to be moved from point A to point B in an optimum path. This path is constrained by parameters such as speed, power-usage, minimization of actuator acceleration and deceleration (that causes wear of the actuators) and collision avoidance. See Fig 1. Keep in mind that cost of production, as well as quality/safety, are value factors to be balanced in this manufacturing example. And the use of AI Robots in this case example is a very real-world example of potential benefit and ethical concerns.

A key part of the solution proposed is an “Extractor” which will build a rationalization of the Deep Learning System into a Rule-Based Decision tree that can be validated against risk analysis/compliance needs, i.e., answering questions related to risks. This is depicted in Fig 2.

### 3.2 Possible Solution: Extraction of Heuristics from Deep Learning Neural Net

The means by which Expert Heuristics are extracted from the Deep Learning Neural Networks has been studied by other teams [12], [13] and [14]. The specific means by which we propose to do so in RUMLSM is an innovative patent pending process [15], [16] and [17]. Expert Heuristic/Rule extraction can be defined as "...given a trained neural network and the data on which it was trained, produce a description of the network's hypothesis that is comprehensible yet closely approximates the network's predictive behavior." Such extraction algorithms are useful for experts to verify and cross-check Neural Network systems. Earlier this year, John Lauchbury, director of DARPA’s Information Innovation Office said, “There’s been a lot of hype and bluster about AI.” They published their view of AI into “Three Waves,” to explain what AI can do, what AI can’t do, and where AI is headed. See Fig 3 [18]. We consider the example we have outlined above falls into this “Third Wave of AI.”

![Figure 1: Deep Learning System Applied to Robot Arm Path Planning (Source: CalvIO Inc)](image-url)
3.3 AI Risk Conceptual Analysis applied to the Case Example

1/ Acceptable-Risk Factor: Consent
The use of a robot (in a protective clean room cell) reduces the need for human operator exposure to Biohazards.

2/ Acceptable-Risk Factor: Informed Consent
However, if the AI directing the robot causes breaches of the clean/safe room (e.g., collisions with the cell walls), then what was thought safe might not be. It is necessary to test any black-box defined behavior will not violate these rules.
3/ Acceptable-Risk Factor: The Affected Population

The affected population might not be limited to the factory; conceivably, an extended exposure could cause health and safety threats to those outside, or violations of FDA regulations, etc. Again, the black-box must be validated against the regulations/laws applicable to the domain.

4/ Acceptable-Risk Factor: Step risk versus State risk

Both state and step risks need to be exposed through the black-box rule extraction process.

5/ Acceptable-Risk Factors: Seriousness and Probability

The seriousness of a biohazard breach can be evaluated in principle, but the probability may need testing. This should be done in simulation (to avoid physical exposure) as well as an assessment of the black-box extracted rule-set.

6/ Acceptable-Risk Factors: Who Determines Acceptable Risk?

There are industry bodies that set standards (e.g., GAMP5) as well as government entities that set regulations in this case example (e.g., US FDA).

7/ Acceptable-Risk Factors: The Wild Card: Existential Risk?

If the biohazard agent was severe enough, as might be possible with nuclear materials and/or live chemical/biological agents, then the impacts could be existential, if the AI goes “rogue.” The severity relates properly to steps 3 and 5 above. The risk of ‘going rogue’ is conceivable, but presumably with complete AI automation of the industrial facility, and absent proper safeguards against hacking or rampancy. A solution may involve a software system over-riding ethical kernel that ensures “no harm.”

4. CONCLUSIONS

We reviewed the concept of AI Risk and picked a well-known one, i.e., the AI Neural Network Black-box problem. We proceeded to outline a means of structuring a dialog of these “Conceptual AI Risks.”

We provided a near-term AI/Robotics case example (Smart Robotics Path Planning for a Pick and Place application for Biohazardous material) and applied the dialog to it; we think the result is an actual beneficial one for highlighting the AI risk concerns and start the process of handling them objectively.

The case example method applies a meta-level/hybrid AI system to “extract” heuristics from the neural network black-boxes (applying a top-down AI system on a bottom-up AI system). The system is based on technology called Really Useful Machine Learning (RUML®). The approach taken is an example of DARPA’s Third Wave of AI. In a sense this process can be considered as a Cybernetic Self-Regulation (extraction of a rationalized model for control).

As such, we believe the resulting techno-philosophy methodology to be a potentially useful early step in the building of tools for conceptualizing and assessing acceptable AI Risk. Further work is needed to develop these concepts, and trial them in real-world applications.

REFERENCES


AUTHORS

Martin Ciupa

Martin Ciupa is the CTO of calvIO Inc., a company (associated with the Calvary Robotics group of companies) focused on simplifying the cybernetic interaction between man and machine in the industrial setting. Martin has had a career in both technology, general management and commercial roles at senior levels in North America, Europe and Asia. He has an academic background in Physics and Cybernetics. He has applied AI and Machine learning systems to applications in decision support for Telco, Manufacturing and Financial services sectors and published technical articles in Software, Robotics, AI and related disciplines.

Keith Abney

Keith Abney is senior fellow in the Ethics + Emerging Sciences Group and senior lecturer in the Philosophy Department of California Polytechnic State University - San Luis Obispo. He is co-editor of Robot Ethics (MIT Press) and the newly published Robot Ethics 2.0 (Oxford UP).
INTENTIONAL BLANK
CNN FEATURES ARE ALSO GREAT AT UNSUPERVISED CLASSIFICATION

Joris Guérin, Olivier Gibaru, Stéphane Thiery and Eric Nyiri

Laboratoire des Sciences de l'Information et des Systèmes (CNRS UMR 7296)
Arts et Métiers ParisTech, Lille, France

ABSTRACT

This paper aims at providing insight on the transferability of deep CNN features to unsupervised problems. We study the impact of different pretrained CNN feature extractors on the problem of image set clustering for object classification as well as fine-grained classification. We propose a rather straightforward pipeline combining deep-feature extraction using a CNN pretrained on ImageNet and a classic clustering algorithm to classify sets of images. This approach is compared to state-of-the-art algorithms in image-clustering and provides better results. These results strengthen the belief that supervised training of deep CNN on large datasets, with a large variability of classes, extracts better features than most carefully designed engineering approaches, even for unsupervised tasks. We also validate our approach on a robotic application, consisting in sorting and storing objects smartly based on clustering.

KEYWORDS

Transfer learning, Image clustering, Robotics application

1. INTRODUCTION

In a close future, it is likely to see industrial robots performing tasks requiring to make complex decisions. In this perspective, we have developed an automatic sorting and storing application (see section 1.1.2) consisting in clustering images based on semantic content and storing the objects in boxes accordingly using a serial robot (https://youtu.be/NpZIwY3H-gE). This application can have various uses in shopfloors (workspaces can be organized before the workday, unsorted goods can be sorted before packaging, ...), which motivated this study of image-set clustering.

As of today, deep convolutional neural networks (CNN) [1] are the method of choice for supervised image classification. Since [2] demonstrated astounding results on ImageNet, all other methods have rapidly been abandoned for ILSVRC [3]. As suggested by [4], performances of CNN are highly correlated to the amount of labeled training data available. Nevertheless, even when few labels are available, several recent papers [5, 6, 7] have shown that CNN can still outperform any other approach by transferring knowledge learned on large datasets. In particular, [5] has shown that extracting features from OverFeat [8] pretrained on ImageNet, and training a simple Support Vector Machine (SVM) [9] classifier on these features to fit the new dataset.
provides better results than many more complex approaches for supervised classification. These results demonstrate that a CNN, trained on a large and versatile dataset, learns information about object characteristics that is generic enough to transfer to objects that are not in the original dataset.

While developing the automatic robotic sorting and storing application, we needed to classify sets of images based on their content, in an unsupervised way. Multiple papers introduced methods to solve unsupervised object classification from sets of images (see section 1.1.1), producing relatively good results. However, we wanted to know if the information from a large and versatile dataset, stored in the weights of a CNN, could be used straightforwardly to outperform state-of-the-art algorithms at unsupervised imageset classification. The goal of this paper is to answer the following question: How good are features extracted with a CNN pretrained on a large dataset, for unsupervised image classification tasks? To do so, we use a similar approach to [5], consisting in applying classic clustering methods to features extracted from the final layers of a CNN (see section 2 for more details) and comparing it with state-of-the-art image set clustering algorithms [10, 11] on several public datasets.

The intuition behind such approach for unsupervised object classification is that, as it works with SVM [5], the CNN must project data in a feature space where they are somehow linearly separable. Thus, simple clustering algorithms such as K-means might be working well. However, this study is interesting as the performance of such simple clustering algorithms often depends on the notion of distance between points, on which we remain uncertain.

1.1. Previous work

1.1.1. Image-set clustering

Given a set of unlabeled images, the image-set clustering problem consists in finding subsets of images based on their content: two images representing the same object should be clustered together and separated from images representing other objects. Figure 1 illustrates the expected output from an image-set clustering algorithm in the case of our robotics application. This problem should not be confused with image segmentation [12], which is also sometimes called image clustering.

Image-set clustering has been widely studied for two decades. It has applications for searching large image database [13, 14, 15], concept discovery in images [16], storyline reconstruction [17], medical images classification [18], ... The first successful methods focused on feature selection and used sophisticated algorithms to deal with complex features. For instance, [19] represents images by Gaussian Mixture Models fitted to the pixels and clusters the set using the Information Bottleneck (IB) method [20]. [21] uses features resulting from image joint segmentation and sequential IB for clustering. [11] uses Bags of Features with local representations (SIFT, SURF, ...) and defines commonality measures used for agglomerative clustering. Recently, image-set clustering algorithms have shifted towards using deep features. [10] uses deep auto-encoders combined with ensemble clustering to generate feature representations suitable for clustering. [22, 18] learns jointly the clusters and the representation using alternating optimization [23].
1.1.2. Robotic automatic sorting application

Together with showing that deep features + simple clustering outperforms other approaches on unsupervised object classification, we apply this pipeline to solve an automatic smart robot sorting application first introduced in [24]. The idea is that the robot, equipped with a camera, visualizes a set of objects placed on a table. Given the number of storage boxes available, it needs to figure out the best way of sorting and storing the objects before physically doing it. The approach proposed in this paper exploits semantic information (deep features) while [24] uses a computer vision algorithm to extract shapes and colors. A video of the application can be seen at (https://youtu.be/NpZIwY3H-gE). An example of inputs/output of the application is shown in Figure 1. The robustness of this application is also investigated in section 4 by changing the lighting conditions, the position and orientation of the objects as well as the background. For this robustness validation, we built a dataset that appears to be a challenging one for image-set clustering (http://www.usine-agile.fr/datas/22-cnn-datas-1.html).

![Clustering + Robot sorting](image)

Figure 1: Robotic application description

1.2. Contributions

The main contribution of this paper is to convey further insight into deep CNN features by showing their scalability to unsupervised classification problems. We also propose a new baseline on several image clustering tasks.

Other contributions include the implementation of an application combining unsupervised image classification with robotic sorting. The method proposed to solve this problem, constitutes a new step towards autonomous decision-making robots. The dataset introduced in this paper, which is relatively challenging for image-set clustering, is also a contribution that can be used in further research to investigate robustness to background and lighting conditions for image clustering algorithms.

2. CLUSTERING IMAGES WITH DEEP FEATURE EXTRACTION

2.1. Pipeline description

The pipeline we propose for image set clustering is fairly straightforward. It consists in extracting deep features from all the images in the set, by using a deep convolutional neural network pretrained on a large dataset for image classification and then apply a "standard" clustering algorithm to these features. We initially tried this approach as a first step towards developing a
better clustering algorithm, however, it appears that this simple approach outperforms state-of-the-art algorithm at image-set clustering.

To implement this unsupervised image classification pipeline, we first need to answer four questions:

- What dataset should be used for pretraining?
- What CNN architecture should we use?
- Which layer output should be chosen for feature extraction?
- What clustering algorithm should be used?

As of today, ImageNet is the only very large labelled public dataset which has enough variety in its classes to be a good feature extractor for a variety of tasks. Moreover, there are plenty of CNN pretrained on ImageNet already available online. Hence, we will use a CNN pretrained on ImageNet. The three other questions are answered experimentally. We use the VOC2007 [25] test set without labels, which is new for the pretrained net, to compare performances of the different options.

To ease development and accelerate implementation, we compare the Keras [26] implementations of ResNet50 [27], InceptionV3 [28], VGG16, VGG19 [29] and Xception [30] with the pretrained weights provided by Keras. For the clustering algorithms, we use the scikit-learn [31] implementations of K-means (KM) [32], Minibatch K-means (MBKM) [33], Affinity Propagation (AP) [34], Mean Shift (MS) [35], Agglomerative Hierarchical Clustering (AC) [36], DBScan (DBS) [37] and Birch (Bi) [38]. For each CNN, the layers after which the features are extracted can be found in Table 1 (Layers names are the same as in the Keras implementations).

In the image-set clustering problem, the expected classes are represented by objects present on the picture and for this reason we need semantic information, which is present in the final layers of the network. Thus, we only choose layers among the last layers of the networks. On the one hand, the last one or two layers might provide better results as their goal is to separate the data (at least for the fully-connected layers). On the other hand, the opposite intuition is also relevant as we can imagine that these layers are too specialized to be transferable. These two contradictory arguments motivated the following experiment.

We also note that the test set of VOC2007 has been modified for this validation. We removed all the images presenting two or more labels in order to have ground truth to compute Normalized Mutual Information (NMI) scores. Indeed, if an image possesses several labels we cannot judge if the clustering pipeline classified it properly or not. We note VOC2007-SL (single label) the modified VOC2007 test set.

2.2. Hyperparameters Choice

To answer the questions stated above, we try to cluster the VOC2007-SL set using all combinations of CNN architectures, layer choices and clustering algorithms. To compare performances, we use NMI scores. We also report clustering time for completeness. Only scikit-learn default hyperparameters of the different clustering algorithms are used, which illustrate the simplicity of this approach. For KM and MBKM, as the results depend on random initialization, experiments are run ten times and reported results are averaged over the different runs.
Looking at the results, we choose Agglomerative Clustering on features extracted from the final layer of an Xception CNN pretrained on ImageNet for image-set clustering. This pipeline is then compared to state-of-the-art methods in the next section in order to see how transferable CNN ImageNet features are for unsupervised categorization.

3. Validation on Several Public Datasets

3.1. Datasets Description

The efficiency of the proposed method is demonstrated by comparing it to other recent successful methods (see section 3.2) on several public datasets which characteristics are described in Table 2.

Table 1: NMI scores (in black) and time in seconds (in blue, italics) on Pascal VOC2007-SL test set using different CNN, different output layers and different clustering algorithms. (Layers names are the same as in the Keras implementations).

<table>
<thead>
<tr>
<th>Layer</th>
<th>KM</th>
<th>MBKM</th>
<th>AP</th>
<th>MS</th>
<th>AC</th>
<th>DBS</th>
<th>Ei</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception V3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mixed9</td>
<td>0.108</td>
<td>0.105</td>
<td>0.219</td>
<td>0.153</td>
<td>0.110</td>
<td>0</td>
<td>0.110</td>
</tr>
<tr>
<td>mixed10</td>
<td>0.468</td>
<td>0.401</td>
<td>0.442</td>
<td>0.039</td>
<td>0.595</td>
<td>0</td>
<td>0.595</td>
</tr>
<tr>
<td>avg pool</td>
<td>0.674</td>
<td>0.661</td>
<td>0.621</td>
<td>0.024</td>
<td>0.686</td>
<td>0</td>
<td>0.686</td>
</tr>
<tr>
<td>Resnet 50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg pool</td>
<td>0.678</td>
<td>0.641</td>
<td>0.587</td>
<td>0.043</td>
<td>0.640</td>
<td>0</td>
<td>0.640</td>
</tr>
<tr>
<td>VGG 16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>block4 pool</td>
<td>0.218</td>
<td>0.085</td>
<td>0.133</td>
<td>0.124</td>
<td>0.277</td>
<td>0</td>
<td>0.277</td>
</tr>
<tr>
<td>block5 pool</td>
<td>0.488</td>
<td>0.048</td>
<td>0.262</td>
<td>0.194</td>
<td>0.530</td>
<td>0</td>
<td>0.530</td>
</tr>
<tr>
<td>fc1</td>
<td>0.696</td>
<td>0.458</td>
<td>0.421</td>
<td>0.187</td>
<td>0.617</td>
<td>0</td>
<td>0.617</td>
</tr>
<tr>
<td>fc2</td>
<td>0.691</td>
<td>0.611</td>
<td>0.551</td>
<td>0.085</td>
<td>0.673</td>
<td>0</td>
<td>0.673</td>
</tr>
<tr>
<td>VGG 19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>block4 pool</td>
<td>0.293</td>
<td>0.139</td>
<td>0.124</td>
<td>0.135</td>
<td>0.234</td>
<td>0</td>
<td>0.234</td>
</tr>
<tr>
<td>block5 pool</td>
<td>0.522</td>
<td>0.321</td>
<td>0.250</td>
<td>0.198</td>
<td>0.540</td>
<td>0</td>
<td>0.540</td>
</tr>
<tr>
<td>fc1</td>
<td>0.670</td>
<td>0.471</td>
<td>0.449</td>
<td>0.188</td>
<td>0.628</td>
<td>0</td>
<td>0.628</td>
</tr>
<tr>
<td>fc2</td>
<td>0.672</td>
<td>0.615</td>
<td>0.557</td>
<td>0.083</td>
<td>0.671</td>
<td>0</td>
<td>0.671</td>
</tr>
<tr>
<td>Xception</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>block13 pool</td>
<td>0.376</td>
<td>0.264</td>
<td>0.351</td>
<td>0.044</td>
<td>0.473</td>
<td>0</td>
<td>0.473</td>
</tr>
<tr>
<td>block14 act</td>
<td>0.574</td>
<td>0.428</td>
<td>0.508</td>
<td>0.071</td>
<td>0.634</td>
<td>0</td>
<td>0.634</td>
</tr>
<tr>
<td>avg pool</td>
<td>0.692</td>
<td>0.636</td>
<td>0.636</td>
<td>0.052</td>
<td>0.726</td>
<td>0</td>
<td>0.726</td>
</tr>
</tbody>
</table>
The clustering tasks involved by these datasets are different from each other (Face recognition, grouping different objects, recognizing different pictures of the same object). In addition, the content of the classes differs from the ones in ImageNet. For these reasons, the four datasets constitute a good benchmark to quantify the robustness of transfer learning for unsupervised object categorization.

### 3.2. Results Comparison

We propose a comparison with the results reported in the following papers dealing with image set clustering:

[11] proposes different clustering algorithms applied on bags of features. In Table 3, we note "BoF" the best results obtained by such pipeline on the different datasets.

[10] proposes a method called infinite ensemble clustering (IEC). In the paper, IEC algorithm is compared to several other deep clustering algorithms and ensemble clustering algorithms. In Table 3, we report the best results obtained using Deep Clustering (DC) and Ensemble Clustering (EC) for each datasets. We note that for VOC2007-5-ML, [10] also uses deep features as clustering inputs (the CNN used is not reported).

[22] proposes a method called Joint Unsupervised Learning (JULE) of Deep Representations and Image Clusters, based on Alternating optimization between clustering and weight optimization of the CNN feature extractor. Results from this work are reported in Table 3.

For each dataset groundtruth is known as they are intended for supervised classification. We compute both NMI scores and purity for each dataset/method pair.

Table 3 shows that features extracted by the final layer of Xception combined with Agglomerative Clustering performs better than or close to state-of-the-art methods at unsupervised object classification as well as fine-grained image clustering (ORL). Results on the ORL dataset are interesting as they show that pretrained Xception is able to classify different faces without supervision, although ImageNet does not deal with human faces at all.

1The poor results with DB-Scan might come from the default parameters. We might get better results using different configurations, but this is out of the scope of this paper.

2The data used for VOC2007 in [10] are irrelevant for clustering with hard assignment. The VOC2007 subset used in [10] contains images with several objects but only one label. However, we still ran our clustering method on this data to be able to compare results. This second modified VOC2007 set is denoted VOC2007-5-ML (5 classes - multiple labels)
This is an important result as it shows that, with today's methods, given an unlabeled image-set, we can extract more information from a large labeled dataset, with a large variety of classes, than from the set itself. It is better than hand-engineered approaches (BoF) as well as unsupervised trained deep networks (Deep clustering and Ensemble clustering). It also raises the question of how the representation learning problem should be handled. Indeed, although less satisfactory from a research perspective, it might be more appropriate to work on the creation of a larger database and train networks on it so that it can be used as a knowledge base for many other problems.

We underline the very good results of JULE ([22]) at clustering COIL100. Regarding the results of this methodology on Scene clustering [18], it appears that fine tuning feature extraction using alternating optimization is a good way of improving clustering results. However, the simple approach proposed here still keeps the advantage of being very fast (as it only requires to evaluate the network once for each sample and apply AC), which is useful for our application for instance.

It is also interesting to notice that [10] is also using CNN features for clustering VOC2007-5-ML. The fact that our pipeline outperform largely DC and EC for this dataset illustrate that when using transfer learning for a new problem, careful selection of the CNN used is important, which is often neglected.

### 3.3. A Word on Scene Clustering

The problem studied in this paper is the one raised by the robotic application, unsupervised objects sorting. For this reason, we compared our result at different object categorization tasks as well as fine-grained classification for clustering of similar objects. Another interesting image classification problem is the one of scene clustering, studied in [18] on two datasets ([42, 43]). For this task, the pipeline proposed in this paper cannot perform as well as for object classification. Indeed, ImageNet does not contain any class which requires to group objects together. Thus, although the features are good at supervised scene classification, without further training they are not able to group objects together as such behaviour is not encoded in the final layers of the CNN. However, this issue is not inherent to the method defined and we believe that with bigger and more versatile datasets, results would be as good as any other method.
4. APPLICATION VALIDATION

In the setting tested initially (Figure 1), where the set of objects to cluster is composed of screw drivers, at keys, allen keys and clamps, and where the background is a table, the success rate of the application described in section 1.1.2 is 100%. Although for certain classes (flat and allen keys) the intra-cluster similarity is high, it is not the case for the two others. This task is also difficult to solve because we carried out the experiments in a shopfloor under unmastered lighting.

For further testing of the application robustness, we have built a dataset for pixel-based object clustering. The full dataset, together with its description, can be found at (http://www.usine-agile.fr/datas/22-cnn-datas-1.html) and example images can be seen on Figure 2. The dataset statistics are reported in table 4. This dataset is difficult because some classes have low intra-cluster similarity (usb) and extra-cluster similarity between some classes is relatively high (pens/screws). The lighting conditions as well as the background also change within the dataset, which makes the task even harder.

Table 4: Several key features about the constructed dataset.

<table>
<thead>
<tr>
<th>Problem type</th>
<th>Image Size</th>
<th># Classes</th>
<th># Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object recognition</td>
<td>640 x 480</td>
<td>7</td>
<td>560</td>
</tr>
</tbody>
</table>

For each conditions, we randomly pick one out of the several pictures (different position/orientation) of each object. The results reported in the top subtable of Table 5 are averaged over 100 random combinations. The clustering results are not perfect, looking at the misclassifications, the main source of error comes from classes with low intra-cluster similarity (pens, usb) and from background containing sharp edges (conditions 4 and 5).

Figure 2: Example images from the robustness validation dataset. Objects in the same row are expected to be clustered in the same group. The five different background/lighting conditions are represented in this figure.

We also carry out an experiment of fine grained classification with this dataset. Within each class, we try to group together the pictures representing the exact same object. Purity results can be found in the bottom subtable of Table 5. An interesting fact can be noticed about object...
recognition within a category. Classes responsible for decreasing the clustering quality in the first subtable are objects with the highest purity in the second table. Such remark makes sens as low intra-cluster similarity is good for the second task but harmful for the first one.

Looking at the results, this robustness validation dataset appears to be a challenging one for image-set clustering and could be used to validate further research in the field.

Table 5: Clustering accuracies for different background, lighting and orientation conditions on the tool clustering dataset.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Grouping by category</th>
<th>Recognizing object inside a category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Condition</td>
<td>1</td>
</tr>
<tr>
<td>Purity</td>
<td></td>
<td>0.85</td>
</tr>
<tr>
<td>NMI score</td>
<td></td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. CONCLUSION AND PERSPECTIVES

5.1. Conclusive Remarks

This paper extends the interesting work of [5, 6, 7] about the transferability of CNN features. It shows that, even for unsupervised classification tasks, features extracted from deep CNN trained on large and diverse datasets, combined with classic clustering algorithms, can compete with more sophisticated and tuned image-set clustering methods. The fairly simple and naive pipeline proposed outperforms the best results reported in recent work, which raises the question of which research direction should be chosen to reach generic knowledge. Are efforts spent in developing image representation extractors more useful than simply building larger and more diverse datasets?

This approach is used to implement a robotic application using unsupervised image classification to store objects smartly. To validate this application, we also built a challenging dataset for image clustering that is made available to the research community.

5.2. Future Work

The proposed improvements mainly go in the direction of the robotics application, which is still not robust enough to adapt perfectly to very different looking objects within a cluster and to difficult backgrounds and lighting conditions. If we want to make it work in difficult...
environments, the clustering pipeline needs to be improved. One possible direction is to tune the final clustering algorithm, indeed, the scikit clustering algorithms are used without any parameter tuning, setting hyperparameters to their default values.

The sorting application can also be improved by introducing automatic image segmentation, which would make it more suitable for practical uses. To do this, we could use a pretrained region proposal network [44] and cluster objects in the proposed regions.

ACKNOWLEDGEMENTS

This work is also supported by the European Union's 2020 research and innovation program under grant agreement No.688807, project ColRobot (collaborative robotics for assembly and kitting in smart manufacturing).

The authors would like to thank Jorge Palos, Gil Boye de Sousa, Leon Sixt and Harshvardan Gazula for their precious advices and technical support.

REFERENCES


AUTHORS

Joris Guérin received the diplome d'ingénieur (equivalent to M.Sc. degree) from Arts et Metiers ParisTech and the M.Sc. in Industrial Engineering from Texas Tech University, both in 2015. He is currently a Ph.D student at Laboratoire des Sciences de l'Information et des Systèmes (LSIS), at Arts et Métiers ParisTech, Lille, France. His current research focuses on Clustering and Reinforcement Learning for Robotics manipulation.

Olivier Gibaru is currently full professor at the Department of Mathematics and Computer Science at ENSAM, Lille campus. He obtained his PhD in applied mathematics in 1997. His main research interests includes: applied mathematics, estimation for robotic applications, geometry, control engineering and high precision mechanical systems. He is the coordinator of the EU Horizon 2020 ColRobot project www.colrobot.eu. He is an active member of the SMAI-SIGMA group which is a national learned society dedicated to Applied Mathematics for the Industrial Applications.

Stéphane Thiery received the Ph.D degree in Automatics from University of Nice-Sophia Antipolis, France, in 2008. He was a post-doctoral fellow in the NON-A team in INRIA-Lille, for eight months in 2009-2010, and joined Arts et Métiers ParisTech Engineering School, as assistant professor in Applied Mathematics and Automatics, in 2010. His current research includes machine learning, real-time parameters estimation, and control of mechanical systems.

Eric Nyiri received the Ph.D degree in Computer Science from University of Lille I, France, in 1994. He joined Arts et Métiers ParisTech Engineering School, as an assistant professor in Applied Mathematics and Computer Science, in 1995. His initial research domain was L1 interpolation and approximation. In 2010, he joined the LSIS Lab and his current research includes machine learning and path planning for robots. Since 2016, he is a member of the COLROBOT European project.
COMPARISON OF BANKRUPTCY PREDICTION MODELS WITH PUBLIC RECORDS AND FIRMOGRAPHICS

Lili Zhang\(^1\), Jennifer Priestley\(^2\) and Xuelei Ni\(^3\)

\(^1\)Program in Analytics and Data Science, Kennesaw State University, Georgia, USA
\(^2\)Analytics and Data Science Institute, Kennesaw State University, Georgia, USA
\(^3\)Department of Statistics, Kennesaw State University, Georgia, USA

ABSTRACT

Many business operations and strategies rely on bankruptcy prediction. In this paper, we aim to study the impacts of public records and firmographics and predict the bankruptcy in a 12-month-ahead period with using different classification models and adding values to traditionally used financial ratios. Univariate analysis shows the statistical association and significance of public records and firmographics indicators with the bankruptcy. Further, seven statistical models and machine learning methods were developed, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, Bayesian Network, and Neural Network. The performance of models were evaluated and compared based on classification accuracy, Type I error, Type II error, and ROC curves on the hold-out dataset. Moreover, an experiment was set up to show the importance of oversampling for rare event prediction. The result also shows that Bayesian Network is comparatively more robust than other models without oversampling.

KEYWORDS

Bankruptcy Prediction, Public Records, Firmographics, Classification, Oversampling

1. INTRODUCTION

Bankruptcy prediction has been studied since the 1960s, to improve decision making related to business operations conducted with reliable counterparties [3]. For example, investors want to make investments to organizations that have high potential to succeed. Banks want to lend to the organizations that are less likely to default. Business entities want to do business and build relationships with the ones that can prosper and survive in a long term. Hence, it is valuable to foresee the possibility of the bankruptcy of a business customer or partner.

To improve the accuracy of bankruptcy prediction, researchers and practitioners have pursued two primary paths of study. First, explore important variables for bankruptcy prediction. For example, the predictive ability of financial ratio variables has been thoroughly studied. Second, improve the methodologies used for the bankruptcy prediction, benefiting from the development of both the algorithm theories and computation infrastructure. Besides significant variables and high-performance methods, we observe that appropriate data sampling before modeling is also
important for improving bankruptcy prediction, considering that frequently the proportion of bankruptcy cases is substantially lower than the proportion of non-bankruptcies.

In this paper, we aim to make contributions from all above perspectives. First, we explore the impacts of public records and firmographics on bankruptcy prediction to add values to widely used financial ratio variables. Both univariate analysis and multiple variable analysis were conducted to measure statistical association and significance. With significant variables selected, we comprehensively compare seven classification models from the statistics and machine learning domains, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, Neural Network, and Bayesian Network. The performance of the models are evaluated on the hold-out dataset. The overall classification accuracy, Type I error, Type II error, and ROC curves are evaluated. Finally, we demonstrate the importance of oversampling for the rare event prediction like bankruptcy prediction, and demonstrate the robustness of the Bayesian Network for rare event modeling.

The paper is structured as follows. In Section 2, related work is reviewed. In Section 3, the data processes are described. In Section 4, the univariate analysis between the dependent variable and each individual input variable is performed. In Section 5, the models are developed, diagnosed, evaluated, and compared. In Section 6 and 7, conclusions and future work are discussed.

2. RELATED WORK

Because of its importance in business decisions like investment and loan lending, the bankruptcy prediction problem has been studied through deriving significant predictors and developing novel prediction models. Altman proposed a set of traditional financial ratios, including Working Capital/Total Assets, Retained Earnings/Total Assets, Earnings before Interest and Taxes/Total Assets, Market Value Equity/Book Value of Total Debt, and Sales/Total Assets, and used them in the multiple discriminant analysis for the corporate bankruptcy prediction [2]. Those financial ratios were widely adopted and extended later [13][4]. Amircame up with some novel financial ratio indicators, including BookValue/Total Assets, Cashflow/Total Assets, Price/Cashflow, Rate of Change of Stock Price, and Rate of Change of Cashflow perShare, in addition to Altman’s ones, for a neural network model, and increased the prediction accuracy by 4.04% for a three-year-ahead forecast [4]. Everett et al. studied the impact of external risk factors (i.e. macro-economic factors) on small business bankruptcy prediction and proposed a logistic regression model [7]. Chavaet al. demonstrated the statistical significance of industry effects by grouping firms into finance/insurance/real estate, transportation/communications/utilities, manufacturing/mineral, and miscellaneous industries [6].

From the methodology perspective, various statistical methods, machine learning algorithms, and hybrid models have been applied and compared for the bankruptcy prediction problem. Odom et al. proposed the first neural network model for bankruptcy prediction [13]. Zhang et al. showed that the neural network performed better than logistic regression and were robust to sampling variations [17]. Shin et al. found that the support vector machine outperformed the neural network on small training datasets [14]. Min et al. applied support vector machine with optimal kernel function hyper parameters [12]. Zibanez had showed the acceptable prediction ability of decision tree on the bankruptcy prediction problem and determined the most important financial ratios [8]. Zikeba et al. proposed and evaluated a novel gradient boosting method for learning an ensemble of trees [18]. Sun et al. studied the application of Bayesian network on the bankruptcy prediction problem in respects of the influence of variable selection and variable discretization on the model performance [15]. Ahn et al. presented a hybrid methodology by combining rough set theory and neural network [1]. Huang et al. proposed a hybrid model by incorporating static and trend analysis in the neural network training [9]. Kumar et al. provided a comprehensive review on both
the financial ratio variables and methods used for the bankruptcy prediction from 1968 to 2005, discussed merits and demerits of each method, and listed some important directions for future research [11]. Bellovary et al. reviewed 165 existing studies for the bankruptcy prediction and made some suggestions, where one suggestion was that the model accuracy was not guaranteed with the number of factors [5].

Most models proposed for bankruptcy prediction in the literature were directly developed on the dataset with a balanced proportion of bankruptcy and non-bankruptcy observations. However, data imbalance is a common issue in practice. Kim et al. proposed a geometric mean based boosting algorithm to address the data imbalance problem in the bankruptcy prediction, but only compared it with other boosting algorithms to show its advantage [19]. Zhou studied the effect of sampling methods for five bankruptcy prediction models, but the models were not tuned to their optimal hyperparameters [20].

The models applied to the bankruptcy prediction utilize a variety of algorithms. Logistic Regression formulates a function between the probability of the event ($\hat{p}$) and input variables ($x_1, x_2, ..., x_n$) defined as:

$$\hat{p} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n)}}$$

The coefficients ($\beta_1, \beta_2, ..., \beta_n$) in the function are estimated by optimizing the maximum likelihood function defined as below, where $y$ is the actual value with the event denoted as 1 and the nonevent denoted as 0.

$$\max \{\log \hat{p} + (1 - y) \log(1 - \hat{p})\}$$

Decision Tree defines hierarchical rules by searching for optimal splits on input variables based on the Entropy or Gini index. The Entropy and Gini index of an input variable are defined below, where $x$ is a given input variable, $1, ..., k$ are levels in the dependent variable, and $p(i|k)$ is the conditional probability for the dependent variable taking value $i$ given $x$ [16].

$$\text{Entropy}(x) = -\sum_{i=1}^{k} p(i|k) \log_2(p(i|k))$$

$$\text{Gini}(x) = 1 - \sum_{i=1}^{k} [p(i|k)]^2$$

Random Forest and Gradient Boosting are an ensemble of multiple decision tree models through bagging and boosting, respectively. In Random Forest, each tree is trained independently on a bootstrap dataset created from the original training dataset and then combined to a single prediction model by taking the average of all trees. In Gradient Boosting, each tree is trained sequentially based on a modified version of the original training dataset by utilizing the information of previously trained trees [10]. In tree-based models, a summary of variable importance can be obtained. The importance of each input variable is measured based on the Entropy or Gini reduction by splitting a given input variable. The larger the value is, the more important an input variable is.

Support Vector Machine defines a hyperplane for two-class classification by maximizing the marginal distance. To handle the nonlinear relationship, a kernel function can be first applied to project the input variables to a higher feature space. Neural Network learns the relationship between the dependent variable and input variables by first transforming input variables with an activation function (Tanh, Sigmoid, etc.) through each hidden unit in one or more hidden layers and then adjusting the weights through backpropagation iteratively to minimize a loss function. Bayesian Network represents the probability relationship and conditional dependencies between the dependent variable and input variables via a directed cyclic graph.
3. **DATA**

The bankruptcy indicator, public records and firmographics information of 11,787,287 U.S. companies in the 4th Quarter of 2012and 2013 was collected by a national credit reporting agency, and were approved for use in this study. From the data, a bankruptcy flag indicates whether a corporate is in bankruptcy or in business at the capture time point. Firmographics in the data include industry, location, size, and status and structure. Each corporate is identified by its unique Market Participant Identifier (MPID).Public Records include judgements and liens reported.

From the dataset provided, we aim to answer the following question explicitly, which can provide decision makers with insights into improved bankruptcy prediction.

*Given the public records and firmographics indicators of an organization in one quarter, can we predict its operation status one year in the future?*

To answer the question above, the dependent variable Bankruptcy Indicator Change (i.e. BrtIndChg) was created and is provided in Table 1. Originally, Bankruptcy Indicator (i.e.BrtInd) has two levels, 0 and 1, where 0 indicates that the organization is operating and 1 indicates a bankruptcy. If an organization in business in 2012 went to bankruptcy in 2013, then BrtIndChg was assigned to 1. If the organization was still in business in 2013, then BrtIndChg was assigned to 0.

The raw data had to be cleaned and transformed prior to modeling, to address missing values, abnormal/incorrect values, and correlated variables. The following steps were applied to the data.

1. Only keep observations with the level value 0 in the original 2012 BrtInd.
2. Create the dependent variable BrtIndChg by comparing BrtInd in the dataset of 2012 and 2013 as shown in Table 1.
3. Drop interval variables if the percentage of coded values or missing values is greater than 30%. A value of 30% was selected to optimize the percent of variance explained in the dataset.
4. Drop observations in an interval variable or a categorical variable if the percentage of the abnormal/incorrect values in that variable is less than 5%.
5. Continuous variables were binned into nominal variables. For example, the variable Number of Current Liens or Judgment was binned into Current Liens or Judgment Indicator (i.e. curLiensJudInd) with two levels, 0and 1, where 0 means an organization does not have a lien or judgment currently and 1 means an organization has one or more liens or judgments currently.
6. Retain the variable with the best predictive ability among several correlated variables. For example, based on both the variable definition and the Chi-Square value, the following variables are correlated: Current Liens/Judgment Indicator, Number of Current Liens/Judgment and Total Current Dollar Amounts on All Liens/Judgments. After comparing their performance, only the variable Current Liens/Judgment Indicator was kept.
Table 1. Creation of Dependent Variables

<table>
<thead>
<tr>
<th>BrtInd 2012</th>
<th>BrtInd 2013</th>
<th>BrtIndChg</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

After the data was cleaned, the variables in Table 2 were prepared for further analysis and modeling. As described above, the bankruptcy is a rare event, which can be further confirmed by the distribution of the dependent variable BrtIndChg, as shown in Table 3. In our dataset, there are 0.12% of observations going into bankruptcy from 2012 to 2013 and 99.88% of observations staying in business from 2012 to 2013. Because the proportion of event cases is much less than the proportion of nonevent cases, we need to consider oversampling to have sufficient event cases to train the model and achieve better performance, which will be discussed in detail in Section 5.

Table 2. Variables for Analysis and Modelling.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPID</td>
<td>Nominal</td>
<td>Market Participant Identifier</td>
</tr>
<tr>
<td>BrtIndChg</td>
<td>Binary</td>
<td>Bankruptcy Indicator Change</td>
</tr>
<tr>
<td>curLiensJudInd</td>
<td>Nominal</td>
<td>Current Liens/Judgment Indicator</td>
</tr>
<tr>
<td>histLiensJudInd</td>
<td>Nominal</td>
<td>Historical Liens/Judgment Indicator</td>
</tr>
<tr>
<td>Industry</td>
<td>Nominal</td>
<td>Industry</td>
</tr>
<tr>
<td>LargeBusinessInd</td>
<td>Nominal</td>
<td>Large Business Indicator</td>
</tr>
<tr>
<td>Region</td>
<td>Nominal</td>
<td>Geographical Region</td>
</tr>
<tr>
<td>PublicCompanyFlag</td>
<td>Nominal</td>
<td>Public Company Flag</td>
</tr>
<tr>
<td>SubsidiaryInd</td>
<td>Nominal</td>
<td>Subsidiary Indicator</td>
</tr>
<tr>
<td>MonLstRptDatePlcRec</td>
<td>Interval</td>
<td>Number of Months Since Last Report Date on Public Records</td>
</tr>
</tbody>
</table>

Table 3. Frequency of Dependent Variable.

<table>
<thead>
<tr>
<th>BrtIndChg</th>
<th>Frequency</th>
<th>Percent(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1031</td>
<td>0.12</td>
</tr>
<tr>
<td>0</td>
<td>843330</td>
<td>99.88</td>
</tr>
</tbody>
</table>

4. EXPLORATORY ANALYSIS

To examine the statistical association and significance between each individual input variable and the dependent variable, bivariate analysis was performed. The results of odds ratio and Chi-square test can be found in Table 4. Based on the Chi-Square results, all the variables are significantly associated with the dependent variable except the variable Public Company Flag. Based on the odds ratio, we have the following observations regarding their relationship:

- Current Lien/Judgment Indicator: The organizations which currently do not have any lien/judgment is about 47.1% less likely to go into bankruptcy in the following year than those which currently have liens or judgments.
- Historical Lien/Judgment Indicator: The organizations which did not have any lien/judgment is about 32% less likely to go into bankruptcy in the following year than the ones which historically had liens or judgments.
- Large Business Indicator: The organizations which are not large are about 45.8% less likely to go into bankruptcy in the following year than the ones which are large.
- Subsidiary Indicator: The organizations which are not subsidiaries are 74.5% more likely to go into bankruptcy in the following year than those organizations which are subsidiaries.
• **Industry**: By using the industry group 8 as the reference level, the organizations in the industry group 3 are about 2 times more likely going to the bankruptcy in the following year than the ones in the industry group 8.

• **Region**: By using the region group 9 as the reference level, the organizations in the region group 2 are about 55.7% less likely to go into bankruptcy in the following year than the ones in the region group 9.

• **Number of Months Since Last Report Date on Public Records** (i.e. MonLstDatePlcRec): Figure 1 shows that the distribution of MonLstDatePlcRec is very different in different levels of BrtIndChg, indicating their strong relationship.

### Table 4. Univariate Odds Ratio and Chi-Square p-value.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Odds Ratio</th>
<th>95% Confidence Interval</th>
<th>Chi-Square p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>curLiensJudInd0 vs 1</td>
<td>0.529</td>
<td>[0.447, 0.627]</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>histLiensJudInd 0 vs 1</td>
<td>0.680</td>
<td>[0.601, 0.768]</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>LargeBusinessIndN vs Y</td>
<td>0.542</td>
<td>[0.474, 0.620]</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>LargeBusinessIndU vs Y</td>
<td>0.202</td>
<td>[0.165, 0.249]</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>PublicCompanyFlagN vs Y</td>
<td>0.295</td>
<td>[0.104, 0.838]</td>
<td>0.065</td>
</tr>
<tr>
<td>PublicCompanyFlagU vs Y</td>
<td>0.370</td>
<td>[0.138, 0.989]</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>SubsidiaryIndN vs Y</td>
<td>1.745</td>
<td>[0.997, 3.053]</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>SubsidiaryIndU vs Y</td>
<td>0.411</td>
<td>[0.261, 0.648]</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Industry1 vs 8</td>
<td>1.538</td>
<td>[0.947, 2.496]</td>
<td></td>
</tr>
<tr>
<td>Industry2 vs 8</td>
<td>3.085</td>
<td>[1.118, 8.514]</td>
<td></td>
</tr>
<tr>
<td>Industry3 vs 8</td>
<td>2.079</td>
<td>[1.545, 2.797]</td>
<td></td>
</tr>
<tr>
<td>Industry4 vs 8</td>
<td>1.971</td>
<td>[1.365, 2.847]</td>
<td></td>
</tr>
<tr>
<td>Industry5 vs 8</td>
<td>1.648</td>
<td>[1.136, 2.392]</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Industry6 vs 8</td>
<td>2.421</td>
<td>[1.704, 3.439]</td>
<td></td>
</tr>
<tr>
<td>Industry7 vs 8</td>
<td>1.386</td>
<td>[1.033, 1.859]</td>
<td></td>
</tr>
<tr>
<td>Industry9 vs 8</td>
<td>1.348</td>
<td>[1.012, 1.795]</td>
<td></td>
</tr>
<tr>
<td>Industry10 vs 8</td>
<td>0.885</td>
<td>[0.216, 3.629]</td>
<td></td>
</tr>
<tr>
<td>IndustryU vs 8</td>
<td>0.473</td>
<td>[0.343, 0.651]</td>
<td></td>
</tr>
<tr>
<td>Region1 vs 9</td>
<td>0.699</td>
<td>[0.479, 1.019]</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Region2 vs 9</td>
<td>0.443</td>
<td>[0.358, 0.549]</td>
<td></td>
</tr>
<tr>
<td>Region3 vs 9</td>
<td>0.627</td>
<td>[0.505, 0.779]</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Region4 vs 9</td>
<td>0.913</td>
<td>[0.686, 1.215]</td>
<td></td>
</tr>
<tr>
<td>Region5 vs 9</td>
<td>0.636</td>
<td>[0.525, 0.772]</td>
<td></td>
</tr>
<tr>
<td>Region6 vs 9</td>
<td>1.203</td>
<td>[0.928, 1.558]</td>
<td></td>
</tr>
<tr>
<td>Region 7 vs 9</td>
<td>1.084</td>
<td>[0.875, 1.343]</td>
<td></td>
</tr>
<tr>
<td>Region 8 vs 9</td>
<td>1.194</td>
<td>[0.920, 1.549]</td>
<td></td>
</tr>
<tr>
<td>MonLstRptDatePlcRec</td>
<td>0.971</td>
<td>[0.969, 0.973]</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
5. METHODOLOGY

To better train and evaluate the models, the dataset was first oversampled and then split into training dataset and validation dataset, where the training dataset was used for training the models and the validation dataset was used as the hold-out dataset for evaluating the performance of models. Seven different models were developed, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, Bayesian Network, and Neural Network. Their respective performances were then evaluated by overall accuracy, Type I error, Type II error, and ROC curve.

Figure 1. Boxplot of MonLstRptDatePlcRec by BrtIndChg

5.1. Sampling

Sampling was done in three steps.

(1) Oversampling: The proportion of the events is 0.12%, as indicated in Table 3. To have sufficient event cases to train the model and achieve better performance, the oversampling technique is used to adjust the proportion of event observations and non-event observations to 50% versus 50%, which include all the event observations and an equal number of randomly selected non-event observations. That ends up with 1031 bankruptcy observations and 1031 non-bankruptcy observations.

(2) Training Dataset and Validation Dataset Split: The out-of-sample test is used by evaluating the models on the hold-out data set. Hence the dataset is split into training and validation by 70% versus 30%, respectively.

(3) Oversampling Adjustment: Prior probability and inverse prior weights are applied to the results to adjust oversampling.

5.2. Model Development and Evaluation

The models were developed using SAS Enterprise Miner. All variables in Table 4 are specified as initial inputs for all models. Every model is tuned to their best performance by trying different hyper parameter values.

In Logistic Regression, backwards selection is used to select significant variables with the significance level set to 0.05. The multi variate odds ratio and Chi-Square p-value of the resulting model can be found in Table 5. The significant variables include curLiensJudInd, histLiensJudInd,
Large Business Ind, Region, and MonLstDatePlcRec. Their multivariate odds ratio is consistent with their univariate odds ratio. For example, univariate odds ratio shows that curLiensJudInd is negatively associated with the dependent variable, which is the same as indicated by the multivariate odds ratio of curLiensJudInd.

Decision Tree, Gradient Boosting, and Random Forest are all tree-based models. Entropy is used as the criteria of searching and evaluating candidate splitting rules for Decision Tree, while Gini index is used for Gradient Boosting and Random Forest. The important variables selected by these models include MonLstDatePlcRec, Region, Industry, curLiensJudInd, histLiensJudInd, and LargeBusinessInd. Their importance measure can be found in Table 6. Note that for Decision Tree and Gradient Boosting, the importance measure presented here is the total Entropy or Gini reduction, while for Random Forest, the importance measure is the marginal Gini reduction.

Table 5. Multivariate Odds Ratio and Chi-Square p-value.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Odds-Ratio</th>
<th>Chi-Square p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>curLiensJudInd0 vs 1</td>
<td>0.573</td>
<td>0.0046</td>
</tr>
<tr>
<td>histLiensJudInd0 vs 1</td>
<td>0.508</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>LargeBusinessIndN vs Y</td>
<td>0.796</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>LargeBusinessIndU vs Y</td>
<td>0.332</td>
<td></td>
</tr>
<tr>
<td>Region1 vs 9</td>
<td>1.067</td>
<td></td>
</tr>
<tr>
<td>Region2 vs 9</td>
<td>0.411</td>
<td></td>
</tr>
<tr>
<td>Region3 vs 9</td>
<td>0.583</td>
<td>0.0002</td>
</tr>
<tr>
<td>Region4 vs 9</td>
<td>0.839</td>
<td></td>
</tr>
<tr>
<td>Region5 vs 9</td>
<td>0.558</td>
<td></td>
</tr>
<tr>
<td>Region6 vs 9</td>
<td>0.858</td>
<td></td>
</tr>
<tr>
<td>Region 7 vs 9</td>
<td>0.881</td>
<td></td>
</tr>
<tr>
<td>Region 8 vs 9</td>
<td>1.261</td>
<td></td>
</tr>
<tr>
<td>MonLstRptDatePlcRec</td>
<td>0.976</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Table 6. Variable Importance.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Decision Tree</th>
<th>Gradient Boosting</th>
<th>Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>MonLstRptDatePlcRec</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.0911</td>
</tr>
<tr>
<td>Region</td>
<td>0.2423</td>
<td>0.2880</td>
<td>0.0048</td>
</tr>
<tr>
<td>Industry</td>
<td>0.1663</td>
<td>0.3516</td>
<td>0.0110</td>
</tr>
<tr>
<td>curLiensJudInd</td>
<td>0.1550</td>
<td>0.0820</td>
<td>0.0024</td>
</tr>
<tr>
<td>histLiensJudInd</td>
<td>0.1192</td>
<td>0.1205</td>
<td>0.0038</td>
</tr>
<tr>
<td>LargeBusinessInd</td>
<td>0.0308</td>
<td>0.2752</td>
<td>0.0100</td>
</tr>
</tbody>
</table>

In Support Vector Machine, linear kernel function performs better than polynomial kernel function. In Neural Network, Tanh is used as the activation function in the hidden layer while Sigmoid is used in the output layer. Its architecture can be found in Figure 2. In Bayesian Network, the significant variables selected by G-Square with the significance level 0.2 include MonLstDatePlcRec, Region, Industry, curLiensJudInd, histLiensJudInd, LargeBusinessInd, and SubsidiaryInd. The resulting Bayesian Network can be found in Figure 3.
The accuracy, Type I error, and Type II error of all models are summarized in Table 7. Overall speaking, there is always a tradeoff between Type I error and Type II error, where a large Type I error causes less profits by classifying organizations with low bankruptcy risk into high risky ones and a large Type II error brings more losses by classifying organizations with high bankruptcy risk into low risky ones. For example, according to Type II error, Support Vector Machine performs the best, but it gives the worst overall accuracy and Type I error in the meantime. Practitioners are suggested to select the model by making a balance among model characteristics (accuracy, Type I error, Type II error, interpretability, etc.) based on their expectations. For example, Neural Network, Bayesian Network, and Logistic Regression give the same Type II Error in this case, Bayesian Network and Logistic Regression may be favored than Neural Network because of their high interpretability.

Table 7. Performance of Models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Type I Error</th>
<th>Type II Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support Vector Machine</td>
<td>73.39%</td>
<td>40.32%</td>
<td>12.90%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>75.16%</td>
<td>36.45%</td>
<td>13.87%</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td>74.51%</td>
<td>31.29%</td>
<td>17.42%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>75.80%</td>
<td>29.35%</td>
<td>19.03%</td>
</tr>
<tr>
<td>Neural Network</td>
<td>74.52%</td>
<td>29.35%</td>
<td>19.35%</td>
</tr>
<tr>
<td>Bayesian Network</td>
<td>74.35%</td>
<td>31.93%</td>
<td>19.35%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>74.35%</td>
<td>31.61%</td>
<td>19.35%</td>
</tr>
</tbody>
</table>

To more comprehensively compare these models, ROC curves on both the training and validation dataset are provided in Figure 4 and Figure 5, respectively. For all models, there is no large
difference between training ROC and validation ROC, so there is no overfitting. Moreover, all models overall perform similar, because there is no large gap among their ROCs.

5.3. Experiment without Oversampling

To show the influence of oversampling in the rare event prediction, the dataset and models were fitted without oversampling. The resulting performance measures on the validation dataset can be found in Table 8. All models, except Bayesian Network, classify all bankruptcy observations to non-bankruptcy, as indicated by Type II error, although they have high overall accuracy 99.88% which is exactly the proportion of non-bankruptcy observations in the original data. For Bayesian Network without, its overall accuracy and Type II error decreased by 9.92% and 2.03%, respectively, while its Type I error increased by 3.66%. The ROC curve in Figure 6 further shows that all other models perform no better than random selection, except Bayesian Network.

Figure 4. ROC Curve on Training Dataset

Figure 5. ROC Curve on Validation Dataset
Table 8. Performance of Models without Oversampling.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Type I Error</th>
<th>Type II Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support Vector Machine</td>
<td>99.88%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>99.88%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td>99.88%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>99.88%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Neural Network</td>
<td>99.88%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Bayesian Network</td>
<td>64.43%</td>
<td>35.59%</td>
<td>17.32%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>99.88%</td>
<td>0%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Figure 6. ROC Curve on Validation Dataset without Oversampling

6. DISCUSSIONS AND CONCLUSIONS

Based on the univariate analysis and multivariate analysis, the impacts of public records and firmographics indicators were comprehensively studied. With them as input variables of different classification models, the model results show that public records and firmographics indicators play an important role in the bankruptcy prediction. This may serve as a reference for practitioners and researchers to include these information in the bankruptcy prediction model.

Different classification models generate quite different Type I/II error, although their overall accuracy is similar. Support Vector Machine gives the lowest Type II error, and Logistic Regression gives the lowest Type I error. Regarding the interpretability, Logistic Regression, Decision Tree and Bayesian Network might be favorable choices. We also find that on the dataset with small/medium size, simple models may outperform complicated models like ensemble models and Neural Network. Practitioners may handle the tradeoff between Type I error and Type II error as well as the model interpretability and accuracy based on their expectations.

For rare event prediction, the oversampling is necessary before modeling to achieve better performance. Bayesian Network is quite robust for rare event prediction without oversampling, compared to other classification models.
7. Future Work

In this study, we only focused on the public records and firmographics indicators. In the future, we may collect other information like financial ratios to further reduce Type I/II error, and test the model performance in a wider time spanning.

References


AUTHORS

Lili Zhang is currently a Ph.D. candidate in Analytics and Data Science at Kennesaw State University, working on the data mining, machine learning, and data management.
INTENTIONAL BLANK
Predicting Players' Performance in One Day International Cricket Matches Using Machine Learning

Kalpdrum Passi and Niravkumar Pandey
Department of Mathematics and Computer Science,
Laurentian University,
Sudbury, Canada

Abstract

Player selection is one of the most important tasks for any sport and cricket is no exception. The performance of the players depends on various factors such as the opposition team, the venue, his current form etc. The team management, the coach and the captain select 11 players for each match from a squad of 15 to 20 players. They analyze different characteristics and the statistics of the players to select the best playing 11 for each match. Each batsman contributes by scoring maximum runs possible and each bowler contributes by taking maximum wickets and conceding minimum runs. This paper attempts to predict the performance of players as how many runs will each batsman score and how many wickets will each bowler take for both the teams. Both the problems are targeted as classification problems where number of runs and number of wickets are classified in different ranges. We used naïve bayes, random forest, multiclass SVM and decision tree classifiers to generate the prediction models for both the problems. Random Forest classifier was found to be the most accurate for both the problems.

Keywords

Naïve Bayes, Random Forest, Multiclass SVM, Decision Trees, Cricket

1. Introduction

Cricket is a sport played by two teams with each side having eleven players. Each team is a right blend of batsmen, bowlers and all-rounders. The batsmen’s role is to score maximum runs possible and the bowlers have to take maximum wickets and restrict the other team from scoring runs at the same time. All rounders are the players who can both bat and bowl and they contribute by scoring runs and taking wickets. Each player contributes towards the overall performance of the team by giving his best performance in each match. Each player’s performance varies with factors like the team he is playing against and the ground at which the match is being played. It is important to select the right players that can perform the best in each match. The performance of a player also depends on several factors like his current form, his performance against a particular team, his performance at a particular venue etc. The team management, the coach and the captain analyze each player’s characteristics, abilities and past stats to select the best playing XI for a given match. In other words, they try to predict the players' performance for each match.

In this paper, we predict the players’ performance in One Day International (ODI) matches by analyzing their characteristics and stats using supervised machine learning techniques. For this, we predict batsmen’s and bowlers’ performance separately as how many runs will a batsman score and how many wickets will a bowler take in a particular match.

Natarajan Meghanathan et al. (Eds) : CCSEA, NCOM, AIFU, DKMP, EMSA, SIPRO, SEA - 2018
pp. 111–126, 2018. © CS & IT-CSCP 2018
DOI : 10.5121/csit.2018.80310
2. LITERATURE REVIEW

An extensive online search produced very few articles related to players’ performance prediction in the game of cricket. A very small number of researchers have studied the performance of cricket players. Muthuswamy and Lam [1] predicted the performance of Indian bowlers against seven international teams against which the Indian cricket team plays most frequently. They used back propagation network and radial basis network function to predict how many runs a bowler is likely to concede and how many wickets a bowler is likely to take in a given ODI match. Wikramasinghe [2] predicted the performance of batsmen in a test series using a hierarchical linear model. Barr and Kantor [3] defined a criterion for comparing and selecting batsmen in limited overs cricket. They defined a new measure \( P(\text{out}) \) i.e. probability of getting out and used a two-dimensional graphical representation with Strike Rate on one axis and \( P(\text{out}) \) on another. Then they define a selection criterion based on \( P(\text{out}) \), strike rate and batting average of the batsmen. Iyer and Sharda [4] used neural networks to predict the performance of players where they classify batsmen and bowlers separately in three categories – performer, moderate and failure. Based on the number of times a player has received different ratings, they recommend if the player should be included in the team to play World Cup 2007. Jhanwar and Paudi [5] predict the outcome of a cricket match by comparing the strengths of the two teams. For this, they measured the performances of individual players of each team. They developed algorithms to model the performances of batsmen and bowlers where they determine the potential of a player by examining his career performance and then his recent performances. Lemmer [6] defined a new measure called Combined Bowling Rate to measure the performance of bowlers. The combined bowling rate is a combination of three traditional bowling measures: bowling average, strike rate and economy. Bhattacharjee and Pahinkar. [7] used this combined bowling rate to analyze the performance of bowlers in Indian Premier League (IPL). They also determined other factors that affect the performance of bowlers and applied multiple regression model to identify the factors that are empirically responsible for the performance of bowlers. Mukharjee. [8] applied Social Network Analysis to rate batsmen and bowlers in a team performance. He generated a directed and weighted network of batsmen-bowlers using player-vs-player information available for test and ODI cricket. He also generated a network of batsmen and bowlers using the dismissal record of batsmen in the history of cricket. Shah [9] also defined new measures to measure players’ performance. The new measure for batsmen takes into account the quality of each bowler he is facing and the new measure for bowlers considers the quality of each batsman he is bowling to. The aggregate of individual performance of a batsman against each bowler is the total performance index of the batsman. Similarly, the aggregate of individual performance of a bowler against each batsman is the total performance index of the bowler. Parker, Burns and Natarajan. [10] defined a model for valuation of players for IPL auction. Their model considered factors like previous bidding price of the player, experience of the player, strike rate etc. Prakash, Patvardhan. and Lakshmi [11] defined batting index and bowling index to rank players’ performance for their models to predict outcomes of IPL matches. Ovens and Bukiet [12] applied a mathematical approach to suggest optimal batting orders for ODI matches.

Our work is probably the first generalized approach to predict how many runs will a batsman score and how many wickets will a player take on a particular match day. Muthuswamyand Lam [1] carried out a similar study predicting how many wickets will a bowler take using neural networks but their work was limited to eight Indian bowlers and is difficult to generalize for all the bowlers in the world. We used some supervised machine learning algorithms to build prediction models that can be used to predict the performance of any player in a given match.
3. DATA AND TOOLS

We obtained all our data from www.cricinfo.com using scraping tools, parsehub [13] and import.io [14]. For batting, we considered matches played from January 14, 2005 to July 10, 2017. The senior most player during this span was SR Tendulkar, so we collected innings by innings list of the performance of all the batsmen from December 18, 1989 when he played his first ODI match. For bowling, we considered matches played from January 2, 2000 to July 10, 2017. The senior most player during this span was PA de Silva, so we collected innings by innings list of the performance of all the batsmen from March 31, 1984 when he played his first ODI match. Since the past stats of the players such as average, strike rate etc. are not available directly online for each match they played, we calculated from the innings by innings list for each match. We imported all the data in MySQL tables and used php to manipulate them.

For predictive analytics, we used Weka and Dataiku. Both these tools are a collection of machine learning algorithms for data mining and also provide some preprocessing functionalities. All the results in this study have been obtained from Weka 3-9-1-oracle-jvm and Dataiku Data Science Studio on Mac OS 10.11.6 and Windows 10.

4. DATA PREPROCESSING

4.1 Calculating the Traditional Attributes

As mentioned in the previous section, the stats of the players such as average, strike rate etc. are not available directly for each game, we calculated these attributes from the innings by innings list using aggregate functions and mathematical formulae. These attributes are generally used to measure a player’s performance. These attributes are as follows:

4.1.1 Batting Attributes

No. of Innings: The number of innings in which the batsman has batted till the day of the match. This attribute signifies the experience of the batsman. The more innings the batsman has played, the more experienced the player is.

Batting Average: Batting average commonly referred to as average is the average number of runs scored per innings. This attribute indicates the run scoring capability of the player.

\[
\text{Average} = \frac{\text{Runs Scored}}{\text{Number of times dismissed}}
\]

Strike Rate (SR): Strike rate is the average number of runs scored per 100 balls faced. This attribute indicates how quickly the batsman can score runs.

\[
\text{Strike Rate} = \left( \frac{\text{Runs Scored}}{\text{Balls Faced}} \right) \times 100
\]

Centuries: Number of innings in which the batsman scored more than 100 runs.

Fifties: Number of innings in which the batsman scored more than 50 (and less than 100) runs.

Zeros: Number of innings in which the batsman was dismissed without scoring a single run.

Highest Score: The highest runs scored by a batsman in any (single) innings throughout his career.
4.1.2 Bowling Attributes

No. of Innings: The number of innings in which the bowler has bowled till the day of the match.

Overs: The number of overs bowled by a bowler.

Bowling Average: Bowling average is the number of runs conceded by a bowler per wicket taken.

\[
\text{Average} = \frac{\text{Number of runs conceded}}{\text{Number of wickets taken}}
\]

Bowling Strike Rate: Bowling strike rate is the number of balls bowled per wicket taken.

\[
\text{Strike Rate} = \frac{\text{Number of balls bowled}}{\text{Number of wickets taken}}
\]

Four/Five Wicket Haul: Number of innings in which the bowler has taken more than four wickets.

4.2 Rating The Attributes

Each traditional attribute is given a rating from 1 to 5 based on its value with 1 being the minimum and 5 being the maximum.

These ratings are as follows:

No. of Innings:

For Consistency:
1 – 49: 1
50 – 99: 2
100 – 124: 3
125 – 149: 4
>=150: 5

For Form:
1 – 4: 1
5 – 9: 2
10 – 11: 3
12 – 14: 4
>=15: 5

For Opposition:
1 – 2: 1
3 – 4: 2
5 – 6: 3
7 – 9: 4
>=10: 5

For Venue:
1: 1
2: 2
3: 3
4: 4
>=5: 5
Batting Average (for all derived attributes):
  0.0 - 9.99: 1
  10.00 - 19.99: 2
  20.00 - 29.99: 3
  30.00 - 39.99: 4
  >=40: 5

Batting Strike Rate (for all derived attributes):
  0.0 - 49.99: 1
  50.00 - 59.99: 2
  60.00 - 79.99: 3
  80.00 - 100.00: 4
  >=100.00: 5

Centuries:
  For Consistency:
    1 – 4: 1
    5 – 9: 2
    10 – 14: 3
    15 – 19: 4
    >=20: 5
  For Form:
    1: 1
    2: 2
    3: 3
    4: 4
    >=5: 5
  For Opposition:
    1: 3
    2: 4
    >=3: 5
  For Venue:
    1: 4
    >=2: 5

Fifties:
  For Consistency:
    1 – 9: 1
    10 – 19: 2
    20 – 29: 3
    30 – 39: 4
    >=40: 5
  For Form & Opposition:
    1 – 2: 1
    3 – 4: 2
    5 – 6: 3
    7 – 9: 4
For Venue:
1: 4
>=2: 5

Zeros:

For Consistency:
1 – 4: 1
5 – 9: 2
10 – 14: 3
15 – 19: 4
>=20: 5

For Form & Opposition:
1: 1
2: 2
3: 3
4: 4
>=5: 5

Highest Score (For Venue Only):
1 – 24: – 1
25 – 49: 2
50 – 99: 3
100 – 150: 4
>=150: 5

Overs:

For Consistency:
1 – 99: 1
100 – 249: 2
250 – 499: 3
500 – 1000: 4
>=1000: 5

For Form & Opposition:
1 – 9: 1
10 – 24: 2
25 – 49: 3
50 – 100: 4
>=100: 5

For Venue:
1 – 9: 1
10 – 19: 2
20 – 29: 3
30 – 39: 4
>=40: 5
Bowling Average (for all derived attributes):
  0.00 - 24.99: 5
  25.00 - 29.99: 4
  30.00 - 34.99: 3
  35.00 - 49.99: 2
  >=50.00: 1

Bowling Strike Rate (for all derived attributes):
  0.00 - 29.99: 5
  30.00 - 39.99: 4
  40.00 - 49.99: 3
  50.00 - 59.99: 2
  >=60.00: 1

Four/Five Wicket Haul:

  For Consistency:
    1 – 2: 3
    3 – 4: 4
    >=5: 5

  For Form, Opposition & Venue:
    1 – 2: 4
    >=3: 5

4.3 Calculating The Derived Attributes

We have derived four attributes for batsmen and bowlers to reflect their abilities and characteristics. These four attributes have been derived by taking a weighted average of the traditional attributes described above. The weights were determined using analytic hierarchy process developed by Thomas L. Saaty [15] [16]. These attributes are described below:

4.3.1 Consistency

This attribute describes how experienced the player is and how consistent he has been throughout his career. All the traditional attributes used in this formula are calculated over the entire career of the player.

Formula for batting:

\[
\text{Consistency} = 0.4262 \times \text{average} + 0.2566 \times \text{no. of innings} + 0.1510 \times \text{SR} + 0.0787 \times \text{Centuries} + 0.0556 \times \text{Fifties} - 0.0328 \times \text{Zeros}
\]

Formula for bowling:

\[
\text{Consistency} = 0.4174 \times \text{no. of overs} + 0.2634 \times \text{no. of innings} + 0.1602 \times \text{SR} + 0.0975 \times \text{average} + 0.0615 \times \text{FF}
\]

4.3.2 Form

Form of a player describes his performance over last one year. All the traditional attributes used in this formula are calculated over the matches played by the player in last 12 months from the day of the match.
Formula for batting:

\[ \text{Form} = 0.4262 \times \text{average} + 0.2566 \times \text{no. of innings} + 0.1510 \times \text{SR} + 0.0787 \times \text{Centuries} + 0.0556 \times \text{Fifties} - 0.0328 \times \text{Zeros} \]

Formula for bowling:

\[ \text{Form} = 0.3269 \times \text{no. of overs} + 0.2846 \times \text{no. of innings} + 0.1877 \times \text{SR} + 0.1210 \times \text{average} + 0.0798 \times \text{FF} \]

4.3.3 Opposition

Opposition describes a player’s performance against a particular team. All the traditional attributes used in this formula are calculated over all the matches played by the player against the opposition team in his entire career till the day of the match.

Formula for batting:

\[ \text{Opposition} = 0.4262 \times \text{average} + 0.2566 \times \text{no. of innings} + 0.1510 \times \text{SR} + 0.0787 \times \text{Centuries} + 0.0556 \times \text{Fifties} - 0.0328 \times \text{Zeros} \]

Formula for bowling:

\[ \text{Opposition} = 0.3177 \times \text{no. of overs} + 0.3177 \times \text{no. of innings} + 0.1933 \times \text{SR} + 0.1465 \times \text{average} + 0.0943 \times \text{FF} \]

4.3.4 Venue

Venue describes a player’s performance at a particular venue. All the traditional attributes used in this formula are calculated over all the matches played by the player at the venue in his entire career till the day of the match.

Formula for batting:

\[ \text{Venue} = 0.4262 \times \text{average} + 0.2566 \times \text{no. of innings} + 0.1510 \times \text{SR} + 0.0787 \times \text{Centuries} + 0.0556 \times \text{Fifties} + 0.0328 \times \text{HS} \]

Formula for bowling:

\[ \text{Venue} = 0.3018 \times \text{no. of overs} + 0.2783 \times \text{no. of innings} + 0.1836 \times \text{SR} + 0.1391 \times \text{average} + 0.0972 \times \text{FF} \]

4.3.5 Other Input Attributes

Batting Hand: The dominant hand of the batsman while batting. It has two possible values: Left or Right.

Bowling Hand: The dominant hand of the batsman while bowling. It has two possible values: Left or Right.

Batting Position: The number at which the batsman bats in the batting order.

Match Type: The type of the match. This attribute has four possible values: Normal, quarter-final, semi-final or final.

Match Time: The time at which the match is played. There are two possible values: Day or Day-night.
Strength of opposition: This is the batting/bowling strength of the opposition team. It is the average of the consistency measure of the batsmen/bowlers of the opposition team.

Ven: The relative venue for the teams. It has three possible values: Home, Away or Neutral.
Oppo: The opposition team.

The following attributes are used only for predicting runs.

Role: The playing role of the player. It can take following values:

- Opening Batsman (OBT) – The two batsmen who usually bat at position one or two are called opening batsmen.
- Top Order Batsman (TOB) – The batsmen who usually bat at position three or five are called top order batsmen.
- Middle Order Batsman (MOB) - The batsmen who usually bat at position five to eight are called middle order batsmen.
- Batsman – The batsmen who usually bat at different positions are categorised simply as batsmen here.
- Allrounder – The players who are equally skilled at both batting and bowling are called allrounders.
- Batting Allrounder – The players who can both bat and bowl but are more skilled at batting than bowling, are called batting allrounders.
- Bowling Allrounder – The players who can both bat and bowl but are more skilled at bowling than batting, are called bowling allrounders.
- Bowler – The players who are expert bowlers but not so skilled at batting, are categorised as bowlers.

Captain: This is a binary attribute indicating whether a player is captain of the team.

WK: This is a binary attribute indicating whether a player is a wicketkeeper.

Innings: This attributes indicates if it is the first or the second innings of the match.

Tournament: The type of tournament in which the match is being played. It can take the following values:

- Two Team Tournament (TT)
- Three-Four Team Tournament (TFT)
- Five Team Tournament (FT)

Toss: Indicates whether the player’s team won or lost the toss.
Pressure: Indicates mental and psychological pressure on the player. It takes values from 1 to 5. Its value depends on the type of match being played and the teams that are playing the match. The values are defined as follows:

- Normal matches: 1
- Quarter Finals: 3
- Semi Finals: 4
- Finals: 5

Above values are incremented by 1 if the match is between India and Pakistan or Australia and England as these countries are strong rivals of each other.

Host: The country in which the match is being played.

Ground: The ground on which the match is being played.

4.3.6 Outputs

Both the problems are treated as classification problems.

Runs are predicted in five classes:

- 1 – 24: 1
- 25 – 49: 2
- 50 – 74: 3
- 75 – 99: 4
- >=100: 5

Wickets are predicted in three classes:

- 0 - 1: 1
- 2 – 3: 2
- >=4: 3

4.3.7 Data Cleaning

A large number of values of Opposition and Venue were zero. This is because a player has not played any match against a particular team or at a venue before the day of play. We treated such values as missing values and replaced them with the class average of corresponding attributes.

4.3.8 Class Imbalance

We observed that majority of the records fall within class 1 in both batting and bowling. This created a major imbalance in the distribution of values and affected the performance of the learning algorithms. To solve this problem, we applied an oversampling technique Supervised Minority Oversampling Technique (SMOTE) [17] on minority classes to make all the classes equally distributed. SMOTE over-samples minority classes by creating synthetic example tuples [17].
5. LEARNING ALGORITHMS

For generating the prediction models, we used supervised machine learning algorithms. In supervised learning algorithms, each training tuple is labeled with the class to which it belongs [18]. We used naïve bayes, decision trees, random forest and multiclass support vector machines for our experiments. These algorithms are explained in brief.

5.1 Naïve Bayes

Bayesian classifiers are statistical classifiers that predict the probability with which a given tuple belongs to a particular class [18]. Naïve Bayes classifier assumes that each attribute has its own individual effect on the class label, independent of the values of other attributes. This is called class-conditional independence. Bayesian classifiers are based on Bayes’ theorem.

Bayes Theorem: Let X be a data tuple and C be a class label. Let X belongs to class C, then

\[
P(C|X) = \frac{P(X|C)P(C)}{P(X)}
\]

where;

- P(C|X) is the posterior probability of class C given predictor X.
- P(C) is the prior probability of class.
- P(X|C) is the posterior probability of X given the class C.
- P(X) is the prior probability of predictor.

The classifier calculates P(C|X) for every class Ci for a given tuple X. It will then predict that X belongs to the class having the highest posterior probability, conditioned on X. That is X belongs to class Ci if and only if

\[
P(C_i|X) > P(C_j|X) \quad \text{for} \quad 1 \leq j \leq m, j \neq i.
\]

5.2 Decision Trees

Decision tree induction is the process of creating decision trees for class-labeled training tuples [18]. A decision tree is basically a tree structure like a flowchart [18]. Each internal node of the tree represents a test on an attribute and each branch is the outcome of the test. Each leaf node is a class label. The first node at the top of the tree is the root node. To classify a given tuple X, the attributes of the tuple are tested against the decision tree starting from the root node to the leaf node which holds the class prediction of the tuple. Ross Quinlan introduced a decision tree algorithm called ID3 in his paper [19]. Later he introduced a successor of ID3 called C4.5 in [20] to overcome some shortcomings such as over-fitting. Unlike ID3, C4.5 can handle both continuous and discrete attributes, training data with missing values and attributes with differing costs. In a basic decision tree induction algorithm, all the training tuples are at the root node at start. The tuples are then partitioned recursively based on selected attributes. The attributes are selected based on an attribute selection method which specifies a heuristic procedure to determine the splitting criterion. The algorithm terminates if all the training tuples belong to the same class or there are no remaining attributes for further partitioning or all training tuples are used. ID3 uses the attribute selection measure called information gain, which is simply the difference of the information needed to classify a tuple and the information needed after the split. These two can be formulated as follows:
Expected information needed to classify a tuple in the training set $D$

$$Info(D) = - \sum_{i=1}^{m} p_i \log_2(p_i)$$

where $p_i$ is the nonzero probability that a tuple in $D$ belongs to class $C_i$.

Information needed after the splitting (to arrive at the exact classification)

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

where $A$ is the attribute on which the tuples are to be partitioned.

Then, information gain

$$Gain(A) = Info(D) - Info_A(D)$$

The attribute with the highest information gain is selected as the splitting attribute.

C4.5 uses gain ratio as the attribute selection measure. Gain ratio is an extension to information gain in a sense because it normalizes information gain by using a split information value;

$$SplitInfo_A(D) = - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \left( \frac{|D_j|}{|D|} \right)$$

Then,

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)}$$

The attribute with the highest gain ratio is selected as the splitting attribute.

### 5.3 Random Forest

Random Forests is an ensemble method for classification and regression [18]. Random forests are a set of decision trees where each tree is dependent on a random vector sampled independently and with the same distribution of all the trees in the forest [21]. The algorithm generates a number of decision trees creating a forest. Each decision tree is generated by selecting random attributes at each node to determine the split [18]. Tim Kam Ho introduced the first method for random forests using random subspace method in his paper [22]. Later, Breiman Leo extended the algorithm in his paper [21] and this method was official known as Random Forests. The general procedure to generate decision trees for random forests starts with a dataset $D$ of $d$ tuples. To generate $k$ decision trees from the dataset, for each iteration $k$, a training set $D_i$ of $d$ tuples is sampled with replacement from the dataset $D$. To construct a decision tree classifier, at each node, a small number of attributes from the available attributes are selected randomly as candidates for the split at the node. Then Classification And Regression Trees (CART) [23] method is used to grow the trees. The trees are then grown to maximum size and are not pruned. CART is a non-parametric decision tree induction technique that can generate classification and regression trees. CART recursively selects rules based on variables’ values to get the best split. It stops splitting when it detects that no further gain can be made or some pre-determined stopping conditions are met.
5.4 Support Vector Machine

Vladimir Vapnik, Bernhard Boser and Isabell Guyon introduced the concept of support vector machine in their paper [24]. SVMs are highly accurate and less prone to overfitting. SVMs can be used for both numeric prediction and classification. SVM transforms the original data into a higher dimension using a nonlinear mapping. It then searches for a linear optimal hyperplane in this new dimension separating the tuples of one class from another. With an appropriate mapping to a sufficiently high dimension, tuples from two classes can always be separated by a hyperplane. The algorithm finds this hyperplane using support vectors and margins defined by the support vectors. The support vectors found by the algorithm provide a compact description of the learned prediction model. A separating hyperplane can be written as:

\[ W \cdot X + b = 0 \]

where \( W \) is a weight vector, \( W = \{w_1, w_2, w_3, ..., w_n\} \), \( n \) is the number of attributes and \( b \) is a scalar often referred to as a bias. If we input two attributes \( A_1 \) and \( A_2 \), training tuples are 2-D, (e.g., \( X = (x_1, x_2) \)), where \( x_1 \) and \( x_2 \) are the values of attributes \( A_1 \) and \( A_2 \), respectively. Thus, any points above the separating hyperplane belong to Class \( A_1 \):

\[ W \cdot X + b > 0 \]

and any points below the separating hyperplane belong to Class \( A_2 \):

\[ W \cdot X + b < 0 \]

SVM was originally used for binary classification. However, several multiclass SVM algorithms have also been developed. In Weka, we used LIBSVM package developed by Chih-Chung Chang and Chih-Jen Lin [25]. The package can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm. LIBSVM is an easy to use package to apply multiclass SVM and has gained a wide popularity in machine learning.

6. RESULTS AND DISCUSSIONS

We used different sizes of training and test sets to find the best combination that gives the most accuracy. We used four machine learning algorithms: Naïve Bayes, Decision Trees, Random Forest and Support Vector Machine in our experiments. The results are tabulated below. Table 1 shows the accuracies of the algorithms for predicting runs and table 2 shows the accuracies of the algorithms for predicting wickets.

As we can see, Random Forest builds the most accurate prediction models for both batting and bowling in all the cases. Also, the accuracy of the models increases as we increase the size of the training dataset for all algorithms except in case of Naïve Bayes for batting where the accuracy decreases as we increase the size of the training set. Random Forest predicts runs with the highest accuracy of 90.74% when we use 90% of the dataset for training. Similarly, Random Forest predicts wickets with highest accuracy of 92.25% when we use 90% of the dataset for training. On the other hand, Naïve Bayes predicts runs with the least accuracy of 42.5% when we use 90% of the dataset for training. Naïve Bayes predicts wickets too with the least accuracy of 57.05% when we use 60% of the dataset for training. Decision Trees performs reasonably well with the maximum accuracy of 80.46% and the minimum accuracy of 77.93% for predicting runs. It predicts wickets with the maximum accuracy of 86.5% and the minimum accuracy of 84.4%, which is again reasonably well against the performance of Random Forest. The prediction models of SVM for predicting runs showed the maximum accuracy of 51.45% with 90% training data and the minimum accuracy of 50.54% with 60% training data. Also for wickets, SVM had the maximum accuracy of 68.78% with 90% training data and the minimum accuracy of 67.45% with
60% training data. Thus, surprisingly SVM was beaten by Random Forest and Decision Trees on both the datasets.

Table 1 Predicting runs

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy (%)</th>
<th>60% train 40% test</th>
<th>70% train 30% test</th>
<th>80% train 20% test</th>
<th>90% train 10% test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>43.08</td>
<td>42.95</td>
<td>42.47</td>
<td>42.50</td>
<td></td>
</tr>
<tr>
<td>Decision Trees</td>
<td>77.93</td>
<td>79.02</td>
<td>79.38</td>
<td>80.46</td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>89.92</td>
<td>90.27</td>
<td>90.67</td>
<td>90.74</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>50.54</td>
<td>50.85</td>
<td>50.88</td>
<td>51.45</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Predicting wickets

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy (%)</th>
<th>60% train 40% test</th>
<th>70% train 30% test</th>
<th>80% train 20% test</th>
<th>90% train 10% test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>57.05</td>
<td>57.18</td>
<td>57.48</td>
<td>58.12</td>
<td></td>
</tr>
<tr>
<td>Decision Trees</td>
<td>84.40</td>
<td>85.12</td>
<td>85.99</td>
<td>86.50</td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>90.68</td>
<td>91.26</td>
<td>91.80</td>
<td>92.25</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>67.45</td>
<td>67.53</td>
<td>68.35</td>
<td>68.78</td>
<td></td>
</tr>
</tbody>
</table>

Muthuswamy and Lam [1] achieved an accuracy of 87.10% with backpropagation network (BPN) and 91.43% with radial basis function network (RBFN) table 3. They predicted wickets in two classes: 0 or 1 and 2 or more. Their study was limited only to eight Indian bowlers who played ODI matches for India since year 2000 against seven countries. Our approach works for any player in the world and is also applicable to new players who will be playing for their country in future. Moreover, they considered only three input parameters: An ID assigned manually to a bowler, an ID assigned manually to the opposition team and the number of overs bowled by the bowler. We carried out a detailed study of the stats and characteristics of the players by considering a number of attributes that can potentially impact the performance of a player in any match.

Table 3 Predicting wickets using BPN and RBFN [1]

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPN</td>
<td>87.10</td>
</tr>
<tr>
<td>RBFN</td>
<td>91.43</td>
</tr>
</tbody>
</table>

7. CONCLUSION AND FUTURE WORK

Selection of the right players for each match plays a significant role in a team’s victory. An accurate prediction of how many runs a batsman is likely to score and how many wickets a bowler is likely to take in a match will help the team management select best players for each match. In this paper, we modeled batting and bowling datasets based on players’ stats and characteristics.
Some other features that affect players’ performance such as weather or the nature of the wicket could not be included in this study due to unavailability of data. Four multiclass classification algorithms were used and compared. Random Forest turned out to be the most accurate classifier for both the datasets with an accuracy of 90.74% for predicting runs scored by a batsman and 92.25% for predicting wickets taken by a bowler. Results of SVM were surprising as it achieved an accuracy of just 51.45% for predicting runs and 68.78% for predicting wickets.

Similar studies can be carried out for other formats of the game i.e. test cricket and T20 matches. The models for these formats can be shaped to reflect required characteristics of the players; e.g. batsmen need to have patience and ability to play longer innings in test matches whereas score more runs in less overs in T20 matches. Similarly, bowlers need to have stronger wicket taking abilities in test matches and better economy rate i.e. conceding less runs in T20 matches. Moreover, attempts can be made to improve accuracies of the classifiers for ODI matches.

REFERENCES


AUTHORS

Kalpdrum Passi received his Ph.D. in Parallel Numerical Algorithms from Indian Institute of Technology, Delhi, India in 1993. He is an Associate Professor, Department of Mathematics & Computer Science, at Laurentian University, Ontario, Canada. He has published many papers on Parallel Numerical Algorithms in international journals and conferences. He has collaborative work with faculty in Canada and US and the work was tested on the CRAY XMP's and CRAY YMP's. He transitioned his research to web technology, and more recently has been involved in machine learning and data mining applications in bioinformatics, social media and other data science areas. He obtained funding from NSERC and Laurentian University for his research. He is a member of the ACM and IEEE Computer Society.
IMPLEMENTING UHF RFID READER ON SMARTPHONE PLATFORM FOR IOT SENSING

Douglas Lautner, Xiayu Hua, Scott DeBates and Shangping Ren

Department of Computer Science, Illinois Institute of Technology,
Chicago, IL 60616, USA

Device+ Research and Technology, Lenovo Motorola,
Chicago, IL 60654, USA

ABSTRACT

As a core component of the Internet of Things technology (IoT), Radio Frequency Identification (RFID) tagged items will add billions, perhaps trillions, of objects to the Internet. As a result, uses of Ultra High Frequency (UHF) RFID sensing become massive ranging from logistics, retail and healthcare to homes and even entire smart cities. Under this trend, mobile UHF RFID scanners also need to evolve. Consumers will interact with their surroundings via tagged RFID items taking full advantage of the advancing IoT. For mainstream consumer smartphones, unfortunately, UHF RFID connectivity has yet to be fully integrated. The major challenges are: 1) the compatibility of an RFID reader module to the host platform, 2) Radio Frequency (RF) signal coexistence interference between the RFID reader and other sensor/RF technologies, and 3) the unacceptable high current drain caused by RFID active scanning. In this paper, we present a design and implementation of a novel modular UHF RFID scanning subsystem, the UHF RFID reader module, on a Motorola Moto-Z smartphone. This module is fully integrated with an Android 7.0 Operating System (OS) and directly interconnects with the low-level smartphone hardware and software framework. With the new antenna design and the signal spectrum analysis, we guarantee the RF isolation of the Mod with the smartphone’s other native wireless components and sensors. Our design and implementation also address the current drain issue and extends the battery life of Moto-Z smartphone up to 30.4 hours with IoT RFID scanning.

KEYWORDS

UHF RFID, Smartphone design, mobile system architecture, mobile sensing

This work is supported by NSF CAREER 381211; NSF CNS 381367; and NSF CPS 381396.

1. INTRODUCTION

The advance of the Internet of Things (IoT) technology introduces new application scenarios to smartphone designers and manufactures. For example, using smartphones to gather information from nearby IoT objects through RFID and providing surrounding status to the user is one of the potentials. Thanks to the massive production of UHF RFID tags (from 8.9 billion tags sold in
2015 to 10.4 billion in 2016 [1]), its price becomes affordable [2], hence transforming common everyday items of a wide variety, such as jeans or brushes, into IoT objects [3]. Likewise, smartphone shipments have increased every year since 2006 with an estimated global market of 2.1 billion units shipped in 2021 [4]. With high-performance computing power and constant Internet connectivity, nowadays smartphones are already capable to associate mobile computing with the IoT [5]. Once equipped with a UHF RFID reader, a smartphone is able to interact with the surrounding objects tagged by UHF RFID, gather and process information from those tags locally or on the internet, then represent feedback to the user via its multimedia user interfaces.

However, to-date there has been no successful widespread deployment of a fully integrated UHF RFID reader on mainstream smartphones. Common UHF RFID reader solutions are either stand-alone devices or are deployed as add-on implementation on smartphones that are not fully integrated with the hardware and software of a smartphone. For example, a UHF RFID reader on microSD which can be inserted into a smartphone through external memory slot was developed in [6]. Another solution utilizes the proprietary audio port to connect a UHF RFID reader with a smartphone [7]. Current integrated solutions, such as Invengo XC-1003 mobile IoT device and Chafon CF-H802 run on older Android platforms, are more industrial and are not widely distributed [8], [9]. Moreover, since hardware and software of these devices are specially designed, they are expensive and their functionality are monotonous. Through the research and prototyping experience in our preliminary works, we have identified three major challenges that causes this absence of RFID reader on a smartphone:

1) No existing interface that enables the UHF RFID reader to establish a direct high-speed connection to the lower level smartphone hardware without significant modifications on the phones hardware framework.

2) Severe signal interferences amongst different RF sensors and communication modules while the RFID reader antenna is active.

3) Unbearable current drain to a consumer smartphone when the UHF RFID reader is up and running.

In this work, we address the three challenges and empower a commercial smartphone with UHF RFID reading capability. Specifically, we design and implement a fully functional consumer UHF RFID Reader Mod (URRM), as illustrated in Fig. 1, based on a mainstream Android-driven smartphone, Motorola Moto-Z, which has been "sold around the world" [10]. By leveraging the advantage of the unique modularization features of Moto-Z, i.e. the support of third-party Mods [11], we are able to utilize the GreyBus protocol to connect our RFID reader module, i.e. the Impinj RS500 RFID module, to the phones lower-level hardware framework without changing the phones native hardware. Comparing to add-on implementations that connect through USB, audio jack or memory card ports [9,10], direct connection to lower-level hardware framework provides with high energy-efficiency, low communication overhead as well as wider applicable system support such as power management or chip-level control and scheduling. To guarantee URRMs RF performance and avoid interference with phones native wireless functionalities, we design an antenna for the URRM which isolates to the phones native wireless signal system. We further utilize the power-efficient sensor fusion core on the URRMs hardware infrastructure to manage the scanning and data processing tasks. We minimize the impact of the high current drain of RFID scanning to the phones battery life by optimizing the control and management system and coupling to an external power source. Our experiments prove that our prototype overcomes
these three challenges and hence becomes practical and ready-to-manufacture. In addition, our design and implementation of the key hardware and software architecture is extensible and scalable for continued research. Other IoT sensors on different mobile platforms, such as WiFi or Bluetooth modules on NuttX-driven Raspberry Pi platforms, can also apply our solutions.

The rest of the paper is organized as follows. We discuss related work in Section II. Section III presents the design and implementation of the portable UHF RFID reader architecture. In Section IV we study the radio performance of the UHF RFID reader Mod with respect to signal isolation, sensing range and sensitivity. We conclude in Section V.

2. RELATED WORK

The fast growing IoT technology and its expanding scope of applications demand for a wide spectrum of both the sensor type and the sensing content [12]. In [13], researchers developed CARISMA (Context-Aware Reflective Middleware System for Mobile Applications), a software framework for IoT contextual sensing and awareness tasks on mobile systems. This framework is generic with respect to mobile operation systems and sensor types, hence it can be implemented on different smartphone platforms and work with different types of sensors. Similar middleware system, the SOCAM (Service Oriented Context-Aware Middleware), was also developed by researchers in [14]. To utilize the advantages of heterogeneous sensing sources in IoT scheme, the e-SENSE [15] mobile sensing system enables ambient intelligence using wireless multi-sensor networks for making IoT oriented context-rich information available to applications and services. e-SENSE combines body sensor networks (BSN), object sensor networks (OSN), and environment sensor networks (ESN) to capture context in the IoT paradigm. Later, Hydra3 [16] was developed as an IoT middleware that aims to integrate wireless devices and sensors into ambient intelligence systems. Hydra3 identifies context reasoning rule engine, context storage, context querying, and event/action management as the key components of a context-aware framework. In the IoT sensing scheme, RFID reader plays a key role. In [17], researchers illustrate eleven different categories of use cases that need to utilize RFID technology to interact with the IoT surroundings. In [18], researchers introduce a semi-passive, reconfigurable UHF RF identification (RFID) sensing tag operating as the generic sensing platform (GSP). The tag is highly configurable and can be dynamically switched between a Continuous data transmitting platform (Online mode) or a Data logging platform (Offline mode). Researchers in [19] develop a suite of web-based, user-level tools and applications designed to empower users by facilitating their understanding, management, and control of personal RFID data and privacy settings. Besides researches on RFID technology itself, research topics of implementing RFID based IoT systems on other application fields, such as massive manufacture workflow optimization [20], Geofencing item tracking [21] and industrial-level quality control [22], have drawn increasing attention not just from computer developers and researchers, but from all the industrial and science community. IoT is advancing fast, RFID and their combined technologies are also becoming ubiquitous and playing a critical role of identifying objects. However, smartphones, one of the most convenient and widely used portable computing devices, is still out of the picture of research on RFID based IoT systems. Although it matches all the requirements from both the hardware and the software perspective, severe RFID integration challenges remain.
3. SYSTEM DESIGN AND IMPLEMENTATION

The core concept of our system design is the URRM and its control and support components. In particular, our system architecture design contains five subsystems: the UHF RFID reader Mod (URRM) subsystem, the Mod control and management subsystem, the Mod support subsystem, the Battery subsystem and the Antenna subsystem. The following subsections discuss the five subsystems in details, followed by a discussion of the system workflow.

A. UHF RFID Reader Mod Subsystem

To embed a RFID reader module into the low-level of a native smartphone system, the major challenge is to establish a physical connection from the RFID module directly to the low-level smartphone system without significantly modifying the native system. We achieve this goal by designing the URRM for a Moto Z smartphone, both of which are showed in Fig. 1, respectively. In particular, we embedded the RFID reader module into the MotoMod platform [23] with a HAT (Hardware Attached on Top) adapter board [24]. Due to the modularization design of the Moto-Z mobile system, the URRM utilizes GreyBus [25] protocol to communicate the Moto Zs hardware layer through Serial Peripheral Interface (SPI) once the URRM is attached to a Moto Zs backboard. The picture and the brief architecture of the Moto-Z system and the URRM are depicted in Fig.1(a) and Fig.1(b), respectively.

GreyBus is developed by Google and Motorola in 2014, it supports hot plug and play communications between the GreyBus Master and GreyBus Slave. The Moto Zs Mod physical hardware interface contains 16 metal touch points which supports USB2.0, USB3.1, SPI, Mobility Display Port (MyDP), Inter-IC Sound (I2S) and other mainstream communication protocols. Up to 15 Mps of high speed data exchange with a dedicated transmit and receive Direct Memory Accesses (DMA) between Master and Slave devices. To drive up the reader module, we use Universal Synchronous/Asynchronous Receiver/Transmitter (UsART) to port the reader Mod to the HAT board. To improve the energy efficiency, we further design a two-layer architecture which utilizes both the MuC on the MotoMod, i.e., the Mod Control and Management Subsystem, and the Application Processor (AP) on the Moto-Z, i.e., the Mod Support subsystem, to control and manage the RFID reader module jointly. Moto-Z smartphone system is equipped with a
SnapDragon 820 CPU as its AP which has 2.2GHz main frequency and 512 KB + 1 MB L2 cache. It is designed to support the latest version of Android system, i.e., Android 8.0, and smartphone applications. However, the current drain is high (28.8mA in full performance and 4.8mA in idle state). For scan-related tasks such as conducting IoT sensing or processing the raw data, the AP is over-powerful and hence becomes very energy-inefficient. The MuC on the MotoMod, on the other hand, has higher energy-efficiency (4.3mA current drain in full performance and 1.1uA in idle state). This system is powered by a STM-32L4 based core that has 80MHz main frequency and 1MB flash memory. It’s computing capacity is less powerful comparing to the AP, but it is sufficient to handle the scanning routings such as sending commands or parsing events to or from the RFID module. In our stress test, the MuC can process no less than 1200 RFID tags per second without depleting its computing power and buffer. In our design, we keep computing and control oriented tasks of UHF RFID sensing, such as the scanning routings or raw data process procedures, on the MCU and create the Mod Control and Management subsystem. For tasks that can only run on Moto Z, such as the scan application with Graphic User Interface (GUI), we shift them to the AP and create Mod support subsystem. Both subsystems are discussed in the following two subsections.

B. Mod Control and Management Subsystem

The MuC and its baseline software framework [26] are the core components of the MotoMod system and provide Mod developers a POSIX and ANSI standards develop environment. The software framework is built upon a real-time operating system, the NuttX Real-time operation system (NuttX) [27], to provide basic I/O control, task scheduler, power management and drivers for specific circuits running on the MotoMod. In particular, the slave-end GreyBus protocol is implemented in the NuttX driver layer and is abstracted as a virtual file. All the data transmissions between Moto Z and MotoMod are implemented through reading or writing this virtual file. The system architecture is illustrated in Fig.1(b).
To control the RFID reader module, we deploy the sensor fusion architecture modified from in our previous works [28], [29], which is illustrated in Fig.2, on the baseline software framework. The core components of the sensor fusion architecture are: Power Manager, Sensor Manager, Algorithm Manager, Application Processor Interface and Wireless Manager. We specify their functions separately.

Power Manager: responsible for managing the tasks power mode votes and configuring the MuC’s power mode based on the votes. The power manager also controls and monitors the status of the Mods battery.

Sensor Manager: responsible for configuring the sensors and managing the raw sensor data obtained from the sensors.

Algorithm Manager: responsible for configuring, scheduling, and providing the raw sensor data for all data processing algorithms.

Application Processor Interface: responsible for all communications between the sensor fusion core and the application processor. Its tasks include managing features turned on/off by the application processor, and interrupting applications when sensor data is available.

Wireless Manager: responsible for configuring the UHF RFID reader IC, and write/reading data of surrounding UHF RFID tags.

The design of the Mod Control and Management subsystem has two advantages. First, the architecture is extensible which means not only the RFID reader, but other similar wireless technologies, such as Bluetooth, WiFi, and NFC, can be deployed on the Mod and can work jointly following the same approach. Second, all the drivers and implementation details are encapsulated on the MuC and are transparent to the smartphone, which means even a URRM with different configurations, like with a different RFID reader module, is attached on, systems on the smartphone are not affected.

C. Mod Support Subsystem

On the smartphone side, we design a vertical system framework running on Android system which communicates through from the top (Application) layer of the Android system to the bottom (GreyBus protocol over the SPI transport) layer of the smartphone system. All functionalities of the whole framework are encapsulated and embedded with an API library, the Motlib, and are exposed to the Androids application layer. Our application, as well as any other third-party applications, utilizes the framework by simply including the Motlib library.

The architecture of the vertical system, including the Mod support subsystem and Mod control and management subsystem, are illustrated in Fig.3. For the Mod support subsystem, each component in this framework from top layer to bottom layer is discussed as follows:

RFID Application an Android application with a GUI that allows users to interact with the URRM by sending commands and displaying the reader Mod status and scan results. This application utilizes functions in the Motlib library to interact with the Mod Manager layer.
Mod Manager: The Mod Manager is a component of the Hardware Abstract Layer (HAL) in the NuttX system. It encapsulates the implementation details of hardware under NuttX and maps system calls from the Application Layer to the device driver functionalities. In our case, the URRM is registered to the Mod Manager with its all functionalities. Plus, Mod Manager can directly update the firmware of the MuC in URRM.

![Architecture of URRM and Mod Support Subsystems](image)

GreyBus Stack (master end): GreyBus is running on the Kernel layer of the Android system. GreyBus defines multiple class protocols that are used to communicate and provide function to specific hardware. In our design, we enable three protocols: RAW, Battery and Power Transfer Protocol (PTP) which enable us to use our own designed communication package format (RAW), enable URRM to manage its own battery (Battery) and enable Moto Z can provide power to URRM (PTP).

SPI Driver: SPI Driver is running in the Android kernel. It controls the SPI Bus which is the physical communication interface between the MotoMod and Moto-Z mobile device for GreyBus communications to the MuC.

This design has two advantages. First, it allows Android applications to flash the MuC’s firmware, either for driver updates or for upgrades of functionality, algorithms or even for the baseline system. This feature brings more flexibility and convenience to the application developer since without this function, such update or upgrade procedures can only be done manually by opening up the Mods enclosure and flash new firmware through J-tag or CC-debugger that is attached to the MuC. Second, the power management function in the Mod Support subsystem allows us to embed the external power source in the URRM and enable URRM to gain power supply from either the external power source or the Moto Z. The detailed implementation of the external power source is discussed in the next subsection.
D. Battery Subsystem

In our design, the RFID Mods MuC is responsible for controlling the power management. We use a MAX17050X chipset for monitoring the battery when in-use and charging. To maximize device power when RFID scanning is active, we couple a regenerative 3.7 volt battery with a capacity of 1500 mAH in the RFID Mod with the Moto Z through the MuC controller. Total energy capacity available is 5000 mAH when combined with Moto Z’s native battery capacity of 3500mAH. Utilizing the MuC’s power management component, our configuration uses the Mods battery until the battery level reaches 5%, at that point the MuC will switch in the Moto Zs battery for continuous use ensuring no degradation in performance.

<table>
<thead>
<tr>
<th>UR Scanning mode</th>
<th>Measured Current Drain</th>
<th>Time (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background Scanning (Sensor Fusion Core only)</td>
<td>291mA</td>
<td>1824</td>
</tr>
<tr>
<td>Foreground scanning (application running and display on)</td>
<td>388mA</td>
<td>1422</td>
</tr>
</tbody>
</table>

Fig. 4. Battery Life Evaluation

The method we use to measure the battery consumption is a NI USB-6251 DAQ to monitor current consumption at the battery contacts of the Moto Z. Once the DAQ is connected, we place the phone into an airplane mode to disable all wireless processing, and measure the current consumption with the UHF RFID application launched but not on. This provides us the idle screen current drain baseline measurement. We then enable the UHF RFID reader Mod. The configuration of the scanning is 10sec on and 10sec idle. The results in Fig.4 show that our design is capable of scanning for an entire day in the background. As we continue to develop and refine existing algorithms we anticipate further reductions in battery consumption.

E. Antenna Subsystem

When active, the reader Mod actively project energy to the RFID receiver, such as RFID tags, and capture their echoes as responses. Therefore, the performance of the reader as well as the interference amongst other wireless communication modules could be severe and becomes one of the main blockage for the deployment of RFID readers on smartphone platform. We provide our antenna design here and will evaluate the antennas performance in the next section.

We design the antenna to operate at 902MHz-928MHz, and an ambient distance of one meter for small passive tags, and three meters for active, or larger UHF RFID tags. We use Rogers 3010 material [30] to reduce the effects of increasing the Z height in our design. The antenna dimensions are 80x52x1.2mm. The feed impedance of the antenna was 50 Ohms. The direction of the antenna’s radiation is perpendicular to the Moto Z’s plane, and the polarization is across the Moto Z’s antennas. The isolation between the UHF RFID antenna and the Moto Z’s communication frequencies is larger than 30dB. The radiation efficiency of our design from 902MHz-928MHz was 5 - 10%.
F. System Workflow

Once the UHF RFID reader Mod is attached to the back of the Moto Z, the Mod transmits its manifest file from the MuC to the Motorola Mod Manager. The Motorola Mod Manager then parses the manifest file and determines what type of Mod is connected and its communication capabilities to expose for applications use. After the reader Mod is registered, applications can use Motorola Mod Manager to send RFID commands over the RAW protocol to the Mod. Once the reader Mod receives the RAW packet it will parse the data and determine which Manager (Sensor, Algorithm, Wireless) gets the payload. The Manager will then determine which driver or algorithm to send the payload. The driver or algorithm then parses the payload to perform the requested action. Upon a completion of the requested action a returned status (success or failure) is returned to the application.

4. URRMS SIGNAL FREQUENCY ISOLATION ANALYSIS

As a part of the smartphone wireless system, the reader Mod works in a very close range with other native sensors and wireless modules. Therefore, the radio signal interference becomes a key feature which determines if the reader Mod is practical for commercial use or is just a prototype. Moreover, the radio performance of the reader Mod is another key feature as a sensor that needs to be evaluated, too short sensor range or too low sensitivity may significantly reduce the applicable range of the reader Mod in real-world scenarios. In the following subsections, we experimentally evaluate these two key features.

A. RF Signal Interference Validation

In Section I, we discuss about the antenna design in the hardware level that isolates the RFID signal from other sensors. In specific, as antennas of different wireless components are tightly coupled in a small enclosure, we need to confirm that our new RFID reader has no 1) Radiated spurious emissions [31] and 2) Inter-modulation distortion [32], [33]) against the components in the original system. In this section, we experimentally proof the signal isolation effectiveness of our design from the radio spectrum perspective.

Our experiment testbed is setup as follow. We use a RF shielded box (Will Technology model SH-120) to provide a signal-free environment and use two developer version Moto Z smartphones to perform data transmission via different wireless modules. One smartphone equips with the RFID reader Mod. A spectrum analyzer (Rohde & Schwarz model FSP13) is connected to the RF shield box to monitor and record the RF signals in the box from frequency 13Mhz to 6Ghz. Another antenna is wired through the shield and is implanted into the enclosure to provide cellular signal. The experiment testbed design is shown in Fig.5(b) and its setup illustrated in Fig.5(b).

To perform the interference test on a wireless module or sensor, we take the following steps. First, we place RFID tags into the RF shielded enclosure. Then, we engage the RF modules on both smartphones and start transferring data from one smartphone to another or to the implanted antenna. The data volume is 10G, hence its transmission provides us enough time to place both phones into the RF shield box and complete the setup. In the third step, we engage the reader Mod and keep it scanning the RFID tags. We then close the shield box and record the spectrums with and without the reader Mod. Following sections illustrates the test results for each wireless
module/sensor, respectively. We label the start, peak and end of the spectrum with green dot and specific frequency and decibel values.

![Diagram of testbed for frequency isolation analysis](image1)

Fig. 5. Testbed for Frequency Isolation Analysis

**B. Radio Spectrum of RFID Scanning Signal**

In Fig. 6 we show our design that is active at frequency band ranging from 864 MHz to 944 MHz with energy peak at 912 MHz. The strength at the peak is 15 dB at distance zero meters.

![UHF RFID spectrum graph](image2)

Fig. 6. UHF RFID Spectrum

1) **Cellular Signal Interference:** Testing for Cellular Interference is performed by connecting the Moto-Z with our UHF RFID reader Mod, and a live Verizon SIM. In both cases, with or without engaging the RFID reader, the radio spectrum of the Cellular voice call signal remains the same, as illustrates in Fig. 7(a), which indicates that there is no interference between RFID reader signal and the Cellular signal. For the Cellular data streaming, there is also no interference. The signal spectrum of data streaming with and without RFID reader
scanning both follow the pattern depicted in Fig. 7(b), all datasets streamed without retransmission.

![Verizon Voice Call](image1.png) ![Verizon LTE](image2.png)

(a) Cellular Voice (b) Cellular LTE

Fig. 7. Frequency analysis I

2) Bluetooth Signal Interference: To test Bluetooth interference, we capture the signal spectrums of all four stages of BT communications, i.e., BT scanning, advertising, paring and data transmission. The accumulated peak spectrums of all stages of BT communication, both with and without RFID scanning activity, are followed the same pattern depicted in Fig. 8(a). The test results in Fig. 8 show no signal interference with Bluetooth, and we do not find transmission errors with the dataset received on the PC side, which indicate there is no interference between BT and RFID reader Mod.

3) Near Field Communication (NFC) Signal Interference: Testing for NFC interference is performed by connecting the Moto Z with Mod, and using a NXP NFC tag. While reading the tags memory, RFID background scanning is running. The test results in Fig. 8(b) show no signal interference with NFC, and neither errors when reading or writing to the NFC tag.

4) WiFi Signal Interference: For WiFi test, we measure the two modes of WiFi communication: in 2.4G channel and in 5G channel. Both test are performed by set one smartphone as a hotspot and use another one to connect to the hotspot. Experimental data shows that for both modes, WiFi communication has no interference to the RFID reader Mod. The specific spectrum data are illustrated in Fig. 9(a) and Fig. 9(b) for 2.4G mode and 5G mode, respectively.

As illustrated in the five experiments above, our UHF RFID reader Mod has no interference on mainstream smartphone connectivity systems, which means our design is able to coexist with the native wireless sensors and communication modules.
C. Sensor Performance Evaluation

As a IoT sensor, the sensing range and the signal attenuation are two key parameters which determine the quality of service of IoT sensing. In the following, we design two experiments to evaluate these two features, respectively. For these two experiments, the environmental temperature is 18.3 Celsius, the target RFID tag is SMARTAC belt RFID paper tag [34] and the experiment lab is covered with radio absorb material to eradicate the radio reflection from wall faces. Battery is charged to 70%. Each value is the average value of 50 readings.

Our first experiment determines the sensing range of the UHF RFID reader Mod. From range 0 meter to 5 meters, we use our Mod to send 150 scanning signals at each distance and count the
successful Electronic Product Code (EPC) reads from RFID tag. Fig.10(a) illustrates the Successful EPC read count/rate of a RFID tag at different distances. As the picture shows, the signal picking up ability reduces when the distance increases but still very reliable before 3.5 meters. From 3.5 meters to 4.5 meters, it degrades dramatically and tags with range longer than 4.5 meter cannot be recognized. We determine the sensing range as 4 meters where the expected retry time is four, which is acceptable to most IoT sensing scenarios.

Our second experiment evaluates the signal sensitivity and attenuation of our reader Mod. In this experiment, we measure the RSSI (Received Signal Strength Indicator) provide from the reader IC from range 0 Meter to 4.2 Meter. The result is depicted in Fig.10(b) where the solid line is the actual value and the dot line is the trendline. With the result from the previous experiment, the attenuation distribution shows that the reader Mod has stable radio performance when signal strength is higher than -65 dB (3.5 meters) and signals that are weaker than -70dB cannot be detected. One interesting finding is that, in average, the attenuation increases with the grows of the distance, but it also shows dramatic decrease in some distance pints. In our case, there are three: 0.97 Meter, 3.18 Meter and 4.21 Meter. We are not able to explain this phenomenon for now, but we will further study it in our future work.

5. CONCLUSION

In this paper, we design and implement a UHF RFID reader Mod with direct connection to a mainstream smartphone, Moto-Z, without changing the hardware of the phone. The experimental data proves that our design of the UHF RFID reader Mod can coexist with other smartphone sensors and communication components with no conflict. Also, the battery life with active RFID scanning is extended significantly due to the control system and the external power source design.

Our research and architecture has been tested for commercial deployment and we have highlighted the key hardware and software architecture components which are extensible and scalable for continued research.

Through our work, mobile devices will take advantage of UHF RFID capabilities by evolving into sensory data accumulators passing raw and fused sensor states to cloud infrastructure for further compute-intensive processing. In our future work, we will optimize the algorithms running on the sensor fusion core to further reduce the power consumption to the acceptable range enabling us to remove the external power source. Next we will introduce RFID as a Sensor (RaaS), providing even wider sensing spectrum and more comprehensive environmental
information. Coupled with a highly optimized Sensor Fusion Core (SFC) and Bluetooth Low Energy (BLE) these sensors jointly provide a real-time, always-on and comprehensive sensory system optimized for a mobile devices limited battery life. Optimized wireless monitoring algorithms will be developed minimizing power consumption while guaranteeing an applications Quality of Service (QoS) requirements. We will also further study the signal sensitivity and attenuation behaviors when the reader Mod is activated.

REFERENCES


IMAGE QUALITY ASSESSMENT- A SURVEY OF RECENT APPROACHES

Noor Al Madeed, Zainab Awan and Somaya Al Madeed

College of Computer Science and Engineering, Qatar University, Doha, Qatar

ABSTRACT

Image Quality Assessment (IQA) is the process of quantifying degradation in image quality. With the increased image-based applications, IQA deserves extensive research. In this paper, we have presented popular IQA methods for the three types namely, Full Reference (FR), No Reference (NR) and Reduced Reference (RR). The paper gives comparison of the approaches in terms of the database used, the performance metric and the methods used.

KEYWORDS

Full reference, image quality assessment, no reference, reduced reference

1. INTRODUCTION

Image quality assessment refers to the evaluation of the quality of a distorted image with respect to the original image. There are two main categories of IQA; namely, subjective methods and objective methods. The former is difficult and time-consuming for real-time assessment needs. Subjective methods require human intervention and are very expensive. The later refers to the algorithmic models to estimate the image quality. For the real-time assessment we require automated or objective methods. There are three dimensions of objective IQA, namely, Full Reference FR, Reduced Reference RR and No Reference NR. FR IQA methods need the whole information of the reference image, while NR IQA methods stipulate that distorted image be assessed without any information of the reference image. RR methods are a compromise between FR and NR methods that require partial information of the reference image.

The rest of the paper has been organised into following sections. Section 2 defines image quality, section 3 explains image quality assessment, section 4 outlines types of IQA, and section 5 highlights few benchmark databases, followed by section 6 that discusses IQA approaches, followed by the discussion and conclusion.

2. IMAGE QUALITY

Image quality has been defined [1] as a composite of three factors, fidelity, perception and aesthetics. Fidelity is the exactness of a distorted image relative to its original image. Perception is inspired from characteristics of human visual systems HVS. This type of metric considers for example visual attention [2], contrast masking [3] etc. Aesthetics is subjective and may contain [1] visual constancy, visual attention and visual fatigue etc.
3. IMAGE QUALITY ASSESSMENT

Due to the continuous development in the multimedia technologies, images have become a great source of information. While transmitting an image over wired or wireless network channels, its quality is degraded due to the noise/distortions introduced. The image on the receiver side is no longer the same as the one sent at the sender side. IQA is hard as the algorithm needs to be reliable, fast, robust and should require less information from the reference image.

4. THREE TYPES OF IQA

IQA methods are categorized into the following three categories. Methods are categorized based on the amount of information needed from a reference/original image.

4.1. Full-Reference IQA

Full Reference QA algorithms have the original/reference image available at hand for the comparison with a distorted version of the same image. These kinds of methods are best suited for in-lab testing as in the real use cases the reference image is not available. Figure 1 is reproduced from [1].

4.2. No-Reference IQA

In NR QA methods, the reference image is not available at all. QA algorithm must be able to evaluate the quality of a distorted image as that is the only information available at hand. Figure 2 is reproduced from [1].
4.3. Reduced-Reference IQA

RR QA methods serve as a compromise between NR and FR methods. These methods require partial information from the reference image in the form of features. That partial information is sent to the receiver side through an ancillary channel. The goal here is to reduce the features required. Figure 3 taken from [4] indicates the framework for reduced reference IQA.

![Figure 3: RR IQA Framework [4]](image)

5. BENCHMARK DATABASES FOR IQA

In this section we will discuss some standard databases available for the evaluation of IQA methods. In order to evaluate a newly proposed IQA metric with the ones that are currently existing, one needs to perform experiments on the publicly available databases for the purpose. The databases contain reference images for which various distorted images are produced by using different distortion types. Each distorted image has as associated subjective score. To evaluate a new metric you need to compute the objective score of each distorted image and compute correlations between the objective and subjective scores.

5.1. CSIQ Database

This database [5] contains 30 reference images where each image is introduced with 6 types of distortions at 4 to 5 different levels of distortions. The distortions types include JPEG compression, JPEG-2000 compression, global contrast decrements, additive pink Gaussian noise, and Gaussian blurring. The subjective ratings are given in Differential Mean Opinion Scores (DMOS). The database is freely downloadable for the experimentation.

5.2. TID2013 Database

This database [6] contains larger test images set than its earlier release TID2008. It contains 25 reference images, 24 distortion types and 5 levels for each distortion type, which makes 3000 test images in total. The images and associated MOS values are freely available for the download and research purposes.

5.3. Tampere Image Database 2008 (TID2008) Database

This database [7] contains 25 reference images and 17 types with 4 levels of distortions for each image, in total 1700 images and subjective ratings are reported in the form of Mean Opinion Scores (MOS). The database is available freely for investigation of the existing and new metrics.
5.4 Laboratory for Image and Video Engineering (LIVE) Database

This database [8], [9] contains 29 reference images, each image distorted with 5 distortion types, total of 799 distorted images. Subjective ratings are provided in the form of DMOS.

5.5 IVC Database

This database [10] contains 8 original images distorted at 3 different processing and 5 different compression rates. In total there are 120 distorted test images available.

5.6 Cornell A-57 Database

This database [11] is relatively a small database. It contains 3 reference images, 6 distortion types at 3 different distortion levels. This database is publicly available for research.

5.7 Wireless Imaging Quality (WIQ) Database

This database [12] contains 7 reference images and 80 distorted images.

6. STATE-OF-THE-ART APPROACHES

This section discusses some recent approaches that have been proposed for the three types of assessment methods.

Table 1: Full reference approaches

<table>
<thead>
<tr>
<th>Serial No</th>
<th>Year</th>
<th>Method</th>
<th>Database</th>
<th>Performance Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2011</td>
<td>In spatial domain [15]</td>
<td>Standard MATLAB Vegetable Image</td>
<td>SCC</td>
</tr>
<tr>
<td>4</td>
<td>2015</td>
<td>Binocular Visual Characteristic [16]</td>
<td>NBU, LIVE I, LIVE II, MICT and CML</td>
<td>PLCC, SRCC and RMSE</td>
</tr>
<tr>
<td>5</td>
<td>2010</td>
<td>Gabor Filters [17]</td>
<td>Their own database explained in the paper</td>
<td>PCC, RMSE MOS, SCC and Outlier Ratio</td>
</tr>
</tbody>
</table>

6.1. Full Reference Approaches

Table 1 contains the studies that have been reviewed for FR methods. In this paper [13], a novel framework for calculating Image Quality Metric (IQM) is given in discrete wavelet domain using Haar wavelet. The framework is applicable to map based such as SSIM that generate quality maps and nonmap based metrics for example PSNR as a final score, with reduced complexity and improved accuracy. The framework has been applied to various well-known assessment methods SSIM, VIF, PSNR and AD. The framework provides trade-off between accuracy and complexity, hence, could be efficiently applied in wavelet based image/video processing applications.

This paper [14] presents a machine learning expert based on SVM classification and regression, known as Machine Learning based IQM (MLIQM), shown in Figure 4. The feature is composed
of local quality features on which multi class SVM classifier is applied to classify test image into one of the five quality classes as recommended by ITU. After this SVM regression is applied to score the quality of the test image. The proposed technique is compared with four FR IQA techniques i.e. VSNR, PSNR, VIF and MS-SSIM, have been reported to perform better than them.

Figure 4: Framework for quality score calculation [14]

This paper [15] proposes new FR IQA metric for objective assessment in spatial domain. This metric is a combination of four factors, namely, sensitivity, edge performance index, contrast improvement and structural correlation. The proposed framework is shown in Figure 5. The proposed metric is reported to be more reliable than SSIM across different distortion types.

Figure 5: Spatial domain quality metric [15]

This paper [16] proposes a framework that gives IQA method for stereoscopic 3D images. The proposed methodology consists of the two phases; training and quality estimation. Binocular receptive field properties are learnt to simulate in simple and complex cells in cortex and to project their impact in quality estimation.

This paper [17] proposed FRIQA for the images distorted with local geometric distortions. The framework is shown in Figure 6. At first, the displacement field and a set of features by using Gabor filters from the image structure are extracted. Local quality scores are then calculated by evaluating the impact of displacement field on the features. Error pooling is performed to calculate the overall quality metric.
Table 2: Reduced reference approaches

<table>
<thead>
<tr>
<th>Serial No</th>
<th>Year</th>
<th>Method</th>
<th>Database</th>
<th>Performance Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2015</td>
<td>Deep Learning, Restricted Boltzmann Machine Similarity Measure RBMSim [18]</td>
<td>LIVE Multiply Distorted Image Quality and CSIQ</td>
<td>SROCC and Pearson Correlation Coefficient</td>
</tr>
<tr>
<td>2</td>
<td>2012</td>
<td>Entropic Differencing [19]</td>
<td>Tampere and LIVE</td>
<td>SROCC</td>
</tr>
<tr>
<td>5</td>
<td>2012</td>
<td>2012 Structural Similarity Estimation [22]</td>
<td>LIVE, Cornell, IVC, Toyama-MICT, TID 2008</td>
<td>CSIQ, PLCC, MAE, RMS, SROCC, KRCC</td>
</tr>
<tr>
<td>6</td>
<td>2014</td>
<td>Phase Information in Complex Wavelet Domain [23]</td>
<td>LIVE</td>
<td>Correlation Coefficient (Prediction Accuracy) and Rank-Order Correlation(Prediction Monotonicity)</td>
</tr>
<tr>
<td>7</td>
<td>2015</td>
<td>Entropy Differences in DCT Domain [24]</td>
<td>CSIQ, Toyama and LIVE</td>
<td>Pearson Linear and Spearman Rank-Order Correlation Coefficient (SROCC)</td>
</tr>
</tbody>
</table>

6.2. Reduced Reference Approaches

Table 2 contains the studies that have been reviewed for RR methods. Novel stochastic RR-IQA metric [18], automatic stochastic procedure capable of assessing the quality of distorted images, independent of the type of image of distortions. Compared with subjective IQA algorithms and two benchmark databases. Restricted Boltzmann Machines are stochastic neural networks and contain two layers of neurons, the hidden layer h and the visible layer v, used for unsupervised
learning. The learning algorithm used for BMs is called Contrastive Divergence CD. Framework is called Gaussian Bernoulli Restricted Boltzmann Machine (GRBM). Each pixel from the RI is modelled on the visible layer v, with individual RGB value encoded as one neuron. For easy weight manipulation the image was split into sub-images. The distorted image DI quality can be assessed by initializing the v of GRBM learned on the RI by the sub images from DI, this is called vDI. After this step, Gibbs sampling is performed to infer the values of hidden neuron and then values of visible neurons are reconstructed based on the hidden neuron values. The reconstructed values are called vDI*. The Root Mean Squared Error RMSE distance (or any other metric) between vDI and vDI* reflects the quality of distorted image. GRBM trained on the RI has to be sent to the client side and very less information is required from the reference image. The performance is comparable with widely known FR- IQA metrics and its computation time is less. So, it can be applied to video/multimedia applications to be able to do real-time performance evaluation of network services.

In [19] the quality score is estimated by measuring the changes in weighted entropies of reference and distorted images in the wavelet domain. Wavelet coefficients of natural image and distorted are modelled as GSM distributions. The quality metric is obtained as a distance between reference and natural image approximation of a distorted image. The proposed algorithm performs much better than FR MSE algorithm.

In this paper [20] the IQA metric is based on Discrete Cosine Transform DCT coefficients, the method follows three steps, shown in Figure 7. In the first step, block-based DCT is applied, as a result of which we have DC coefficients where the most of the energy of the image resides and AC coefficients which can be easily ignored because energy is very low within these coefficients. The DCT coefficients are reorganized by decomposing them into 10 sub bands. Then the coefficients of the same sub bands from different blocks are grouped together such that there is a structural similarity between all the sub bands. The reorganized DCT coefficients look like wavelet image representation. In the second step, generalized Gaussian density GGD curve fitting is done to model DCT coefficient distribution GGD model efficiently represent the coefficient histogram for each DCT subband. In addition to two standard parameters of GGD another parameter, prediction error is introduced. In the last step, city block distance is computed between actual distribution and the fitted distribution. The proposed method has been tested on LIVE database and compared with the well-known RR WNISM and FR PSNR and SSIM. The proposed method outperforms the RR WNISM and FR PSNR but slightly underperforms the FR SSIM.

![Figure 7: RR IQA DCT based metric [20]](image)

In [21] a novel RR IQA algorithm based on image statistics is proposed. Image statistics are modelled in pixel domain which is based on gradient distribution. RR quality features are extracted based on statistical model. The quality is measured by determining the similarity between reference and distorted image. The proposed method is reported to perform as good as widely known FR PSNR.
In this paper [22], an approximation attempt for RR SSIM metric from FR SSIM has been made by divisible normalization transform (DNT) of the natural image. In addition to quality assessment, image repair concept has been introduced by matching the sub band statistical properties of the distorted image with those of the reference image. Features are extracted from the reference image by applying multi-scale linear transform (wavelet transform) and then DNT representation is calculated by dividing each coefficient of wavelet by a local energy measure based on neighbouring coefficients. For the efficient summarization of the statistical properties of the reference image zero mean GGD is fitted over DNT coefficients. For the estimation of RR SSIM, the effect of the distortions on the statistical properties need to be coherent with corresponding FR SSIM. The proposed method is compared with FR (PSNR and SSIM) and RR (Wavelet Marginal and DNT marginal) and the proposed metric is highly competitive in most of the cases.

In this paper [23] IQA is based on the Complex Wavelet Transform (CWT) as it carries not only the magnitude information but also the phase information. Choosing the right RR features for IQA is critical, they have to be representative and yet small. Here different features are used namely, KLD, the standard deviation, the kurtosis, the skewness in the distribution of the relative phase and directional information from CWT coefficients. Generalized complex wavelet coefficient known as dual-tree complex wavelet transform (DT-CWT) has been used rather than simplistic CWT. The central idea of the algorithm is that the different distortion types affect the distribution in different ways. The distribution of phases of CWT coefficients is uniform and therefore contain no information. But the distribution of relative phases of different images follows a similar pattern when affected by a single distortion type. In case of different distortion type the pattern would be different. General Regression Neural Network (GRNN) has been used to predict the quality of distorted image by regression. The performance has been compared with 2 FR-IQA and 5 RR-IQA metrics and the proposed method is reported to be as good as popular FR-IQA metric MS-SSIM while using smaller features. To assess the quality of the distorted image we need the histograms of the relative phase of the distorted and the original image and then compute the Kullback-Liebler divergence.

In this paper [24], Figure 8, the differing importance of spectral sub-bands in DCT domain of images have been exploited by considering the fact that human eye corresponds differently to different bands. Quality of image should be assessed separately in each sub-band. Quality of the distorted image is computed via Shannon entropy differencing of DCT co-efficients between the reference and the distorted image. The overall quality of the distorted image is calculated by the weighted summation over entropy difference of all the sub bands. On the sender side the feature extraction is performed in three steps i.e. firstly, block-based DCT of the RI, secondly, the coefficient rearrangement and lastly the entropy calculation of each band. The same process is performed on the receiver side with the DI. The entropy features of RI at the sender side are sent to the receiver side over the ancillary error free channel which is required at the receiver side for the differencing. Higher prediction accuracy was attained using less reference data, similar frequency bands were merged into one term. The proposed algorithm was compared with already existing RRED, WNISM and RRVIF and overall outperforms over the three methods.
6.3. No Reference Approaches

Table 3 contains the studies that have been reviewed for NR methods. The proposed framework [25] is a combination of two existing NR-IQA approaches, shown in Figure 9. The two approaches are called free energy based distortion metric FEDM and structural degradation model SDM. The newly proposed metric no reference free energy and structural degradation model NFSDM is reported to outperform two FR metrics, SSIM and PSNR and NR metrics BLIINDS-II [26] and DIIIVINE [27]. This metric makes use of a relatively novel feature which is a non-linear
combination of free energy and structural degradation information. This metric has been tested on LIVE database.

This paper [28] proposes shearlet based no reference image quality assessment SHANIA framework, Figure 10. The method makes use of natural scene statistics NSS model in shearlet domain [29]. The central idea behind approach is that the statistical properties of natural images remain constant in shearlet domain and change for the distorted images. The natural parts from the distorted image are used as reference and quality score is computed between reference and the distorted parts. The problem is mapped into a classification problem. The features are obtained from the distorted and reference parts and softmax classifier is used to obtain the quality score. The proposed framework is comparable in performance to three FR metrics including PSNR, SSIM and Multi-Scale SSIM and four NR methods BIQI, DIIVINE, BLINDS II and BRISQUE.

The paper [30] proposes NR-IQA framework based on visual codebooks which is a generalized framework not particular for some types of distortions, as shown in Figure 11. The proposed framework is reported to outperform FR-IQA SSIM and PSNR on LIVE database. The proposed method makes use of local feature extracted by using Gabor-filters. After this stage, visual codebook is constructed by forming a large set of block based Gabor feature vectors for the training images. Codebook C is constructed from that set by using clustering algorithm. Images
are represented in the form of codewords taken from codebook. Afterwards, regression is applied to calculate the quality score.

Figure 11: Visual Codebooks based NR IQA Framework [30]

The proposed framework [31] is for multiple domain images in real time applications, shown in Figure 12. There are three steps involved in this method, (1) local feature extraction, (2) global feature extraction and (3) regression model. In this approach step 1 and 3 unlike previous approaches are not treated independently rather they are linked together by back projection. Back projection allows to choose less number of discriminative features. The method utilizes relatively small set of features obtained from raw image patches directly which makes it a fast method. It has been applied to the document image datasets to prove its cross-domain applicability.

Figure 12: Framework for NR-IQA based on filter learning [31]
The proposed framework [32] utilizes local spatial and spectral features of distorted images, overview in Figure 13. Two types of features have been used in the method, namely, spatial and spectral entropies. The method involves four stages, i.e., pre-processing(down-sampling) of the distorted image, block-based partition and computation of spectral and spatial entropies for each block, percentile pooling from the two feature sets is done and lastly the quality score is computed. The metric is called Spatial-Spectral Entropy based Quality index SSEQ and have been reported to outperform FR SSIM and NR DIIVINE, BLINDS II and BIQI on LIVE and TID2008 database.

Figure 13: Framework for NR-IQA based on spatial and spectral entropies [32]

7. DISCUSSION

The problem of IQA is inherently hard. It has been dealt by using different tools. In this paper we have discussed entropy based methods, gradient based methods and DCT based methods. We have also shown a framework where deep learning has been used to deal with it. The fact that images from different domains have different properties. We represent those properties as features to a machine. Therefore, we cannot have universal set of features that could be applicable to all types of applications. State-of-the-art methods for the three types of IQA have been discussed previously. It shows that we need new IQA metrics that possess greater power than widely used PSNR and MSE.

8. CONCLUSIONS

We have reviewed some state-of-the-art approaches for the three types of IQA. Of the three types, NR IQA poses most challenges. As it requires no information from the reference image at the same time without strict assumptions regarding distortion level and types. Till date no system have been developed that meets the requirements of real-time quality prediction, data-base independence and generalized for all the distortion types and levels.

ACKNOWLEDGEMENTS

This publication was made possible by NPRP grant # NPRP8-140-2-065 from Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.
REFERENCES


THE WEB-BASED EDUCATION JOURNEY: A CONSTANT LIFELINE

Vidhu Mitha

Department of Information Technology, SRM University, Chennai, India

ABSTRACT

E-learning has revolutionized our realm in more than just a listable number of ways. But it took a paradigm shift when it entered the threshold of the varsity system. With the prevailing spoon-feeding era, are the students really ~industry ready? We answer that by confirming a fact; web-based learning has become the oxygen of freshers in the IT Industry instead of the traditional learning done through graduation. Furthermore, are university enforced e-learning assessment systems a true representation of a student’s proficiency? This paper is a peep into what web-based e-learning systems are to a student of today’s world, by giving an overview of university-level e-learning in India deploying an example from SRM University’s organizational framework. It assesses a key e-learning trend, the implementation of which bridges the gap between universities and the industry. It is proposed to provide constructive feedback to the e-learning community and shine some light on areas of scope for future developments.

KEYWORDS

web-based learning, e-learning, university

1. INTRODUCTION

We see trends move toward a NO-PAPER system in education today. But what is the reality?

Ranging from high schools’ e-labs, to finding it’s traces in fully developed experts’ CVs as nano degrees, e-learning has found its way in, everywhere including in India with an estimated $3 billion market. With the rapidly changing definition of e-learning today, it has swept countless fields and has a universal set of users starting from middle schoolers who’ve started submitting their assignments online, to corporate biggies that cannot function without it.

Gone are the days that a student accredited an institution's tutoring for his/her learning. Each student now has a pet go-to e-learning facility to thank for saving their tests and scores! This what tempts us to ask if this was used as a last-minute trick and if it helped? If yes, we will discuss if it is really a true representation of a student’s skill as compared to a hardworking offline-study student.

But on the other hand, is this system being utilized to its maxed potential?
In universities, fully equipped hi-tech labs are available for training graduates. During the college years, we are taught the practice of multiple technologies, coding languages and its practical implementations. But the fact remains that over 90% of it, is learned online. Throughout this learning process, students are being asked to line up to a system bound by the university, to complete the courses they have defined for them. This leads to the said underutilization of e-learning. While sticking to pre-defined university courseware, we lose track of the fast-paced industry's growing needs. This would lead us to the discussion on "bridging the gap".

2. THROUGH THE STUDENT LENS

On an average, a graduate level student uses books less than once a week. But online content is accessed at least 4-5 days a week and multiple times during a study day. The virtual notebook has taken over the blackboard. A few scenarios can help explain how:

When a student does not know a common definition or is unaware of the meaning of certain phrases or terminology used somewhere, he no longer looks for an encyclopedia, but looks up his phone’s google dictionary instead. It is a lot faster than manual searching since it’s just a few clicks away.

Another example could be when an assignment or project must be submitted. The student has many options open to him on the internet with an ocean of similar projects that can easily help set precedence to his project. There are websites which provide free expert reviews of student’s projects and help them create it from scratch, from building blueprints onward and even for free. This was obviously much more economic compared to when we they had to purchase a book for the same purpose.

Yet another instance could be the paramount help e-learning is during competitive exams and university tests. Instead of spending many hours at tedious and monotonous lectures, there are interesting and fun ways of learning concepts via puzzles, questionnaires, online group activities etc. on the internet which have been proven to be more effective due to shortcut methods available, even if it is a last-minute preparation. They also have access to large databases of question papers to refer and study from. And based on each student’s caliber, there are specialized test questions available. The mock tests online also help give them an approximate idea of their preparation level and an expected score range.

So, we can see that a student's learning curve is vague but still dependent on the method of e-learning that he chooses. To make a wise choice he must pick the method that is most comfortable to him and suited to his understanding and grasping power. The different choices he has include:

2.1. STATIC LEARNING

This is a one-way street where the student's only interaction with the web is to access the disseminated and stored information or content. This includes but is not limited to PowerPoint slides, E-books, question papers, videos, recorded lectures and related material. On picking up the relevant content from sources such as WhatsApp class groups, University-Notes portals etc., the student conducts domestic self-study with the help of these collectives.
2.2. DYNAMIC LEARNING

This usually consists of a regular online community that is balanced by two poles: The ‘teacher’ and the ‘pupil’.

This is further divided into:

- Private tutoring: The teacher imparts knowledge and ideas to be learned and explains them to the pupil in a one-on-one session.

- Group E-Learn Meets: Two or more people who are well versed in a particular common area of interest get together and exchange ideas to help each other out. Both are equal participants in this exchange.

- The Online Classroom: A qualified individual schedules a class and delivers a lecture to an entire class of interested individuals live via Skype, Blackboard, FaceTime, Google Plus Hangouts, Udacity etc., and also conducts a questionnaire session for doubt discussions.

Most of the higher education institutes and universities in India are adopting the online classroom method. Students are getting on to the E-learning journey either by choice to skill up in new areas or as part of their university curriculum. Once he selects what is personally best for him, he is all set to go on his E-learning journey.

3. ALONG THE WAY

The student starts walking down his selected road, but he encounters a lot of deviations and obstacles.

With universities enforcing online tests and online submissions as a focal part of education, there is the rising threat to authenticity. While e-learning enabled students to work time-effectively, it still had its shortcomings with respect to college level testing. Students could always duplicate content from online easily using plugins for every type of testing software. This is majorly due to the general technical drawback of the online world: there was always a way to break in, hack and alter the functioning of most systems easily. This has made it increasingly easy for students to indulge in malpractice since every step to performing such unethical hacks are present openly online paradoxically.

So, in an era where cybercrimes are a matter of fact, online testing might not be a completely apt representation of a student's true potential and proficiency.

But since it's our best bet for development and still more effective than traditional learning, we move on to analyzing it further. At university level, to be successful at web learning, there are multiple factors that contribute apart from the students learning capacity and tutor/instructor:

- The student’s attitude.
- The content organization or knowledge base.
• An interactive tool or the web-based platform used as media for communication, which is more important than the content itself.

Many universities are launching distance education programs using web-based learning and online libraries. With India being the current third largest online market for education in the world, it has launched NPTEL—a Govt. initiative and website where regular courses from reputed Indian institutions hosted for free. One such similar but private e-learning instance in India is eLab.

3.1. THE REVIEW OF eLAB

SRM university's eLab is an auto-evaluation tool for learning programming. eLab helps learners to practice and acquire programming skills. Sophomore students are warmed up to the new coding languages they would be learning through a semester using this tool. With a step-by-step approach, it levels up in competitiveness with every question. There are several sections present in the tool, to cater to all the subjects the student will be completing in the current semester. Each section within a subject would pertain to a chapter of the subject. Thus, all areas of the student’s study, would be thoroughly tested. Each student is required to complete a set of programs whose problem statements will be defined to them along with example test cases (as shown in the screenshot below). The student is required to analyze the question and code to satisfy the test cases. This system executes the written code and returns a result or highlights the issues with the code if insufficient. The system is also designed to award a completion score in percentage. If the code was ideal and satisfied all conditions, it recorded a 100%. If the code worked in favor of the question but would not satisfy all conditions, it recorded a 33,50 or 75 percent based on how much improvement the code needed. Finally, an average of all these scores makes the final cut to be added to the student's grade point average (GPA).

Initially the access to this tool was limited to devices connected via the university’s private network. Now it has been made available for use from any network. So now students can use it from anywhere including home, to complete their graded college work and submit it as well. But the list of pros doesn’t stop there. Teachers and respective faculty gain well from this tool too. They can check the student’s progress at any given time, review & reframe the question set, and download a progress report for all students that registered under them.

The tool’s access specifications are sealed in the backend, due to which the students cannot change or edit their grades or scores. Additionally, since the same software is used for multiple subjects, it works well to satisfy the university's needs while constantly acting as an effective learning tool for the student learners. Though the tool has a very good infrastructure for web learning, the content must match real-world problems, else it would defeat the purpose of learning programming. With students dealing with university enforced tools with a lot of loopholes, they lag behind when they are sent into the industry.
4. ZOOMING IN ON THE GAP

With the advent of e-learning in colleges, the universities curriculum started to get greased and easy. But this was only one side of the learning curve because on the other end of the curve, were graduates passing out of the universities with no knowledge of the industry whatsoever. They were incomplete products whom the companies seem to devour and shape to suit their specific needs.

According to HR reports, around 1.5 million engineers are released into the job/working market every year. But only about 25% of them are directly employable from the college campus. So, this goes to say that employers today look for more than just the basic skill set in their employees, they look for a long-term relationship with the organization. Which is why they focus on reskilling. Graduate students are required to unlearn and learn to fit into the IT world. So, the educators need to start thinking ahead of the curve.

This is where companies like Embibe and Simplilearn enter the picture with their courses for mid-level professionals. They help create a smooth segway from college to industry. Many businesses empires have been built around helping this university-to-company transition: Coursera, Udacity, edX etc. These e-learning websites primarily focus on delivering certified industry level courses which are now even being prescribed by companies itself! MNCs like Google, Microsoft and IBM now ask their employees to refer to such common e-learning portals instead of conducting intense and tedious training programs. A Brandon-Hall Study proved that learning online required almost 40% to 60% less employee time than learning the same topic in a traditional classroom setup. For instance, when IBM implemented an e-learning program in their company, its employees learned five times more without increasing their time spent in training.
When the freshers felt unprepared for the market, they ran to these web-based systems for a revival. On completion of these online courses from Coursera etc., they would've added jewels of value to their CVs which immediately helped them get hired and have also eased their transition into the walls of the industry since companies preferred such candidates over newer unprepared college minds.

4.1. HOW THE GAP COULD BE HANDLED

- Course content for universities should start to be: prescribed by or delivered directly from industry partners or companies, as the constant updates that are done by college seem insufficient. Therefore, the responsibility comes upon the system to update the syllabus and the pedagogy based on the society's demands.

- Universities should use professional content developers who have the industry experience.

- Web-based learning systems should consider using Artificial Intelligence. The tools should be smart and more interactive in nature. It should learn by itself and adjust the complexity of the delivered course based on the interaction and feedback of the students.

- The evaluation tools should predict the student understanding level and adjust the complexity level and provide the set of questions dynamically based on their response and award accordingly.

- Colleges should create space for students with a more direct inclination to the industry and inculcate that as part of their graded courseware, as outcome-driven academic systems flourish better.

- More web-based courses should be introduced into university curriculum to help industry freshers since they come with no experience of the workplace pressure, workload volumes or co-ordination required at companies at a global level. They need to be thought how to unlearn what they have collected in domestic college environments and learn what real-time systems require.

5. IN CONCLUSION

With the rigorously increasing E-learning appetite, the education system has improved by leaps and bounds as compared to what it was in the past decade. But ever since the comfort of studying on your own couch was introduced, the lazy students sense the main turbulence of studying at their own pace - the procrastination. Students need to have a highly motivated self-learning drive since web-based learning eliminates travel, timings and sometimes deadlines too and makes it all too flexible. But we cannot deny the fact that both the economy and literacy levels have gone up since the advent of web-based e-learning. Furthermore, tailored courses with flexible completion timing do enhance majority of students’ inclination toward learning. The stepping stool into the IT industry has become much larger with e-learning. More and more companies are starting to recognize this trend. It is due to the abstraction feature of web-learning that, from a sea of unrelated information we get straightforward and tailored content delivered to us for learning. Thus, employees are now able to deliver better and effective results at within shorter time spans.
So, web-learning has contributed to nothing but good to the technical revolution. But with just a few more changes and guideline enforcements, in the foreseeable future, e-learning would have revamped the book era entirely and become the new worldwide education standard. Universities would be technology enabled with a much leaner infrastructure. And in this changed paradigm, the learner is the new center of the knowledge-universe with his personalized learning path.

REFERENCES


[3] Prof (Dr) Rajan Saxena, (2016) "Future of jobs and its implications on Indian higher education".


AUTHOR

Vidhu Mitha Goutham is a second-year student pursuing Engineering in Information Technology at SRM University in Chennai, India with a keen interest in voicing to the online community.
INTENTIONAL BLANK
IMPROVED HANDOVER ALGORITHM FOR PROXY MIPv6 BASED ON AAA SERVER

Hewei Yu¹ and Meiling Zhou²

¹School of Computer Science & Engineering, South China University of Technology, Guang Zhou, P.R.China
²GF Fund Management Co. Ltd., Guang Zhou, P.R.China

ABSTRACT

This paper proposes an improved handover algorithm which does not need authenticating again if Mobile Node moves within the same Proxy MIPv6 domain. When MN enters PMIPv6 domain at the first time, it needs to make an AAA authentication. But when MN moves between MAGs in the same domain, it can perform handover procedure without the second times of AAA authentication, and speed the handover process. We built a structure of PMIPv6 including AAA server on NS-2, and proposed an improved handover algorithm for PMIPv6 based on AAA authentication. The simulation results show that the new scheme can effectively reduce the handover latency and ratio of packet loss, and improve network performance.

KEYWORDS

PMIPv6; AAA; handover; fast handover

1. INTRODUCTION

Nowadays, the huge commercial demand for mobile Internet has brought new opportunity and challenge of technology. In 2008, the Internet Engineering Task Force (IETF) proposed a network-based mobility management protocol called Proxy MIPv6[1]. Because of the bright prospect of PMIPv6, researchers all over the world have carried out extensive and in-depth study on PMIPv6, and provides various effective handoff scheme. As a whole, the research on PMIPv6 has many aspects, including security, handover performance etc. On the security mechanism of PMIPv6, RFC5779[2] defined AAA for authentication, authorization and accounting. Document[3] designed and implemented the authentication and authorization of PMIPv6 network system, the testing results proved the validity and feasibility of the certification system. On the handover performance of PMIPv6, document[4] proposed a seamless handover scheme using the neighbor discovery message of IPv6 to reduce the handover latency and packet buffering at the Mobile Access Gateway (MAG) to avoid the on-the-fly packet loss during a handover for PMIPv6. However, it is unable to satisfy the real-time demand of higher business communication. Document[5] proposes a novel low latency handover scheme for PMIPv6 using Media Independent Handover (MIH) services. Document[6] tested the handover performance of PMIPv6 with and without the IEEE 802.21 MIH services in a heterogeneous wireless networks' environment. The results showed that the handover performance had improved if the IEEE 802.21 MIH services is used, while there have some problem such as Signal overload. Of course, A scheme to Reduce Packet Loss during PMIPv6 Handover is called PL-PMIPv6, which is similar to the fast handover for MIPv6, and FH-PMIPv6 was improved based on it.
In the PMIPv6 system include AAA servers, when MN enters PMIPv6 domain at the first time, it is required to be authenticated. This paper proposes a scheme, MN only need to be authenticated once in the same domain. In another word, when MN moves into a new MAG in the same domain, it does not need to be authenticated again, so this schema can minimize the authentication latency and packet loss.

The rest of this paper is arranged as follows: the implementation of AAA server for PMIPv6 is introduced in section 2. Section 3 introduces the improved handover schema to minimize typical authentication procedures. Simulation and results analysis are provided in section 4. This paper is concluded in section 5.

2. IMPLEMENTATION OF AAA SERVER FOR PMIPv6

PMIPv6 introduces two network entities: MAG (Mobile Access Gateway) and LMA (Local Mobility Anchor). LMA is the home agent for the MN in a PMIPv6 domain. It is the topological anchor point for the MN’s home network prefix(es) and is the entity that manages the MN’s binding state. MAG is a function on an access router that manages the mobility-related signalling. It is responsible for tracking the MN’s movements to and from the access link and for signalling the MN’s local mobility anchor.

In this paper, the functions of AAA server are implemented in NS2 using C++ language and TCL scripts. We add two structures that is AAA server and list of policy profile. Figure 1. (a) and Figure 1. (b) show the working flow chart of AAA server.

After AAA server is implemented, Figure 2 shows the signal flow when MN moves from P-MAG (pre-MAG) to N-MAG (new-MAG) in the same PMIPv6 domain.

![Figure 1](image-url)

Figure 1. (a) Create list of policy profile (b) Judging the legality of MN

The whole handover procedure goes as follows:

1) P-MAG sends DeReg PBU message (including MN-HNP and Proxy-CoA) to LMA when it detects that MN is going to leave. Then LMA can release the binding between MN-HNP and Proxy-CoA of MN.

2) LMA starts the delay timer immediately when it receives DeReg PBU message. So LMA allows N-MAG to wait for a period of time before updating the binding cache entry, and replies DeReg
PBA message to P-MAG simultaneously.

3) P-MAG receives DeReg PBA message.

4) When N-MAG detects the attachment of MN, N-MAG obtains information about MN. Then MN sends RS message to N-MAG.

5) N-MAG sends QUERY message to AAA server after received RS message, then gets policy profile of MN.

6) After MN’s successful access authentication, N-MAG sends PBU message to LMA to update the LMA about the current location of MN.

7) Upon receiving the PBU message, LMA updates binding cache entry. LMA sends PBA message including MN-HNP.

8) After receiving the PBA message, N-MAG sets up a tunnel to LMA and adds a default route over tunnel to LMA. N-MAG then sends RA message to MN on the access link to advertise the MN-HNP.

9) When MN receives RA and know that there are no change on L3 interface after checking, it keeps the same IP address and uses it for packet delivery.

Figure 2. Signal flow of Proxy MIPv6 based on Inter-Domain Handover
As signal flow chart shows, the handover schemes for PMIPv6 networks includes deregistration (1-3), authentication (4-5) and registration (6-9) phases.

3. IMPROVED HANDOVER SCHEME

In order to reduce the longer latency and extra cost caused by combining the AAA function while guaranteeing the handover performance, some handover procedure have been proposed as follows: in document[8], the AAA server selects the most suitable LMA based on the load information received from the LMAs. Moreover, the LMA performs admission control based on the number of current mobile nodes registered to the LMA. Document[9] proposed adaptive context transfer schemes for a fast handoff in Proxy MIPv6 which reduces the latency in AAA authentication and specifies context transfer scenarios taking into account proactive and reactive handover. When MNs attempt to handover inter/intra domain, the AAA context information stored in LMAs and MAGs will be used to support the handover without visiting the AAA server. Other documents put forward an configuration file based intra-domain handover method in PMIPv6. Optimize the intra-domain handover performance by expanding the contents of the configuration file and adding the proxy care-of address option. Of course, there are many improved schemes based on AAA server. In this section we discuss our proposed scheme to reduce the handover latency of Proxy MIPv6 which is based on intra domain.

As we know, handover latency in PMIPv6 is caused due to following four latencies[10]:

- Link switching latency
- AAA authentication latency
- MN registration latency
- Latency due to Router Solicitation(RS) message and Router Advertisement(RA) message

In the proposed scheme, the handover latency is reduced due to AAA authentication again in the same domain.

As mentioned previously, the MAG detects the movements of MN and performs mobility-related signalling with the LMA in place of the MN. As Fig3 shows, the serving MAG(i.e.,P-MAG) will send handover initial (HI) message to the target MAG(i.e.,N-MAG) when MN is going to leave. Then N-MAG will be told the MN has been authentication. First , N-MAG stores information (such as the AAA initiation information of MN(i.e.,MN-ID) and the address of LMA ) from HI message. Then N-MAG sends back a handover acknowledge(HACK) message to P-MAG. During this time, MN has not disconnected from P-MAG, so all the data packets are forwarded through P-MAG as before.

P-MAG sends DeReg PBU message to LMA while detects MN is disconnect. The deregistration period is the same as mentioned previously.

When N-MAG detects that MN enters or gets the Route Solicitation(RS) message from MN, it will check whether the MN’s information (including MN-ID and LMAA) have already existed. If N-MAG has MN’s information, it means that MN has been authenticated. So it is not need to registration again. Here we give some analysis about handover latency. Usually handover latency is marked by the maximum time interval between MN receives the last packet from the P-MAG and the first packet from N-MAG. Using some marks such as $T_{DeReg}$ (deregistration), $T_{Auth}$ (authentication), $T_{Reg}$ (registration), $T_{MN-MAG}$ (MN and MAG), $T_{MAG-LMA}$ (MAG and LMA), $T_{MAG-}$
AAA(MAG and AAA) and T_{L2} (layer2 handover), we can get handover latency (HL) of traditional algorithm according to Fig4.1.

\[ HL = T_{DeReg} + T_{L2} + T_{RS} + T_{Auth} + T_{Reg} + T_{RA} \]
\[ = 4T_{MAG,LMA} + 2T_{L2} + 2T_{MN-MAG} + 2T_{MAG,AAA} \quad (1) \]

The handover latency (HL) of proposed algorithm is as the following according to Figure 4 (b).

\[ HL = T_{DeReg} + T_{L2} + T_{RS} + T_{Reg} + T_{RA} \]
\[ = 4T_{MAG,LMA} + 2T_{MN-MAG} \quad (2) \]

Compare equation (1) and (2), you will find that the delay of AAA authentication is eliminated in our algorithm.
4. SIMULATION AND ANALYSIS

We have taken simulation using NS-2 and compared the performance of the proposed scheme with the previous and the ordinary one when MN moves between MAGs in the same domain. In the simulation, we assumed correspondent node (CN) sent packets to MN and MN moved from P-MAG to N-MAG, the packets from CN get to MAG through LMA and router. The simulation model is shown in Figure 5. The time in the simulation is controlled in 20 seconds. CN sent data packets at intervals of 0.05s after 1.0 second. At 14.4 second, MN moved away from P-MAG and went toward N-MAG.
2.1. Handover latency comparison

The handover latency comparison between ordinary (marked by black) and proposed (marked by red) scheme is shown in Fig. 6. The abscissa represents the simulation time, the ordinate represents handover latency. Because of MN has not entered PMIPv6 network at the starting time, there are no handover latency and the delay time is 0s. The packets transfer delay is 0.05s. There are two high vertical lines, the first line represents the latency that MN enters a new Proxy MIPv6 domain and attaches to P-MAG. The graph shows, handover latency is sharply increased up to 1.02s at 1s. It is the same as the ordinary scheme. However, there are visible differences in the second line. The second line represents the latency that MN enters N-MAG. The handover latency of the ordinary scheme is sharply increased up to 0.5s while it is 0.45s in the proposed scheme, which shows that the proposed scheme can reduce the handover latency.

![Figure 6. Handover latency comparison between ordinary and proposed scheme](image)

Table 1 shows some parameters about handover latency in the simulation.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Start time of handover(s)</th>
<th>Finish time of handover(s)</th>
<th>Handover latency(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous scheme</td>
<td>14.4</td>
<td>14.9</td>
<td>0.5</td>
</tr>
<tr>
<td>Proposed scheme</td>
<td>14.4</td>
<td>14.85</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Both of the start time of handover of the two schemes are 14.4s. The proposed scheme has finished handover at 14.85s, and the ordinary one has finished at 14.9s.

Handover latency of the ordinary scheme = 0.5s

Handover latency of the proposed scheme = 0.45s

Obviously, the handover latency of the proposed scheme has been reduced 0.05s (0.5-0.45=0.05s). It is decreased by 10% ((0.05/0.5)*100%=10%) according to our proposal.

2.2. Packet loss rate

Figure 7 shows the number of packet loss comparison of ordinary and proposed scheme. In this simulation, the ordinary scheme lost 9 packets. As while as the proposed scheme lost 8 which is less than the ordinary one.
In the simulation, the largest number of data packet is 383 in two scheme by checking the simulation output trace file. So we can get table 2.

![Figure 7. Number of packet loss of ordinary and proposed scheme](image)

All of the forward data packets are 383.

The number of packet loss are 9 and 8 in ordinary and proposed scheme respectively.

Obviously, the ratio of packet loss of our proposed scheme has been reduced $1 (9-8=1)$. So the ratio of packet loss is decreased by $11.11\% ((1/9)^{*}100\% \approx 11.11\%)$ according to our proposal.

To sum up, the proposed scheme is characterized by completed N-MAG authentication in advance of handover, so there is no need of authentication again. It is clear from the result of simulation, the proposed scheme reduces the handover latency and the ratio of packet loss without increased throughput.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>The number of loss</th>
<th>The number of forward</th>
<th>Packet loss ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous scheme</td>
<td>9</td>
<td>383</td>
<td>$\approx 2.35%$</td>
</tr>
<tr>
<td>Proposed scheme</td>
<td>8</td>
<td>383</td>
<td>$\approx 2.09%$</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

In this paper, we introduce the present development of PMIPv6 and the way to add AAA server to PMIPv6 in NS-2. What’s more, we have proposed an improved handover scheme which can minimize handover latency and ratio of packet loss by omitting operation of AAA authentication again in the same domain. We evaluate the performance of both the ordinary and the proposed scheme in NS-2 using a PMIPv6 system model. The simulation results show that our proposed scheme can effectively reduce the handover latency and ratio of packet loss. However, there are some drawbacks, for example, it is only satiable for intra-domain handover, the security issues is not perfect etc. Further work should include the improvement and simulation jobs aimed at the problems above.

ACKNOWLEDGEMENTS

This study is supported by National Natural Science Foundation of China (No. 61070179), Guangzhou Science & Technology Project (No.2014J4100019), Key Collaborative Innovation Project among Industry, University and Institute of Guangzhou (No. 20160401001).
REFERENCES


AUTHORS

Hewei Yu
Professor of South China University of Technology. His research interests focus on integration technology of heterogeneous wireless networks. He directed several projects supported by the National Natural Science Foundation of China, Science & Technology projects of Guangdong Province, and Guangzhou City.

Meiling Zhou
Got her Master degree from South China University of Technology. Her research direction is handoff technology of Proxy Mobile IPv6. Now she is working in the GF Fund Management Co. Ltd.
ANALYSIS OF MOBILE PAYMENT
INFLUENCING FACTORS

Sitalakshmi Venkatraman
School of Engineering, Construction and Design (IT),
Melbourne Polytechnic, Australia

ABSTRACT

With the exponential proliferation of mobile devices in the consumer market, wireless e-business is emerging as a key area to revolutionise industries. In the past few years, industry has witnessed an increase in the adoption of mobile payment and billing methods that leverage on wireless technologies. Yet, the success of mobile payments in businesses much depends on many factors such as, type of wireless technologies used, security options available, the players involved and their influencing m-business models. This paper examines mobile payments in both technical as well as business perspectives. It identifies and analyses the influencing factors from multi-dimensional aspects that would be useful for adopting mobile payments.

KEYWORDS

Mobile payments, Wireless technologies, Business models. M-commerce

1. INTRODUCTION

Wireless e-business termed as m-commerce has the main advantage of portability with ubiquitous business transactions that can be performed anywhere and anytime [1]. In addition to the key mobile features of location-based and personalised information services, payments embedded in mobile apps or websites provide value-added facility that aid in the business agility of organisations in this competitive business environment [2][3]. With recent developments in wireless communication technologies using 3G evolved to 4G networks, mobile Internet and multimedia streaming have gained popularity and provide increased channels of interaction with customers [4][5]. Hence, organisations expect that m-commerce provides a cost-effective improvement in their operational efficiency, resulting in an increase in profitability. Businesses have started to consider mobile payments as value-added services providing opportunities for their marketing, promotion and branding [6][7].

Apart from business perspectives, looking at consumer trend in the last decade, most consumers have been using mobile phones beyond the traditional services of voice calls and SMS. By the end of 2016, two thirds of the world’s population acquired a mobile subscription with a surge in mobile broadband adoption via smartphones [8]. Mobile devices now account for 60% of the connections with new services and dedicated apps leveraging on mobile internet that could make common man's daily activities quite convenient. Hence, consumers' expectations from these smart devices have risen dramatically and mobile payment services are now being recognised to be important. Many consumers from countries such as, Finland, UK, North America, Japan, Korea and Singapore started to adopt mobile payments more than a decade back [9][10] as their preferred mode of cashless transactions in their everyday lives - while transacting with vending...
machines, online ticketing agencies, public transport systems and merchant stores. A recent survey in UK indicates that 56% of smart phone owners purchased a product using a mobile app transaction in 2015 and mobile banking transactions via apps have now been envisaged by consumers [11]. However, global reports indicate that mobile revenue growth is slow [8][12]. There is lack of research in analysing this situation from current technological and business perspectives and such investigations could have a positive influence on mobile payment adoption [7][13]. This provides the key motivation for the study reported here.

This paper is aimed at providing an analysis of the trends in mobile payments, from technology, business, as well as consumer perspectives for analysing the influencing factors. The remainder of the paper is organised as follows. Section 2 describes the need for the study. The influencing factors from various facets of mobile payments are identified and analysed in section 3. Finally, concluding remarks are provided in section 4.

2. THE NEED FOR THE STUDY

Despite potential advantages of m-commerce, mobile payments face major challenges as financial transactions via mobile devices demands a multi-faceted approach involving technology innovations, apps and service offerings as well as global business trends [14][15]. In addition, with the emergence of player-centric mobile payment services, new mobile payment business models are being introduced [16]. Both remote and proximity type of mobile payment systems have witnessed evolution of technologies such as Infrared (IR) and Bluetooth combined with services such as Interactive Voice Response (IVR), Short Message Service (SMS), Radio Frequency Identification (RFID)/2D bar codes, Wireless application Protocol (WAP), contactless interface to chip-enabled payment technology, Near-Field Communication (NFC) enabled mobile phones, contactless stickers, tags, mobile browsers, mobile apps and m-wallets [5][17].

With the existing plethora of wireless technologies, both merchants and consumers are clouded with a range of options available to them for adopting mobile payments and are not sure if they are taking the road to success or failure [4][7]. In the past decade, mobile payments were used for predominantly non-banking transactions with several failed initiatives such as Dencoin and Google Wallet even in developed countries [18][19].

On one hand, the banking sector is gradually embracing digital payments capitalising on the IT players such as Apple and Google to compete in the market that may shape its future [12]. On the other hand, the success or failure of mobile payments is based on the consumer acceptance [13]. Recent studies conducted worldwide indicate that mobile payment growth has been slower than anticipated [8][12]. This is because, mobile payments are not just a new technology added on to existing consumer networks, but actually have their own set of historical and normative roots in development practices with an emphasis to be given on market subjects and consumer subjects [7]. Previous studies have predominantly focussed on the technology and payment options available to a consumer based on factors such as location (remote/proximity), payment value (micro/macro amount), and charging method (post-paid/pre-paid), and the associated security protocols [20][21][22]. There is lack of literature that analyses the current multi-faceted dimensions of mobile payments [23][24][25]. There is little guidance available for merchants and consumers to review the various mobile payment options and to adopt those mobile payment methods that suit their requirements.

Due to the paucity of research in this area, this paper analyses various influencing factors for adopting mobile payments. This would help businesses and consumers to make a more systematic review before they embark on the right kind of mobile payments that would add value to their everyday activities. This paper takes a modest step in this direction.
3. **INFLUENCING FACTORS FOR MOBILE PAYMENTS**

Although some industries such as, retail, entertainment and banking have been successfully adopting mobile options for billing and payment in certain countries, many industries are still cautious in offering mobile payment methods for their customers. This is partly because there are several issues related to mobile payments in an enterprise. On analysing the influencing factors, these could be broadly classified under three main dimensions:

- mobile payment players,
- payment methods and
- enabling wireless technologies

**Mobile Payment Players**

There are many parties or players contributing to the mobile payment value chain. In order to access a payment service, a user or subscriber initially has to come into an agreement with a network provider for accessing telecommunication and the payment services provided by the specific operator. If mobile internet is used, an additional contract between a subscriber and an ISP is required. However, with the introduction of 3G/4G technologies, mobile internet could be accessed directly through their mobile operator. Recent evolutions have enabled users to adopt new mobile payment and billing models for accessing various networks and services independent of the network provider [5][26]. In Table 1, we have summarised a list of players involved in influencing mobile payments in businesses.

![Table 1. Mobile Payment Players and Influencing Factors](image)

The abovementioned players need not be exactly mapped onto various business parties but could be an aggregation of business entities. For instance, the network service provider could be aggregating the services of mobile network operators, content providers and Internet service providers. Alternatively, there are cases where the network operator covers the functionality of an Internet service provider, while the content provider caters to payment services as well. Proprietary mobile payment solutions are now being introduced as issuer wallets to compete with device manufacturers such as Apple, Samsung, etc. Hence, an evolution of such industry players have a major influence on the consumer adoption of mobile payments.

**Payment Methods**

Though mobile payment services were traditionally managed by mobile network operators for many years using player-centric models, recently we find the emergence of new payment and billing models that allow consumers to access the payment services independent of who the
The issue of mobile payment adoption is further compounded by the types of payment methods supported here. Commonly used payment methods include pre-paid cash card, micro-billing, person-to-person (P2P) and credit card payments. Some of the main factors that influence these payment methods in m-commerce are identified in Table 2.

### Table 2. Payment Methods and Influencing Factors

<table>
<thead>
<tr>
<th>Payment Methods</th>
<th>Influencing Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-paid cash card</td>
<td>Small transaction amount, Top-up methods and Availability to anyone</td>
</tr>
<tr>
<td>Micro-billing</td>
<td>Billing options, Service charge and Transaction maintenance</td>
</tr>
<tr>
<td>Person-to-person (P2P)</td>
<td>Direct credit or debit options, Service charge, Account maintenance and Security</td>
</tr>
<tr>
<td>Credit card</td>
<td>Service provider trust, M-wallets (H/W and S/W integration) and Security</td>
</tr>
</tbody>
</table>

Similar to Internet-based online payment methods that support pre-paid (debit) and post-paid (credit) services, mobile payments require an integrated flexible model that support quality of service (QoS) with one-stop billing schemes for the end-users [21].

### Enabling Wireless Technologies

The aforesaid payment methods could be adopted to make mobile payment transactions using different wireless technologies. Such mobile payments go through a set of wireless security levels which is a key factor in providing consumer trust and confidence [28]. The enabling wireless technologies could support mobile security elements that are either removable or non-removable from the mobile device [24]. Removable security elements could be a single chip, a dual chip or a dual slot, while non-removable security elements could support embedded hardware and software modules, including personalised biometric technologies [13]. In table 3, we identify a list of wireless technologies commonly adopted in businesses and their influencing factors for mobile payment adoption.

### Table 3. Enabling Wireless Technologies and Influencing Factors

<table>
<thead>
<tr>
<th>Enabling Wireless Technologies</th>
<th>Influencing Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radio Frequency Identification (RFID)</td>
<td>Transponder &amp; reader frequencies / range, Encryption methods and Payment methods</td>
</tr>
<tr>
<td>Short Message Service (SMS)</td>
<td>SMS gateway QoS, Payment methods, Merchant port ID and Push technology</td>
</tr>
<tr>
<td>Wireless Access Protocol (WAP) using WIM chip</td>
<td>Dual chip phones, WIM certificate, Payment methods, Security and Account Maintenance</td>
</tr>
<tr>
<td>Infrared (IR)</td>
<td>Point-to-point range, Fault-tolerance and Payment methods</td>
</tr>
<tr>
<td>Bluetooth</td>
<td>Security, Transmission time and Data volume</td>
</tr>
<tr>
<td>Near Field Communication (NFC)</td>
<td>Tokenisation, Mobile phones, NFC reader, Payment methods</td>
</tr>
<tr>
<td>Internet</td>
<td>Mobile Apps for Payments, M-wallets, Security, Payment Methods, Account Maintenance &amp; Fees</td>
</tr>
</tbody>
</table>

A mobile payment system typically espouses several wireless technologies with different possible payment methods and with different player-centric influencing factors. With security being a major concern in mobile payments, security modules with digital signatures are being implemented in the form of WAP Identity Module (WIM) cards. From traditional, simple Interactive Voice Response (IVR) systems adopted in phone banking to enhanced WIM chips in mobile devices and now with sophisticated mobile apps, the use of different wireless technologies
involves different levels of security as the device moves in and out of multiple wireless networks leading to many security issues [17].

4. CONCLUSIONS

The adoption of mobile payment is constricted by multi-faceted issues that can be approached from several angles. This paper identified three main dimensions such as mobile payment players, payment methods and enabled wireless technologies. An analysis resulted in their influencing factors for a successful mobile payment adoption. These influencing factors would facilitate consumers, businesses and researchers to consider the different mobile payment solutions, from technology as well as business perspectives for decision making.

REFERENCES


[12] EY White paper (2017), Mobile money — the next wave of growth, Ernst and Young.


AUTHORS

Dr. Sita Venkatraman earned her PhD in Computer Science, with a doctoral thesis titled “Efficient Parallel Algorithms for Pattern Recognition”, from National Institute of Industrial Engineering in 1993. In the past 30 years, Sita’s work experience involves both industry and academics - developing turnkey projects for IT industry and teaching a variety of IT courses for tertiary institutions, in India, Singapore, New Zealand, and more recently in Australia since 2007. She currently works as Lecturer (Information Technology) at the School of Engineering, Construction & Design, Melbourne Polytechnic, Australia. She also serves as Member of Register of Experts at Australia's Tertiary Education Quality and Standards Agency (TEQSA).

Dr. Venkatraman’s recent research areas are predominantly in Business Intelligence, Data Mining, Cloud Computing, Information Security and Health Informatics. She has published eight book chapters and more than 120 research papers in internationally well-known refereed journals and conferences..
GRC-MS: A GENETIC RULE-BASED CLASSIFIER MODEL FOR ANALYSIS OF MASS SPECTRA DATA

Sara Al-Osimi\textsuperscript{1} and Ghada Badr\textsuperscript{2}

\textsuperscript{1}Department of Computer Science, Shaqra University, Riyadh, KSA
\textsuperscript{2}IRI - The City of Scientific Research and Technological Applications, Alex, Egypt

\textbf{ABSTRACT}

Many studies use different data mining techniques to analyze mass spectrometry data and extract useful knowledge about biomarkers. These biomarkers allow the medical experts to determine whether an individual has a disease or not. Some of these studies have proposed models that have obtained high accuracy. However, the black-box nature and complexity of the proposed models have posed significant issues. Thus, to address this problem and build an accurate model, we use a genetic algorithm for feature selection along with a rule-based classifier, namely Genetic Rule-Based Classifier algorithm for Mass Spectra data (GRC-MS). According to the literature, rule-based classifiers provide understandable rules, but not accurate. In addition, genetic algorithms have achieved excellent results when used with different classifiers for feature selection. Experiments are conducted on real dataset and the proposed classifier GRC-MS achieves 99.7\% accuracy. In addition, the generated rules are more understandable than those of other classifier models.

\textbf{KEYWORDS}

Mass spectrometry, data mining, biomarkers, rule-based classifier, genetic algorithm.

\textbf{1. INTRODUCTION}

Mass spectrometry (MS) is an efficient technique that has been widely used in many disciplines, such as science, engineering, and biology. Recently, MS has been used in the bioinformatics field to identify the amounts of chemical and biological materials in human tissue or serum to use later as biomarkers. These biomarkers can be used as measures for clinical assessments to monitor and predict individuals’ health conditions in order to plan suitable therapeutic interventions [1]. However, because the data generated using the MS technique is so huge and extensive, it is difficult to extract any useful knowledge or biomarkers; Many studies have been done to develop data mining analysis tools (i.e., classification, clustering, correlation analysis, etc.) for the interpretation and extraction of accurate knowledge from MS data. However, the results of most of these studies have not been satisfactory. Even when the studies do achieve good results, experts may struggle to understand them. According to the literature [2], rule-based classifiers yield acceptable results when they are applied to the analysis of discrete data. In addition, these
classifiers have the unique ability to provide very meaningful outcomes. However, unfortunately, rules-based classifiers do not achieve the quality required for analysis of MS data.

In this paper, we propose an efficient and meaningful approach that uses Genetic Algorithms (GAs), namely GRC-MS, to select features and then build a rule-based classification model with the objective of classifying and understanding MS data. We also test our proposed approach on a real dataset of ovarian cancer patients in order to measure the accuracy of the proposed approach. The proposed approach is intended to be a general framework that can be used for the analysis of any MS data or related continuous data. To the best of our knowledge, the combination of rule-based classifiers with GAs as the feature selection technique has not yet been applied to MS data.

This paper is organized as follows: Section 2 provides a background about the MS techniques: preprocessing, some feature selection, and classifiers that are used for MS data. Section 3 refers to some of the studies that use GA technique as a feature selection approach for MS data. In addition, it summarises some of the studies that use rule base techniques as classifiers for MS data. Section 4 explains the steps of our proposed approach, GRC-MS. The experimental setup and results on a real dataset are presented in Section 5. Section 6, discuss the results. Finally, Section 7 concludes the paper and discusses future work.

2. BACKGROUND

Mass spectrometry (MS) is a recently developed technique that is used to identify, analyze, and determine the elements, molecules, and atomic structures of any given sample [3]. MS quickly and accurately determines the relative numbers of molecules present in complex biological or chemical samples by transforming these samples into ions, detecting their mass-to-charge ratios (m/z), and then measuring the intensity of each ion type [4]. This technique is used primarily to study the effects of ionizing energy on sample molecules [3]. It has several beneficial characteristics, such as speed and sensitivity. Moreover, because MS has a variety of possible applications, it is preferable to other analytical methods and, as a result, has progressed rapidly over the last decade. Today, MS is used in a number of applications, such as biochemical problems, pollution control, atomic physics, food control, forensic science, reaction kinetics, geochronology, inorganic chemical analysis, process monitoring, and so on [4].

2.1 Proteomics

Proteomics, a term that is first coined by Australian scientist Marc Wilkins in 1994, is an emerging area in bioinformatics [7]. It provides information about proteins and their interactions in the human body. The major aim of most proteomic studies is the detection of proteins of interest, which are known as biomarkers. The term “biomarkers” refers to protein molecules that facilitate the detection of a particular cell type and that identify cell characteristics, such as cells’ ability to perform their functions. [8]. The discovery of biomarkers in MS data is useful for the early diagnosis of diseases. Most researchers hope to discover novel and powerful diagnostic proteomic tools to detect these biomarkers [8]. Recently, several techniques have been developed for analyzing bodily fluids, such as human serum, human urine, and, in some studies, tumor tissue, to achieve protein profiling. Commonly, the analysis of body fluids is accomplished using MS techniques [9]. Two major techniques are intended for proteomic analysis: MALDI and SELDI. MALDI-TOF MS is a new and widely used technique for discovering biomolecules, such as proteins with molecular masses between 400 and 35000 Da, in samples [6].
2.2 Mass Spectrometry

MS experiments are generally conducted in three main stages: the data generation stage, the data preprocessing stage, and the data analysis stage [4] [5]. In the first stage, MS techniques generate data that are represented as a huge sequences of pairs, called matrix, spectrum, or MS data [4]. This spectrum contains mass-to-charge ratio values and intensity values [6]. The mass-to-charge ratio values (which are represented on the x-axis) depend on the molecular mass detected in the sample, and the intensity values (which are represented on the y-axis) depend on the quantity of molecules detected in the sample (Figure 1) [6]. Depending on the resolution of the MS technique, a spectrum can contain hundreds or thousands of pair values [7]. Data preprocessing involves cleaning the data and improving their quality. On the other hand, during data analysis, data mining or pattern extraction techniques are applied to extract knowledge.

2.3 Data Mining in Mass Spectrometry

Data mining is a well-known approach that is used in science and business to extract useful information from large and complex datasets [7][2]. The steps involved in data mining include (Figure 2) data preparation, feature selection, model development (or pattern recognition), and model assessment. The following section focuses on the basic algorithms used in data mining for application to mass proteomic data. However, as previously mentioned, MS data are high in dimensionality, and they cannot be analyzed through the direct use of data mining techniques. Preprocessing the MS data is a crucial step in improving the data quality—and, thus, improving the quality of the classifier algorithms [6].

- Preprocessing MS Data

MS data or spectra are commonly influenced by errors or noise that occur during the sample preparation or the insertion into the device or by noises generated by the device itself [4]. Using the raw MS data directly for the analysis process is not effective because contaminants like noise, m/z measurement errors, and matrix size affect the results [6] [7]. In addition, because of the dimensional complexity of the spectra, efficient results cannot be obtained through the direct application of data mining algorithms or pattern extraction techniques. Therefore, cleaning the MS data is critical. To achieve clean data, different preprocessing techniques are applied to the MS data before the application of any data mining technique such as reducing noise and
smoothing data, normalization, data reduction by binning, peak extraction, and peak alignment. These techniques can be used alone or in combination [10].

Figure 2. Typical flowchart of the critical steps in data mining and examples of the techniques available for MS data analysis.

- **Feature Selection Techniques**

The MS technique produces high-dimensional data. Compared to the number of samples, a greater number of peaks needs to be analyzed (high features-to-sample ratio datasets) [11]. Most of the problems in analyzing data stem from the size and the complexity of the datasets that are represented in tables of rows and columns. Rows represent records or cases, and columns represent data dimensions, features, and attributes [7]. In the analysis of MS data, the extraction uses the intensity of every peak in the spectrum as a feature. The number of features (or peaks) is usually large (e.g., 17,000 peaks), while the number of samples is usually small (e.g., 140 patients) [12]. However, features often contain noise with a very little or no informational value. Thus, it is necessary to select features from a large set of those likely to be useful in predicting the outputs of interest. To solve this problem, after the data is pre-processed, a feature selection phase step is performed. This step aims to detect the main parts of the spectrum that might provide a
better understanding of the data’s important features, which could be used in the analysis phase [11].

Feature selection techniques can be divided into three categories [12]: filter, wrapper, and embedded. The filter technique analyzes each feature independently and eliminates features one at a time based on how they correlate with the target. Feature subsets are selected based on evaluation criterion, such as information gains. This is a simple and quick process that is sometimes referred to as independent feature selection. Moreover, filter selection methods are relatively computationally efficient [11] [12]. Examples of independent feature selection techniques used with MS or high-dimensional data include: statistical tests (i.e., t-tests [13] [14], Wilcoxon tests [17], χ² tests [18]), information gains [19], and so on. The wrapper techniques simultaneously analyze features in groups or subsets and build the analysis model [11] [12]. Classifiers are used to assess (several) features or feature subsets. Although the process is computationally busy and potentially very time-consuming, since this technique typically requires an evaluation of every scheme at every iteration, it discovers critical information that is typically lost in independent features analysis [16] [20]. Examples of wrapper feature selection techniques that are used with MS or high-dimensional data include: genetic algorithms [21] [22], sequential searches [23], and estimations of distribution algorithms [24]. In embedded techniques, the search for a best set of features is made into the classifier construction. They learn which set of features can best contribute to the accuracy during the creation of the model [12]. These techniques make no distinction between learning and feature selection. Embedded techniques have the advantage of including the interactions with the classification model, while simultaneously being far less computationally intensive than wrapper methods [12]. Examples of embedded feature selection techniques that can be used with MS or high-dimensional data include random forest techniques [25] and weight vector support vector machine techniques [26].

- **Classifiers and Decision Models for MS Data**

For MS Data, usually classification or supervised learning uses to predict or classify new cases. Where, previous knowledge about classes can be used to classify new cases. The previous knowledge is built using a training dataset, which includes input values and their output classes. In the training stage, the training dataset is used to define how the features are to be selected and combined to distinguish among the different classes. In the testing stage, the weighted features are applied to classify a new test dataset. The test dataset’s class is not known, and the dataset has never before been seen by the model. If the model classifies new cases correctly, it is a good model. A wide range of algorithms, such as decision tree algorithms, SVMs, ANNs, and so on, have been developed for classification. In this subsection, we well indicate to some of well-known classifiers that used for MS data.

i. **Decision Tree (DT) Classifier**

Decision tree (DT) a hierarchical tree structure model that is described as a set of rules, presented in a visual form that is very easy to understand. The DT was used by Vlahou et al. [28] to analyze the MS data and it is achieved 80% accuracy in discriminating between ovarian cancer patients and healthy controls. In addition, Su et al. [29] used DT to analyze the MS data and they obtained 85.3% accurate.
ii. Artificial Neural Networks (ANNs) Classifier

Artificial neural networks (ANNs) are another popular classifier used in MS data analysis. Ward et al. [13] used an ANN algorithm to analyses non-cancer and colorectal cancer samples via SELDI to identify colorectal cancer biomarkers. The ANNs with the seven highest peaks obtained 95% sensitivity and 91% specificity. Also, Chen et al. [30] used ANNs to diagnosis colorectal cancer. The proposed approach obtained 91% sensitivity and 93% specificity.

iii. Naive Bayesian (NB) Classifier

The naive Bayesian is “a simple probabilistic classifier based on the Bayesian theorem with the (naive) independence assumption” [31]. Zheng [31] compared the performance of the naïve Bayesian (NB) and the logistic regression (LR) on MS data. They found the average performance of the NB (around 90%) and the logistic regression depended on the amount of training data.

iv. Support Vector Machine (SVM) Classifier

Support vector machines (SVMs) attempt to find the best hyperplane line that separates all class A data points from class B data points. Wagner et al. [32] found that the linear SVM was the only classification method that obtained robust performance (98% accuracy). Also, Prados et al. [19] achieved 97% sensitivity and 71% specificity when used SVM-based model to classify MS data.

- Performance Assessment of Models

The last stage of the data mining modeling process is the assessment or validation of the model. Below, we will discuss the measures of accuracy, sensitivity, and specificity.

The classification of accuracy is calculated by “comparing the ratio of the number of correctly classified samples to the total number of samples in the test data” [33]. However, when the spread of a certain class is greater than that of other classes, the majority class will create unequal results. In this scenario, the accuracy measure will not be true. Most MS data analysis studies have used accuracy to report their results [33] [34].

There are four possible results when test decisions are built for data with two class samples: true-positive, true-negative, false-positive, and false-negative [33]. The true-positive rate is known as sensitivity. It represents the ratio of the number of correctly classified positive samples to the total number of positive samples. When the effect of incorrectly predicting a diseased person as healthy is high, high sensitivity is preferred in medical diagnoses. Specificity refers to the false-positive rate, or the probability that a healthy subject will be incorrectly classified as unhealthy [14]. When a false alarm would result in unwanted tests or treatments, high specificity is desirable here [7]. In very good classification, both sensitivity and specificity should be high, though different levels of these measures are accepted depending on the application. However, it is very hard to compare the results of different studies using only measures of sensitivity and specificity [33]. Up to our knowledge, many approaches failed to achieve high accuracy. Even when high accuracy is obtained, the “black box” nature of these proposed approaches is a major issue. To address this problem and to build an accurate and understandable model, we propose to use a rule-based classifier approach along with using GAs for feature selection.
3. LITREATURE REVIEW

In this section, we explore some of the studies that use GA technique for feature selection on MS data. In addition, we illustrate some studies that use rule-based techniques as a classifier on spectrum data with classifiers.

A. Genetic Algorithm Based Feature Selection for MS Data

One popular algorithm that is used for feature selection purpose is a genetic algorithm (GA). A GA searches for optimal MS data features or biomarkers to use in the mining stage in order to distinguish patients from controls in an accurate way. Here we discuss GA as a feature selection approach for MS data. Many studies have used GA for feature selection before applying a classifier. In 2009, Reynès et al. [35] developed a new model using a GA for feature selection and a very simple tree as a classifier. The GA in this model sought to choose a set of interesting features in a spectrum to achieve the best split points in the tree. First, the authors applied preprocessing steps to the dataset. The dataset contained 162 ovarian cancer samples and 91 control samples. Of these, 46 control samples and 81 cancer samples were randomly chosen for use as a training set; the rest (45 control and 81 cancer samples) were later used for testing. The authors obtained 98% accuracy after building the tree with three different peaks (245 Da, 434 Da, and 649 Da). The major issue in this technique when GAs return large numbers of features the DT become large and difficult to understand.

In 2004, Mohamad et al. [36] proposed a new model for applying a GA to seek and identify potential informative features using an SVM classifier. Experimental results on a breast cancer dataset (which contained 200 samples for training and 77 samples for testing) and a leukemia cancer dataset (which contained 38 samples for training and 34 samples for testing) showed the usefulness of the proposed approach for low- and high-dimension data. The authors obtained 82% accuracy for the breast cancer dataset, with 8 features, and 100% accuracy for the leukemia cancer dataset, with 50 features. In 2004, Li et al. [37] proposed a novel model used a GA for the feature selection stage and an SVM method as a classifier. The MS dataset used included 91 control samples and 162 samples from patients with ovarian cancer. Both feature selection approaches (filter and wrapper) were explored. The results showed 98% accuracy the proposed model was applied with a filter approach.

In 2002, Petricoin et al. [38] used GA for feature selection with a cluster algorithm. The proposed algorithm was applied to a training set containing 50 ovarian cancer samples and 66 control samples. The authors obtained a sensitivity of 100%, a specificity of 95%, and a rounded accuracy of 94%. In 2007 Shah and Kusiak [39] proposed a model using GA for feature selection and DT and SVM as classifiers. They applied the proposed model to three different datasets for ovarian cancer, prostate cancer, and lung cancer. The proposed model had high classification accuracy when applied to the ovarian cancer and lung cancer dataset, such that it was able to recognize the most significant features. Table 1 below summarizes some of the relevant research in this field that used genetic algorithm as feature selection for mass spectrum data.

After we review some studies that using GAs for feature selection in the analysis of MS data we found that most approaches obtained a very good accuracy results. However, there are some major challenges. For example, there is no guarantee that GAs will always simultaneously find the best solution and in the same time the minimum number of discernment features. When a GA
obtains a large number of features, there will be problems using certain classifiers, such as DTs. In such cases, DTs may become very large, complex, and difficult for experts to understand. Some researchers have tried to solve this problem by adding constraints to the GA. This was the case in [39], in which the authors repeated the selection process when the number of selected features was more than 100; however, this process took a long time. Moreover, in [35], the authors added a constant to the fitness function to help it select the fewest number of features possible. However, the constant did not always work in obtaining a minimal number of features.

Table 1. Some of the research using GAs as features selection for analysis MS data

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Year</th>
<th>Disease</th>
<th>Data</th>
<th>Feature Selection Method</th>
<th>Data Mining Algorithm</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reynès et al. [35]</td>
<td>2009</td>
<td>Ovarian Cancer</td>
<td>▪ 253 ovarian cancer serum samples.</td>
<td>DT</td>
<td>98% Accuracy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>▪ 162 samples from patients with ovarian cancer and 91 samples from healthy patients.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mohamad et al. [36]</td>
<td>2004</td>
<td>Breast Cancer</td>
<td>▪ 200 training samples and 77 test samples.</td>
<td>SVM</td>
<td>82% Accuracy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Leukemia Cancer</td>
<td>▪ 38 training samples and 34 test samples.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li et al. [37]</td>
<td>2004</td>
<td>Ovarian Cancer</td>
<td>▪ 253 ovarian cancer serum samples.</td>
<td>SVM</td>
<td>100% Accuracy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>▪ 162 samples from patients with ovarian cancer and 91 samples from healthy patients.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Petricoin et al. [38]</td>
<td>2002</td>
<td>Ovarian Cancer</td>
<td>▪ 216 ovarian cancer serum samples.</td>
<td>Genetic Algorithm</td>
<td>94% Accuracy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>▪ 100 training samples and 116 test samples.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
After excluding all 100% accurate results due to the high chance of over-fitting, we found that the best accuracy achieved was 98.66%, which was obtained by the SVM classifier. Thus, we seek to obtain a better accuracy than this one, while simultaneously building a classifier that is easy to understand. We propose the use of a rule-based classifier, which can be understandable even when GAs return large numbers of features. This is because a rules-based classifier is easier to understandable than a DT, especially with higher numbers of features. Finally, we also seek to obtain higher classifier accuracy than that achieved by the SVM.

### B. Rule-Based Classifier models for MS Data

Several machine-learning classifiers, such as DTs, SVMs, and K-nearest neighbor classifiers, have been used to successfully classify MS data. These have all achieved high predictive accuracy. However, the black-box nature of these classifiers presents major issues for developers [40] [41]. By contrast, the IF-THEN rule-based classifier can obtain satisfactory predictive accuracy, while also being easier to describe and interpret by humans than other classifiers, due to its readable IF-THEN rule structure [42]. The challenge is the extraction of a small, accurate and easy-to-interpret sets of IF-THEN rules from high-dimensional MS data. In the following, we will review various studies that have used IF-THEN rule classifiers to classify of MS data. We will then discuss these papers in order to provide a simple introduction for the development of this type of classifier.

In 2006, Ressom et al. [41] proposed a novel classifier for classifying MS data using a fuzzy IF-THEN rule-based structure. For feature selection, the authors used ant colony optimization
(ACO) with an SVM. They hoped that the combination of these two methods in the feature selection step would improve the quality of the potential biomarker identification and build an accurate fuzzy IF-THEN rules classifier. The authors collected 150 serum samples of hepatocellular carcinoma (HCC) diseases that were taken from Egypt between 2000 and 2002. Of these, 78 samples were taken from patients with HCC, and 72 samples were taken from normal individuals. After they preprocessed the samples, the authors selected 100 samples randomly as a training set, including 50 samples from the HCC patients and 50 samples from the healthy individuals. The remaining samples (28 from the HCC patients and 22 from healthy individuals) were used as a testing set for performance evaluation. The authors applied a combination of ACO and SVM to extract useful biomarkers in the feature selection stage. They found six m/z candidate biomarkers, as follows: 1863.4-1871.3, 2528.7-2535.5, 933.6-938.2, 1737.1-1744.6, 4085.6-4097.9, and 1378.9-1381.2 Da. These six m/z candidate biomarkers were used as inputs to the IF-THEN rules classifier. The prediction accuracy of this classifier was estimated using a four-fold cross-validation method. Then, the authors used the ACO algorithm to select four rules from among the 4095 candidate rules extracted from the training dataset with the candidate biomarkers. The IF-THEN rules distinguished HCC patients from controls in the testing dataset with 91% sensitivity and 89% specificity.

Assareh and Moradi [43] proposed a model that used a t-test to select the best features and a IF-THEN rules classifier to classify the MS datasets. The dataset was for ovarian cancer, and it was made available to the public through the American National Cancer Institute (NCI) website. The ovarian cancer dataset contained 253 samples, of which 91 samples came from healthy individuals and 162 came from ovarian cancers patients. Before addressing these datasets, the authors used preprocessing to clean the datasets to enhance the classifier’s performance. They binned all of the M/Z points as candidate biomarkers and applied a t-test to select the best candidate biomarkers. The t-test eliminated the biomarkers that were locally correlated, since these could correspond to the same peptide. The authors found three m/z candidate biomarkers. The proposed method achieved acceptable accuracy (86.93%) compared to two classification methods: LDA (74.24%) and KNN (68.18%).

In 2011, Wang and Palade [44] proposed a new Multi-Objective Evolutionary Algorithms-based Interpretable Fuzzy (MOEAIF) model. This model used Fuzzy C-Mean Clustering-based Enhanced Gene Selection (FCCEGS) for feature selection with fuzzy IF-THEN rules to analyze high-dimensional data, such as microarray gene expressions and MS data. The proposed model was evaluated on proteomics mass spectroscopy data from an ovarian cancer dataset containing 253 samples (91 from healthy individuals and 162 from ovarian cancer patients). Some preprocessing steps were applied to the dataset. The authors extracted eight fuzzy IF-THEN rules from the dataset (average rule length of two) using six candidate biomarkers. The candidate biomarker MZ6880.2 and the feature MZ18871.5 played important roles in most of the rules. This proposed MOEAIF model achieved 63.75% accuracy. Table 2 below summarizes some of the relevant research in this field.

In reviewing the various research papers using rule-based classifier to analyze MS data, we found that the research related to these rule-based classifiers was still very active. Various researchers had tried to improve the black-box problem of most classifiers while simultaneously achieving high predictive accuracy. Each paper proposed a model for obtaining a certain number of IF-THEN rules that would be easy for experts to understand and manipulate. However, the major challenge is improving rule accuracy by finding the best set of features. Several authors have
attempted to use different feature selection methods; however, up to our knowledge, none has achieved a higher classification accuracy.

Table 2. Research using IF-THEN rules as classifiers for the analysis of MS data.

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Year</th>
<th>Diseases</th>
<th>Data</th>
<th>Features Selection Method</th>
<th>Data Mining Algorithm</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ressom et al. [41]</td>
<td>2006</td>
<td>HCC</td>
<td>▪ 150 serum samples of HCC diseases&lt;br&gt;▪ 78 samples from patients with HCC and 72 samples from healthy patients</td>
<td>ACO-SVM algorithm</td>
<td></td>
<td>91% sensitivity and 89% specificity</td>
</tr>
<tr>
<td>Assareh and Moradi [43]</td>
<td>2007</td>
<td>Ovarian cancer</td>
<td>▪ 253 ovarian cancer serum samples.&lt;br&gt;▪ 162 samples from patients with ovarian cancer and 91 samples from healthy patients</td>
<td>T-test</td>
<td></td>
<td>86.93% Accuracy</td>
</tr>
<tr>
<td>Wang and Palade [44]</td>
<td>2011</td>
<td>Ovarian cancer</td>
<td>▪ 253 ovarian cancer serum samples.&lt;br&gt;▪ 162 samples from patients with ovarian cancers and 91 samples from normal.</td>
<td>Fuzzy C-Mean Clustering based Enhanced Gene Selection method (FCCEGS)</td>
<td>IF-THEN Rule-Based</td>
<td>63.75% Accuracy</td>
</tr>
</tbody>
</table>

4. GENETIC-RULE-BASED CLASSIFIER MODEL FOR MS DATA (GRC-MS): A PROPOSED APPROACH

Given MS datasets of any diseases, the GRC-MS model has the following input and output:

Input: MS data obtained from controls (healthy individuals) and patients.

Output: A set of rules expressed as: I⇒C, where I refers to a set of features or biomarkers and C refers to a class label (i.e., healthy or patient).
The steps of the GRC-MS model are shown in Figure 3. The details of each step are explored in the following subsections:

**STEP 1: Preprocess Raw MS Data**

Each point on a spectrum is represented by two measurements: m/z and the intensity value. Sometimes, these points are affected or distorted by noise. Thus, preprocessing is needed to clean the MS data of noise and contaminants [9]. In addition, the preprocessing step must reduce or decrease the dimensions of the spectrum; this is important later for obtaining an efficient algorithm [33]. In this model, to correct the m/z and intensity values, we use the following steps: (A) Reduce Noise and Amount of Data and (B) Make Spectra Comparable.

**A. Reduce Noise and Amount of Data**

To remove a chemical noise baseline from a spectrum without harming the data is a challenging problem, since the wrong baseline correction may damage the spectrum, resulting in the wrong peak shape, peak position, or peak width [10]. We will use a function to estimate a low-frequency baseline. Then, we will subtract this baseline from the spectrum. Figures 4 show how the function corrects the baseline. These examples were taken from real dataset (ovarian cancer dataset) before and after the baseline’s removal.
To remove electrical noise, it is important to know that spectra usually contain a combination of noises and signals. Thus, a spectrum must be de-noised to improve the validity and precision of the observed m/z values of the spectrum peaks. To accomplish this, we use Lowess smoothing and polynomial filters.

B. Make Spectra Comparable

Normalization of spectra is needed to make MS data independent of experimental differences. Normalization enables us to compare different samples, since the peak values of different spectrum fractions may be incomparable [19]. In this model, we will use the direct normalization function to calculate a re-scaled intensity value as needed. In addition, to make spectra comparable, peak alignment determines which peaks from the different spectra samples correspond to the same peak. For this, we use a sample alignment function that allows us to use a dynamic programming algorithm to assign the observed peaks in each spectrogram to the common mass/charge reference vector, if needed.

STEP 2: Use Genetic Algorithms for Features Selection

After the data preprocessing, we implement a feature selection stage, which seeks to achieve better understanding of the important features of the MS data in order to improve the classification phase later. In our model, we use GAs, which try to find optimal search solutions for problems with large datasets. Running on MS data, GAs attempt to find small sets of biomarkers that separate patient cases from control cases. This set of biomarkers, or features, is called a chromosome, such that every biomarker corresponds to a biological sample’s
measurements at a given \( m/z \) value (Masse). Each chromosome is evaluated by a fitness function that attempts to find the best chromosome (set of biomarkers) for separating patients from controls. GA follows the steps outlined below:

A. **Encoding Solutions (choose from initial chromosome, generation, and population)**

Each “chromosome” (i.e., mathematical entity, not biological) consists of \( d \) different biomarkers or features (called genes) that are initially randomly selected from all features (since most studies used all of the MS data as features after the preprocessing steps).

B. **Determine Fitness of Population**

The fitness value of each chromosome is determined by the chromosome’s ability to classify the training set samples into patient and control groups. In our model, we will use to compute fitness values:

\[
\text{Fitness value} = \text{a posteriori probability} + \text{Error rate of a linear classifier}.
\]

*Note: Repeat from Step C to Step G until terminated.*

C. **Selection Operator (select parents from population)**

The chromosome with the best fitness value is entered into the next generation, and the remaining positions are filled according to the relative fitness of the chromosomes in the parent generation (probabilistically). There are many methods for selecting the best chromosomes; we use the roulette wheel selection method, in which the parents are selected according to their fitness. Chromosomes with greater fitness will be selected more times. Thus, the better a chromosome’s fitness score, the greater its chances of being selected will be.

D. **Crossover Operator (perform crossover on parents to create the new population)**

The crossover can be applied to either single or double points. Each gene has an equal chance of coming from either parent. Our model uses single-point and fraction crossovers to determine the fraction of the next generation population created by the crossover function.

E. **Mutation Operator (perform mutation of population)**

When a chromosome is chosen for transmission to the next generation, a small number of genes are randomly selected for mutation (with probabilities between 0 and 1). Once the number of genes in the chromosome to be mutated has been determined, these genes are randomly selected and replaced with genes that are not already in the chromosome. In our model, we use uniform mutation.

F. **Determine the Fitness of the New Population.**

G. **Check for Termination Criteria.**

The process is terminated when a stopping criterion, such as a specific number of high chromosomes, a maximum number of generations, or a fitness value of 100\%, is obtained. We
use the average relative change in the best fitness function value over generations is less than or equal certain value or maximum number of generations is reached.

**H. Frequent Masses Analysis**

Frequency with which masses were select is then analyze. Then, using different number of masses form top frequency masses many times to dement best number of masses set which gives best rules accuracy.

**STEP 3: Build an IF-THEN Rule-based Classifier**

The IF-THEN rule-based classifier is built from training data using only the top selected features. Then, the IF-THEN rule-based classifier is used to predict the class label (i.e., healthy or patient) for the MS test data. The IF-THEN classification rule is as follows:

R: IF condition (C), THEN Class (C).

Example.

R1: “If biomarker 1 is less than threshold 1 and biomarker 2 is greater than threshold 2 and biomarker 3 is less than threshold 3, then the sample belongs to the patient group.”

R2 is “If biomarker 1 is greater than threshold 1 and biomarker 2 is less than threshold 2, then the sample belongs to the healthy group.”

- The LHS represent the rule condition; it is a conjunction of feature tests (biomarkers).
- The RHS denotes the rule consequent or the class label (healthy or patient).

In our work, we will build an IF-THEN rule-based classifier from a DT. In comparing the IF-THEN rule-based classifier with the decision tree, we found that the IF-THEN rule-based classifier was easier for humans to understand, especially when the DT was very huge. Then, we will assess each IF-THEN rule using rule coverage and accuracy. The Rule Ordering Scheme (i.e., Rule-Based Ordering) will then be apply. In this scheme, rule priorities are determined beforehand, and a decision list is built. This list is order according to rule quality (accuracy and coverage). The match rule that appears at the beginning of the list has the highest priority. In the event that no rule is satisfied by X, a default rule will be define for a default class, based on the training set. This class then becomes the majority class of the MS sample, encompassing all instances that are not cover by rules.

**5. CASE STUDY AND RESULTS**

In order to test and evaluate the accuracy of oGRC-MS model and to ensure that its rules are understandable, we apply the GRC-MS model to real data using MATLAB® software.

**5.1 Dataset**

We rely on open-source an MS dataset of ovarian cancer that is available to public through the clinical proteomics program of the National Cancer Institute (NCI) website(http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp). This dataset is labeled “Ovarian 8-7-02”. The WCX2 protein chip was used to produce this dataset. To generate the spectrum from the samples, the
upgraded PBSII SELDI-TOF mass spectrometer was used. The dataset includes 162 ovarian cancer patients and 91 control (healthy) patients. The produced spectrum can be represented by a curved shape, in which the x-axis shows the m/z ratio (the ratio of the weight of a molecule to its charge) and the y-axis represents the intensity of the same molecule as a measure of the amount of that molecule. These datasets include peak intensity measurements at 15,154 points, as defined by the corresponding m/z values in the range of 0 to 20,000 Da.

5.2 Experimental Setup and Results

The following steps are apply by the GRC-MS model to the previous dataset:

1) Import MS data (raw data), using the xlsread or importdata function to load the data from an Excel® file. In Excel, the data are represented as discrete values, such that the rows show the m/z ratios and the columns represent the samples. The cells (the intersections of rows and columns) represented each molecule’s intensity as a measure of the amount of that molecule in the sample. After this step is finished, we have two variables loaded into MATLAB (MZ and Y). MZ is the mass/charge vector, while Y is the intensity value for all 216 patients (control and cancer).

2) Preprocess the MS data to remove all forms of noise and all artifacts introduced to the data by applying the following functions in the following order:
   • msbackadj function.
   • mslowess function.
   • mssgolay function.
   • msnorm function.

In addition, a grouping vector is created including the type of each spectrogram and the indexing vector. This “labelling” will aid in any further analysis on this dataset.

3) Run Genetic Algorithm.

   a) Create a Fitness Function for the Genetic Algorithm. In our case, the genetic algorithm tests small subsets of m/z values using the fitness function and then determines which m/z values to pass on to or remove from subsequent generations. The fitness function (biogafit) is passed to the genetic algorithm solver using a function handle. It maximizes the reparability of two classes using a linear combination of a posteriori probabilities and linear classifier error rates.

   Fitness value = a posteriori probability + Error rate of a linear classifier

   b) Set Genetic Algorithm Options. The GA function uses an options structure to store the algorithm parameters used to perform minimizations with the GAs. The gatset function creates this options structure. The parameter values set for the GA are as follows:

   - Population size: [50 100 150 200].
   - Maximum number of generations: [50 100 150 200].
   - Number of features: [1-10].
   - Probability of crossover: [0.5 0.6 0.7 0.8].
   - Probability of mutation: [0.02 0.05 0.1].
   - @selectionroulette.
   - @crossoversinglepoint.
   - @mutationuniform.
c) Run GA to Find the Best Discriminative Features. We using the (ga) to start the GA function to decide the best feature values. We run the GA function with different times for all cases as a filter selection approach with DT. Then, we compute the DT correction rate (accuracy), the DT error rate, the DT sensitivity, and the DT specificity using 10-fold cross-validation. We also compute run time. Then, we compare the results to choice best accuracy trees. Table 3 lists the best GAs result along with their parameters. For example, in the first line we achieve 99.2 accuracy when using 200 population size, 50 generations, 0.7 crossover rate, 0.02 mutation Rate and only uses two features. The best results appears at (Table 3).

<table>
<thead>
<tr>
<th>Population Size</th>
<th>Generations</th>
<th>No. Features</th>
<th>Crossover Rate</th>
<th>Mutation Rate</th>
<th>DT. Correct Rate</th>
<th>DT. Error Rate</th>
<th>DT. Sensitivity</th>
<th>DT. Specificity</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>50</td>
<td>2</td>
<td>0.7</td>
<td>0.02</td>
<td>0.992</td>
<td>0.008</td>
<td>1</td>
<td>0.978</td>
<td>105.412</td>
</tr>
<tr>
<td>150</td>
<td>100</td>
<td>5</td>
<td>0.7</td>
<td>0.05</td>
<td>0.992</td>
<td>0.008</td>
<td>1</td>
<td>0.978</td>
<td>79.906</td>
</tr>
<tr>
<td>150</td>
<td>150</td>
<td>5</td>
<td>0.7</td>
<td>0.05</td>
<td>0.992</td>
<td>0.008</td>
<td>1</td>
<td>0.978</td>
<td>77.689</td>
</tr>
<tr>
<td>150</td>
<td>200</td>
<td>5</td>
<td>0.7</td>
<td>0.05</td>
<td>0.992</td>
<td>0.008</td>
<td>1</td>
<td>0.978</td>
<td>79.003</td>
</tr>
<tr>
<td>150</td>
<td>150</td>
<td>8</td>
<td>0.7</td>
<td>0.02</td>
<td>0.992</td>
<td>0.008</td>
<td>0.993</td>
<td>0.978</td>
<td>90.376</td>
</tr>
<tr>
<td>150</td>
<td>200</td>
<td>8</td>
<td>0.7</td>
<td>0.02</td>
<td>0.992</td>
<td>0.008</td>
<td>0.993</td>
<td>0.978</td>
<td>88.52</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>8</td>
<td>0.7</td>
<td>0.1</td>
<td>0.992</td>
<td>0.008</td>
<td>0.993</td>
<td>0.978</td>
<td>21.814</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>8</td>
<td>0.7</td>
<td>0.1</td>
<td>0.992</td>
<td>0.008</td>
<td>0.993</td>
<td>0.978</td>
<td>25.575</td>
</tr>
<tr>
<td>50</td>
<td>150</td>
<td>8</td>
<td>0.7</td>
<td>0.1</td>
<td>0.992</td>
<td>0.008</td>
<td>0.993</td>
<td>0.978</td>
<td>25.31</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>8</td>
<td>0.7</td>
<td>0.1</td>
<td>0.992</td>
<td>0.008</td>
<td>0.993</td>
<td>0.978</td>
<td>25.143</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>8</td>
<td>0.8</td>
<td>0.1</td>
<td>0.992</td>
<td>0.008</td>
<td>0.993</td>
<td>0.978</td>
<td>63.898</td>
</tr>
<tr>
<td>100</td>
<td>150</td>
<td>8</td>
<td>0.8</td>
<td>0.1</td>
<td>0.992</td>
<td>0.008</td>
<td>0.993</td>
<td>0.978</td>
<td>63.675</td>
</tr>
<tr>
<td>100</td>
<td>200</td>
<td>8</td>
<td>0.8</td>
<td>0.1</td>
<td>0.992</td>
<td>0.008</td>
<td>0.993</td>
<td>0.978</td>
<td>64.1</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>9</td>
<td>0.8</td>
<td>0.02</td>
<td>0.992</td>
<td>0.008</td>
<td>0.993</td>
<td>0.978</td>
<td>36.629</td>
</tr>
<tr>
<td>50</td>
<td>150</td>
<td>9</td>
<td>0.8</td>
<td>0.02</td>
<td>0.992</td>
<td>0.008</td>
<td>0.993</td>
<td>0.978</td>
<td>30.34</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>9</td>
<td>0.8</td>
<td>0.02</td>
<td>0.992</td>
<td>0.008</td>
<td>0.993</td>
<td>0.978</td>
<td>30.443</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>9</td>
<td>0.8</td>
<td>0.1</td>
<td>0.992</td>
<td>0.008</td>
<td>1</td>
<td>0.978</td>
<td>23.469</td>
</tr>
<tr>
<td>150</td>
<td>50</td>
<td>9</td>
<td>0.7</td>
<td>0.05</td>
<td>0.992</td>
<td>0.008</td>
<td>0.993</td>
<td>0.978</td>
<td>70.296</td>
</tr>
</tbody>
</table>

4) Frequent Masses Analysis

Using the parameters in the previous table (Table 3), we obtain 42 different masses that give us the best accuracy results. Figure 6 shows the analysis of the frequencies with which the masses are selected, where mass 8073.585 and 244.3685 appear 10 times giving the best accuracy result.
5) **Build multiple DTs From the Training Dataset Using Only Top MZ Values and Extract Frequent Rules.**

This process involves building multiple DTs using different numbers of masses from the top frequency masses every time. Then, determine the most frequent rule in these trees for each number of masses. Steps below from A to J explain this process for each different number of masses from top two to top ten. For example, using only the top two m/z values to build multiple DTs (x1=8073.5852 m/z and x2=437.0239 m/z), which are the values extracted as top two features from the previous step. Then extract frequent rules from the multiple DTs that built by using the training dataset and compute the average coverage and accuracy of each rule using the test dataset. Note that we apply holdout validation 100 times, randomly reserve two-thirds of the dataset for training to build multiple DTs, and extract most frequent rule. The remaining one-third of the dataset is used for testing, the average coverage and accuracy are computed for the most frequent rules every time.

- **R1:** IF MZ (437.0239) >= 1.22269 THEN Class = Cancer.
- **R2:** IF MZ (437.0239) < 1.22269 and MZ (8073.5852) < 0.29102 THEN Class = Cancer.
- **R3:** IF MZ (437.0239) < 1.22269 and MZ (8073.5852) >= 0.29102 THEN Class = Normal.

Last, build a decision list using accuracy values.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Average Coverage Percentage</th>
<th>Average Accuracy Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>100</td>
<td>95.74</td>
</tr>
<tr>
<td>R2</td>
<td>100</td>
<td>38.35</td>
</tr>
<tr>
<td>R3</td>
<td>100</td>
<td>97.62</td>
</tr>
<tr>
<td>Overall</td>
<td>100</td>
<td>98.80</td>
</tr>
</tbody>
</table>

**Table 4. Rules accuracy values**
6. DISCUSSION

The results show that the GRC-MS classifier model achieves very good results when applied to the analysis of ovarian cancer datasets with different numbers of features or masses that used in the model (Table 6). We observed that 437.0239, 244.36855, 8073.5852 and 793.30944 m/z were significantly discriminative masses that can be potential biomarkers for ovarian cancer. Table 7 lists the frequently occurring masses that play important roles in most of the rules.

Table 6. The GRC-MS classifier results.

<table>
<thead>
<tr>
<th>No. Features</th>
<th>No. Rule</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>98.8095</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>98.8142</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>99.2118</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>99.2447</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>99.5731</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>99.6099</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>99.6112</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>99.6016</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>99.7038</td>
</tr>
</tbody>
</table>

Table 7. Frequently occurring masses that play important roles in most of the rules

<table>
<thead>
<tr>
<th>Mass</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>437.0239</td>
<td>9</td>
</tr>
<tr>
<td>244.36855</td>
<td>5</td>
</tr>
<tr>
<td>8073.5852</td>
<td>4</td>
</tr>
<tr>
<td>793.30944</td>
<td>4</td>
</tr>
<tr>
<td>681.86861</td>
<td>1</td>
</tr>
</tbody>
</table>

In Table 8, shows that our GRC-MS classifier model provides highly competitive accuracy (99.7%) when compared to other existing classifier models, when applied to an ovarian cancer dataset. In addition, our model also provides highly comprehensible rules that facilitate the translation of raw data into easy-to-understand knowledge that can help experts.
### Table 8. Results of some existing classifier models

<table>
<thead>
<tr>
<th>Author(s), Year</th>
<th>Feature Selection Method</th>
<th>Data Mining Algorithm</th>
<th>Result (Accuracy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reynès et al. [35], 2009</td>
<td>GA</td>
<td>DT</td>
<td>98%</td>
</tr>
<tr>
<td>Li et al. [37], 2004</td>
<td>GA</td>
<td>SVM</td>
<td>98%</td>
</tr>
<tr>
<td>Petriconi et al. [38], 2002</td>
<td>GA</td>
<td>Cluster</td>
<td>94%</td>
</tr>
<tr>
<td>Shah and Kusiak [39], 2007</td>
<td>GA</td>
<td>DT</td>
<td>94.07%</td>
</tr>
<tr>
<td>Shah and Kusiak [39], 2007</td>
<td>GA</td>
<td>SVM</td>
<td>97.46%</td>
</tr>
<tr>
<td>Assareh and Moradi [43], 2007</td>
<td>T-test</td>
<td>IF-THEN Rule-Based</td>
<td>86.93%</td>
</tr>
<tr>
<td>Wang and Palade [40], 2011</td>
<td>Fuzzy C-Mean Clustering-based Enhanced Gene Selection Method</td>
<td>IF-THEN Rule-Based</td>
<td>63.75%</td>
</tr>
</tbody>
</table>

## 7. CONCLUSION AND FUTURE WORK

Many studies have sought to increase the accuracy of diagnoses by analyzing MS data and finding biomarkers. Some of these studies have proposed approaches capable of high accuracy, sensitivity, and specificity, while other studies have failed to obtain satisfactory results. One major issue remains: How can an accurate model that avoids the “black box” limitation be built? The “black box” produces such problems as a lack of knowledge flow between the system and the expert. To address this problem and build a model capable of yielding accurate diagnoses that are easy for experts to understand, we used a ruled-based technique to build a classifier model to analyze MS data. Recently, significant attention has been paid to the use of rule-based classification techniques because of their unique ability to provide meaningful outcomes.

In addition, we apply a GA in the feature selection stage to increase the quality and accuracy of the p GRC-MS classifier model. In previous research, excellent results have been obtained through the combination of GA with different types of classifiers. In order to test the validity, accuracy, and performance of the GRC-MS model, we conducted an experimental study using open-source databases. In this experiment, we first applied several preprocessing steps to prepare the MS data for the GRC-MS model. These steps included reducing the noise in the data and the amount of data, identifying and extracting peaks, and normalizing and aligning the data. We found that the GRC-MS classifier model enhance the accuracy and meaningfulness of the MS data analysis results. As a future work, we aim to apply the GRC-MS model to another MS dataset or other high-dimension dataset, such as a microarray gene expression dataset. We also aim to develop more effective fitness functions for the GA.

## REFERENCES


DATA SCIENCE METHODOLOGY FOR CYBERSECURITY PROJECTS

Farhad Foroughi and Peter Luksch

Institute of Computer Science, University of Rostock, Rostock, Germany

ABSTRACT

Cybersecurity solutions are traditionally static and signature-based. The traditional solutions along with the use of analytic models, machine learning and big data could be improved by automatically trigger mitigation or provide relevant awareness to control or limit consequences of threats. This kind of intelligent solutions is covered in the context of Data Science for Cybersecurity. Data Science provides a significant role in cybersecurity by utilising the power of data (and big data), high-performance computing and data mining (and machine learning) to protect users against cybercrimes. For this purpose, a successful data science project requires an effective methodology to cover all issues and provide adequate resources. In this paper, we are introducing popular data science methodologies and will compare them in accordance with cybersecurity challenges. A comparison discussion has also delivered to explain methodologies’ strengths and weaknesses in case of cybersecurity projects.

KEYWORDS

Cybersecurity, Data Science Methodology, Data-Driven Decision-making, User Data Discovery, KDD Process, CRISP-DM, Foundational Methodology for Data Science, Team Data Science Process

1. INTRODUCTION

Cybersecurity solutions are traditionally static and signature-based which means they depend on pattern identification by detecting a match between a pre-captured attack or a malware with a new threat [1]. Hence, it needs regular update to deploy new signatures in the product database. For this reason, it is not possible to detect or prevent zero-day attacks.

The traditional solutions along with the use of analytic models, machine learning and big data could be improved by automatically trigger mitigation or provide relevant awareness to control or limit consequences of threats. Furthermore, traditional solutions are very binary with limited advantages compared to predictive models that could predict the possibility of attacks or risky actions based on data analysis techniques. In addition, access to a large amount of data makes it possible to solve challenging and complicated security problems. In accordance with big data and data mining, the more data creates more accurate and precise analysis [1].

This kind of intelligent solutions is covered in the context of Data Science for Cybersecurity. A general definition of data science is the information extraction and knowledge discovery from data by using a scientific approach. Data science could build innovative cybersecurity solutions by utilising new technological advantages of storage, computing and behavioural analytics [2]. In general, cluster storages which are deployed by distributed systems make it easier to collect and store huge amount of data (big data), and cloud computing also make it possible to utilise
complex and sophisticated machine learning techniques to create predictive and analytic models to identify and detect and respond threats. Behavioural analytics also could transform traditional signature-based detection techniques to the new behaviour-based predictive solutions.

According to data analysis potentials in cybersecurity, National Institute of Standards and Technology has developed a framework consists asset risk identification (and threat consequences), information protection, intruders detection, responding to intruders and business recovery [3].

As a result, Data Science provides significant role in cybersecurity by utilising the power of data (and big data), high-performance computing and data mining (and machine learning) to protect users against cybercrimes. For this purpose, a successful data science project requires an effective methodology to cover all issues and provide adequate resources.

In this paper, we are introducing popular data science methodologies and will compare them in accordance with cybersecurity challenges. Section 2 describes general Data Science overview along with its relation to the cybersecurity concept. Section 3 provides information about popular data science methodologies in details. Four different methodologies have been explained in this section. A comparison discussion is delivered in section 4 to explain methodologies’ strengths and weaknesses against each other along with a summary table. In the end, we recommend a methodology that might cover all requirements to provide the most possible efficient data science cybersecurity project.

2. DATA SCIENCE

Data science could enhance and improve decision-making process by providing data-driven predictions. This requires principles, processes and techniques to understand a problem through an automated evaluation and data analysis. A successful data science has to employ data mining algorithms to solve business problems from a data perspective [4].

Data science is a set of basic concept which leads to the principled extraction of information and knowledge from data. It is very similar to the data mining tries to extract this information via technologies and applied and utilised for relationship management and behaviour analysis in order to recognize patterns, values and user interests [4]. CFJ Wu identified differences between traditional pure statistics and modern data science practices in 1997. He described these significant factors as Data Collection, Data Modelling and Analysis, and Problem Solving and Decision Support [5]. Data science is a recursive process which requires iterative performing.

Fayyad says Data Mining is a component of knowledge discovery in database process [6]. The knowledge discovery in database (KDD) was composed in 1989 to refer to the wide practice of obtaining and acquiring knowledge in data to stress the high-level application of certain data mining techniques. According to Fayyad et al. definition, KDD is the process of utilising data mining techniques to draw out knowledge based on the specification of measures and threshold, making use of a database together with any necessary pre-processing, sampling and transforming the database [7]. Therefore, a knowledge discovery process requires at least, Selection, Pre-processing, Transformation, Data mining, Interpretation and Evaluation steps. Knowledge discovery in database in cybersecurity domain interpreted into two major concepts. These concepts are User Data Discovery (UDD) with is a user profiling process and Data-Driven Decision-making which is a decision-making process based on data analysis [8].
2.1. Data-Driven Decision-making

Data-driven decision-making (DDD) is the term for the process and technique of taking decisions based on data analysis and information evaluation rather than strictly on intuition [4].

DDD is not a binary practice to provide all or nothing. It could be employed in cybersecurity domain with different levels of engagement. Provost et al. demonstrate two types of decisions: 1) decisions which are based on data discovery 2) decisions which are based on frequent decision-making processes particularly at considerable dimensions or massive scale. This kind of decision-making processes might gain from an even minor increase in reliability and precision based on information evaluation and data analysis [4].

Figure 1 describes data science and data-driven decision-making relation. Data Science overlaps data-driven decision-making because cybersecurity decisions and choices could increasingly be created instantly and automatically by computer systems [4].

![Data Drive Decision Making](image)

Figure (1): Data-Drive Decision-making through Data Science

Data processing and data engineering are essential to support data science tasks and very beneficial for data-driven decision-making, effective transaction processing and online pattern recognition. Big Data is simply a term for datasets which are very large for conventional data processing and require new methods and technologies. Therefore, big data technologies are in fact utilised for applying data processing to support data mining strategies and data-driven decision-making tasks [4]. Modern efficient cybersecurity solutions are depend on big data because more data creates more accurate and precise analysis [1].

Data analytic thinking is a crucial element of data science. Underlying the comprehensive collection of methods and strategies for mining information is a significantly smaller set of basic concepts comprising data science. Understanding the essential concepts and having a data analytic thinking structure and framework could help cybersecurity researchers to boost data-driven decision-making process.

2.2. User Data Discovery

User Data Discovery (UDD) is the process of producing profile of users from historical information and past details. This particular information might be personal data, academic documents, geographical details or other private activities.

The primary function of user profiling process is capturing user’s information about interest domain. These information may be used to understand more about individual’s knowledge and skills and to improve user satisfaction or help to make a proper decision. Typically, it evolves data mining techniques and machine learning strategies. UDD process is a type of
knowledge discovery in database or the new version, knowledge data discovery model and requires similar steps to be established [8].

User profiling is usually either knowledge-based or behaviour based. The knowledge-based strategy uses statistical models to categorise a user in the closest model based on dynamic attributes. Typically, Questionnaires and interviews could be utilised to acquire this particular user knowledge [9].

Behaviour-based strategy employs the user’s behaviours and actions as a model to observe beneficial patterns by applying machine learning techniques. These behaviours could be extracted through monitoring and logging tasks [9].

Recognizing user behaviour in real time is an important element of providing appropriate information and help to take suitable action or decision in cybersecurity projects. Typically it is a human task that experts would provide in the information security domain, but it is possible to employ user modelling to make this process automatic by using an application or intelligent agent [10].

UDD could obtain appropriate, adequate and accurate information about a user’s interests and characteristics and demonstrate them with minimal user intervention [11] to offer appropriate awareness with relevant mitigation recommendation based on the security situation.

An intelligent cybersecurity solution should take into account the various attributes and features of a user and a security situation to create a customized solution based on the notion and concept of user profile [12].

3. DATA SCIENCE METHODOLOGY

Several theoretical and empirical researchers have considered basic concepts and principles of knowledge extraction from data. These basic methods and fundamental principles are concluded from numerous data analytic studies [4].

Extracting beneficial knowledge from data should be dealt with systematic processes and procedures through well-defined steps.

Data science needs attentive consideration and result evaluation in the context it is used because the extracted knowledge is significant to assist the decision-making process in a particular application [4].

“Breaking the business problem up into components corresponding to estimating probabilities and computing or estimating values, along with a structure for recombining the components, is broadly useful.” [4].

The correlation finding is one of the data science concepts which should be considered in relation to the cybersecurity. It typically provides details on data items that supply information about other data items, particularly, recognized quantities which reduce the uncertainty of unknown quantities [4].

Entities which are identical with regard to known features or attributes, oftentimes are identical with regard to unknown features or attributes. Computing similarity (pattern recognition) is among the primary resources of data science [4]. It is also significant to pay quite close attention to the existence of confounding elements, most likely unseen ones.
A methodology is a general approach that guides the techniques and activities within a specific domain. The methodology does not rely on certain technologies or tools. Instead, a methodology delivers a framework to acquire results by using a wide range of methods, processes and heuristics [13].

Predictive model creation, pattern recognition and underlying discovery problems through data analysis are usually a standard practice. Data science provides plenty of evolving data analysis technologies to constructing these models. Emerging analytics methods and action automation provide strong machine learning models to solve sophisticated analytic problems such as DDDs. To create an appropriate data analytic model it is required to use a data science methodology that could provide and supply a guiding strategy regardless of technology, data volumes or approaches.

There are several methodologies available for data mining and data science problems such as CRISP-DM and SAS SEMMA and KDD process but Gregory Piatetsky confirms CRISP-DM remains the top methodology for data mining projects with 42% in 2014. The KDD Process has used by 7.5% [14].

Rollins demonstrates a foundational methodology which is similar to CRISP-DM, SEMMA and KDD Process but also emphasizes a number of new methods in data science including big data usage, the incorporation of text analytics into predictive modelling and process automation [13]. Microsoft also introduces Team Data Science Process (TDSP) which recommends a lifecycle for data science projects [15].

Before applying any of these methodologies to cybersecurity projects, it might be helpful to review and compare their essential features. For this reason, this paper provides a comparison between KDD Process, CRISP-DM, TDSP and the foundational methodology for data science (FMDS). FMDS and CRISP-DM have been chosen, because they are considered to be the most popular but SAS SEMMA is not in this review because there is a big decline in applying it (from 13% in 2007 to 8.5% in 2014) [14]. The KDD Process has also included because it provides initial and basic requirements of knowledge discovery. TDSP has been chosen because it is customized for machine learning or artificial intelligence projects which are considerably linked to cybersecurity applications.

3.1. KDD Process

The KDD process offered by Fayyad et al. in 1996 [7]. It is the method of using data mining techniques to extract knowledge based on particular measures and thresholds in a database by employing any necessary pre-processing, sampling or data transformation actions [7]. Furthermore, the application domain perception is needed to be considered during the KDD process development, improvement and enhancement. Figure 2 illustrates the KDD process.

The KDD process has 5 steps as the following [7].

1. Selection: It means generating or producing a target data set or concentrating on a subset of variables or data samples in a database.

2. Pre-processing: This phase tries to obtain consistent data by cleaning or pre-processing selected data.

3. Transformation: Reducing feature dimensionality by using data transformation methods is performing in this phase.
4. Data Mining: Trying to recognize patterns of attention or behaviours by using data mining techniques in a specific form should perform in this step. (Typically, prediction)

5. Interpretation/Evaluation: Mined pattern should be assessed and interpreted in the final phase.

3.2. CRISP-DM

In 1996, SPSS and Teradata developed Cross Industry Standard Process for Data Mining (CRISP-DM) in an effort initially composed with NCR, Daimler and OHRA. It is a cycle of six high-level phases which describe the analytic process [16, 17].

CRISP-DM is still a beneficial tool but details and specifics needed to be updated for cybersecurity projects such as those including Big Data. The original site is not active anymore but IBM SPSS modeller still supports it [14].

Figure 3 demonstrates the CRISP-DM six stages, but their sequence is not strict. CRISP-DM is very well documented and there are many case studies which have used it [16].
CRISP-DM’s structured, well defined and extremely documented process is independent of data mining tools and this important factor is crucial to make the project successful [16].

The CRISP-DM is so logical and seems like common sense. There are many methodologies and advanced analytic platforms which are actually based on CRISP-DM steps because the use of a commonly practised methodology gains quality and efficiency. Vorhies says CRISP-DM provides strong guidance for even the most advanced of today’s data science projects [17].

The six phases are the following [16]:

1. Business understanding: It is designed to focus on understanding the problem or project goals and requirements from a business perspective (here is a cybersecurity application) and then transforming this perception into data mining problem description and preliminary approach.

2. Data Understanding: This phase begins with an initial data collection and then with tasks in order to acquaint with data, to recognise data quality problems, to find out primarily insights into the details or even to identify interesting subsets to develop hypotheses for hidden information. It typically creates a set of raw data.

3. Data Preparation: This phase covers all tasks and activities to build the final required dataset from the first raw data.

4. Modelling: Through this stage, modelling techniques and strategies are selected and applied, and their specific parameters and prerequisites should be identified and calibrated regarding the type of data to optimal values.

5. Evaluation: At this point, the obtained model or models which seem to provide high quality based on loss function completely evaluated and the actions executed to ensure they generalise against hidden data and to be certain it correctly archives the key business goals. The final result is the selection of sufficient model or models.

6. Deployment: This stage means deploying a code representation of the final model or models in order to evaluate or even categorise new data as it arises to generate a mechanism for the use of new data in the formula of the first problem. Even if the goal of the model is to provide knowledge or to understand the data, the knowledge acquired have to be organised, structured and presented in a means which could be used. This includes all the data preparation and required steps which are needed to treat raw data to achieve the final result in the same way as developed during model construction.

3.3. Foundational Methodology for Data Science

This methodology has some similarities and consists many features of KDD Process and CRISP-DM, but in addition, it provides a number of new practices such as use of extremely large data volumes, text and image analytics, deep learning, artificial intelligence and language processing [13]. The FMDS’s ten steps illustrate an iterative nature of the problem-solving process for utilising data to discover security insights. Figure 4 demonstrates FMDS process.

Foundational Methodology for Data Science consists the following steps [18]:

1. Business understanding: This very first phase provides a basic foundation for a profitable and effective resolution of a business problem (a cybersecurity challenge in here) regardless of its size and complexity.
2. Analytic approach: As soon as the problem clearly stated, it is required to determine an analytic approach by identifying suitable machine learning technique to solve it and obtain the desired result.

3. Data requirements: The analytic approach that has been chosen in the second phase, defines needed data requirements including specific data content, formats and representations which are instructed by cybersecurity knowledge.

4. Data collection: In this primary data gathering phase, it is required to identify and collect available data resources (structured, unstructured and semi-structured) which are related and applicable to the problem domain. It is important to cover any data collection gap by revising data requirements and gathering brand new and/or additional data.

   It is also significant to use high-performance platforms and in-database analytic functionality to work with huge datasets for sampling and sub-setting to obtain all available data.

5. Data understanding: Descriptive statistics and visualisation methods are useful in this phase to understand data content, evaluate data quality and explore data insights. This could be required to revise the earlier phase to close data collection gaps.

6. Data preparation: This phase comprises all tasks to construct dataset which will be utilised in the modelling phase. These tasks include data cleaning, data merging from several sources, dealing with missing data, data transformation into more useful variables, duplication elimination, and finding invalid values. In addition, feature engineering and text analytics are possible to be utilised to provide new structured variables, defining and enriching the predictors and boosting or improving the model’s reliability, accuracy and precision. A collaboration of cybersecurity knowledge and existing structured variables could be very useful for feature engineering. This phase is probably the most time-consuming stage, but high-performance and parallel computing systems could reduce the time required and prepare data quickly from huge datasets.

7. Modelling: This phase concentrate on predictive or descriptive model development based on earlier described analytic approach and by using the first version of the prepared dataset as a training set (historical data). The modelling process is extremely iterative as it provides
intermediate insights and reputable refinement of data preparation and model specification. It is significantly helpful to try several algorithms with specific parameters to find the ideal model.

8. Evaluation: Before deployment, it is crucial to evaluate the quality and efficacy of the developed model to realize whether it completely and appropriately addresses the cybersecurity problem. This evaluation involves computing of several different diagnostic measures and other outputs including tables and graphs by using a testing set. This testing set is actually independent of the training set but follows the identical probability distribution and has known results.

9. Deployment: Once the developed model approved and accredited in the evaluation phase that covers the cybersecurity challenge appropriately, it should be deployed into the production environment or perhaps in a comparable test environment. Typically, it will be used in a limited and specific way until all performance variables entirely evaluated. Deployment could be as basic as producing a simple report with proper suggestions or perhaps, embedding the model in an elaborated or sophisticated workflow and scoring process handled by a customised application.

10. Feedback: This final phase, collects outcomes from the implemented edition of the analytic model to analyse and feedback its functionality, performance and efficiency in accordance with the deployment environment.

3.4. Team Data Science Process

The Team Data Science Process (TDSP) is a data science methodology to provide efficient predictive analytics. TDSP These solutions are including artificial intelligence and machine learning. It boots data science project agility, team working and learning by employing best practices and successful structures from Microsoft [15]. TDSP supports both exploratory and ad-hoc projects. Figure 5 illustrates TDSP five stages.

![Figure (5): Team Data Science Process Lifecycle](image)

TDSP supports development of projects which have already employed CRISP-DM and KDD process. It is very customizable based on project’s size and dimensions [19].

The TDSP lifecycle consists integrative phases as the following [19].
1. Business Understanding: Initially, a question which describes the problem objectives should be defined clearly and explicitly. Relevant predictive model and required data source/s also have to be identified in this step.

2. Data Acquisition and Understanding: Data collection starts in this phase by transferring data into the target location to be utilised by analytic operations. The raw data is needed to be cleaned and incomplete to incorrect values should be identified. Data summarization and visualization might help to find required cleaning procedures. Data visualization also could help to measure if data features and collected amount of data are adequate over the time period. At the end of this stage, it might be necessary to go back to the first step for more data collection.

3. Modelling: Feature engineering and model training are two elements of this phase. Feature engineering provides attributes and data features which are required for machine learning algorithm. Algorithm selection, model creation and predictive model evaluation are also sub-components of this step. Collected data should be divided into training and testing datasets to train and evaluate machine learning model. It is important to employ different algorithms and parameters to find the best suitable solution to support the problem.

4. Deployment: Predictive model and data pipeline are needed to be produced in this step. It could be a real-time or a batch analysis model depends on the required application. The final data product should be accredited by the customer.

5. Customer Acceptance: The final phase is customer acceptance which should be performed by confirming data pipeline, predictive model and product deployment.

4. DISCUSSION

All data science methodologies where discussed consist four common iterative stages including problem definition/ formulation, data gathering, data modelling and data product development except the KDD process.

![Figure (6): General Data Science Methodology](image)

According to the above explanations, table 1 demonstrates a summary of the presented correspondences.

Comparing the KDD process with the CRISP-DM presents that KDD process does not cover the business understanding and also deployment phases in CRISP-DM methodology.

As it is mentioned above, the business understanding phase provides a comprehension perception of the application domain and pertinent prior knowledge and also objectives of required solution. Furthermore, the deployment phase incorporates knowledge or modelling code into a system or...
application to build a data product. These two significant phases have been missed in KDD process.

Table (1): Summary of Data Science methodologies and their phases

<table>
<thead>
<tr>
<th>KDD Process</th>
<th>CRISP-DM</th>
<th>FMDS</th>
<th>TDSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>Business Understanding</td>
<td>Business Understanding</td>
<td>Business Understanding</td>
</tr>
<tr>
<td>-</td>
<td>Analytic Approach</td>
<td>Data Requirements</td>
<td></td>
</tr>
<tr>
<td>Selection</td>
<td>Data Understanding</td>
<td>Data Collection</td>
<td>Data Acquisition and Understanding</td>
</tr>
<tr>
<td>Pre Processing</td>
<td>Data Understanding</td>
<td>Data Understanding</td>
<td></td>
</tr>
<tr>
<td>Transformation</td>
<td>Data Preparation</td>
<td>Data Preparation</td>
<td></td>
</tr>
<tr>
<td>Data Mining</td>
<td>Modelling</td>
<td>Modelling</td>
<td></td>
</tr>
<tr>
<td>Interpretation/ Evaluation</td>
<td>Evaluation</td>
<td>Evaluation</td>
<td>Modelling</td>
</tr>
<tr>
<td>-</td>
<td>Deployment</td>
<td>Deployment</td>
<td>Deployment</td>
</tr>
<tr>
<td>-</td>
<td>Feedback</td>
<td>Customer Acceptance</td>
<td></td>
</tr>
</tbody>
</table>

Comparing the CRISP-DM with the FMDS illustrates that CRISP-DM has not the analytic approach and the feedback phases. The analytic approach is required to recognize appropriate statistical or machine learning techniques before entering to the data gathering steps. This phase could be very useful to identify the suitable data collection strategy and data resources but it is missed in the CRISP-DM methodology. The feedback phase also has been missed that is very beneficial to optimize the system to achieve high-performance functionality and efficient result.

A comparison between FMDS and TDSP presents that they are very similar but FMDS has more details steps in general. Feedback also is a part of FMDS cycle which could create new requirements to improve the data product in an iterative process but customer acceptance is a stage out of the circle in the TDSP. FMDS detailed stages could be more useful for a wide range of projects but TDSP uses a specific set of Microsoft tools and infrastructure to deliver intelligent applications by deploying machine learning or AI models.

Concerning the remaining phases present the following:

- The Data understanding phase is in both CRSIP-DM and FMDS can be recognized as the collaboration of Selection (Collection) and Pre-Processing phases in the KDD process but Data Acquisition and Understanding in TDSP also covers Transformation stage in the KDD process. However, data requirements which are related to the analytic approach phase and provides required data content are missed in both KDD process and CRISP-DM. Selecting an analytic approach is integrated into the business understanding phase of TDSP but cybersecurity projects might gain more benefits in details from this task when it is an independent step in the FMDS particularly when data resources are separated and requires a different method or level of access.

- Business understanding and problem formulation is an initial phase that makes the data understanding phase more efficient by recommending data formats and representations but data requirement analysis is missed in the CRISP-DM. It might be very crucial in cybersecurity projects particularly when data resources are unstructured or semi-structured.

- The data preparation phase has the similar function as the transformation phase in KDD process and it is included in the Data Acquisition and Understanding phase of TDSP by using some tools.
• The modelling phase might be recognised with data mining phase which is very limited in the KDD process. Modelling phase in TDSP also has the evaluation task included by using some tools such as scoring and performance monitoring, but it is an independent phase in other methodologies.

• The evaluation phase is also included in all three methodologies.

In spite of CRISP-DM strengths, there is more emphasis that should be considered for modern cybersecurity projects.

• It is required that methodology could also handle data blending from several sources. This should be deployed through a completely repeatable process. FMDS and TDSP provide this feature in the Data Preparation and Data Acquisition and Understanding phases. TDSP provides Microsoft tools to support On-Premises, Cloud, Databases and Files but FDMS is independent of any platform or a specific tool. This might makes cybersecurity projects more reliable and efficient.

• Choosing the most appropriate degree of reliability and accuracy for the problem is very crucial to make sure there is no need to spend excessive time on data preparation or modelling to enhance accuracy when it could not be utilised. This feature also included in the FMDS methodology.

• The entire analytic algorithm should be tested and evaluated to make sure there are working in all situations and not just for sample modeller. The evaluation, deployment and feedback cycle in the FMDS could provide this requirement as well as Model training task of Modelling phase in the TDSP. Feedback phase in FMDS might create new data science questions to optimize the cybersecurity product or make new functionalities for it.

• It is significant to consider quality during model simplification by ensuring that decision elements such as missing data reduction, synthetic features generation and unseen data holdout are properly managed. The evaluation, deployment and feedback cycle in the FMDS could bring this need better than simple quality insurance in the evaluation phase of CRISP-DM.

• Data science lifecycle is very well defined in the FMDS and connections are clearly determined between every stage but TDSP’s stages are all linked together (except customer acceptance) and it is possible to move into any stages from anyone else.

• TDSP lifecycle is designed for predictive analytic projects by using machine learning or artificial intelligence models. FMDS is more general and independent of any platform, tool, model or algorithm. Both are functional for exploratory or ad-hoc cybersecurity analytics projects by some customization.

5. CONCLUSION

In conclusion, a cybersecurity data science project requires four general steps. The first step has to be a problem definition by formulating a security challenge. In accordance with problem definition and appropriate formula, it is necessary to gather required information in the second step. The collected information should be employed in the third step and in an analysis process to provide adequate data which is expected to predict or provide a resolution for the defined
problem. The final step is a production step which deploys relevant modules and a system to run
the whole process automatically and regularly when it is needed.

![Diagram: FMDS for Cybersecurity projects](image)

Figure (7): FMDS for Cybersecurity projects

Regarding the general process and in accordance to Table 1, the FMDS covers all beneficial
attributes of CRISP-DM methodology but fills data gathering gaps and also provides extra steps
to optimize and enhance the model and results by mathematical prescriptive analytics and using
high-performance computing. It is also platform and tool independent but TDSP is not. Because it
is designed in details with clearly divided steps, it could be fully customized to fit in any
cybersecurity project.

REFERENCES

https://www.information-management.com/opinion/the-conflict-between-data-science-and-
cybersecurity


Technology, 2014.


1996.


Understanding People Title Properties To Improve Information Extraction Process

Saleem Abuleil and Khalid Alsamara

MMIS Department, Chicago State University, Chicago, USA

Abstract

In this paper, we introduce a new approach to tackle the process of extracting information about people mentioned in the Arabic text. When a person name is mentioned in the Arabic text usually it is combined with a title, in this paper the focus is on the properties of those titles. We have identified six properties for each title with respect to gender, type, class, status, format, and entity existence. We have studied each property, identified all attributes and values that belong to each one of them and classified them accordingly. Sometimes person title is attached to an entity; we have also identified some properties for these entities and we show how they work in a harmony with person title properties. We use graphs for the implementation, nodes to represent person title, person name, entity and their properties, where edges are used to present inherited properties from parent nodes to child nodes.

Keywords

People Titles, Title Properties, NLP, Arabic Language

1. Introduction

Information Extraction (IE), as defined in the Message Understanding Conferences, has been traditionally defined as the extraction of information from a text in the form of text strings and processed text strings that are placed into slots labeled to indicate the kind of information that can fill them. The problem of extracting information from a large document collection can be approached using many different algorithms. The three classic models used in information extraction, (under which all these algorithms can be loosely grouped), are called Rule-based, Pattern Learning, and Supervised Learning. Most of the Arabic Named Entity Recognition (NER) systems use keywords such as titles to tag proper name phrases in the text, once they tag proper name phrase they use either rule-based systems or statistical approach to tag proper names and extract information about them. Using titles to tag proper names in the Arabic text is an important technique that has been used widely, but titles have been used as keywords for the purpose of identifying proper name phrases and tag proper names without studying and exploring their properties. Our technique in this paper is to identify and use title properties and attributes to enhance the result of extracting information about people names in the Arabic text.
2. LECTURER

Al-Kouz [1] presented a framework designed for mining the explicit and implicit lexical semantic information impeded in the structure and the content of Aljazeera.net. Furthermore, it provides an efficient and structured access to the resulted semantic graph, the authors also claim in their paper that Aljazeera.net is professionally edited and has rich semantic structure and it establishes an asset, an impediment and a challenge for research in Arabic Natural Language Processing. Abdallah [2] proposed a simple method for integrating machine learning with rule-based systems and implement this proposal using the state-of-the-art rule-based system for NERA, the experimental evaluation shows that their integrated approach increases the F-measure by 8 to 14% when compared to the original (pure) rule-based system and the (pure) machine learning approach, and the improvement is statistically significant for different datasets, more importantly, their system outperforms the state-of-the-art machine-learning system in NERA over a benchmark dataset. Abdul Hamid [3] introduced simplified yet effective features that can robustly identify named entities in Arabic text without the need for morphological or syntactic analysis or gazetteers, a CRF sequence labeling model is trained on features that primarily use character n-gram of leading and trailing letters in words and word n-grams, the proposed features help overcome some of the morphological and orthographic complexities of Arabic.

Abuleil [4] presented a new technique to extract names from the text by building a database and graphs to represent the words that might form a name and the relationships between them. First, they mark the phrases that might include names, second they build graphs to represent the words in these phrases and the relationships between them, and third, they apply rules to find the names. Benajiba [6] investigated the impact of using different sets of features in three discriminative machine learning frameworks, namely, support vector machines, maximum entropy and conditional random fields for the task of named entity recognition, they explore lexical, contextual and morphological features and nine data-sets of different genres and annotations; they measure the impact of the different features in isolation and incrementally combine them in order to evaluate the robustness to noise of each approach.

Chen [8] described their system for the CoNLL-2012 shared the task, which seeks to model coreference in Onto Notes for English, Chinese, and Arabic; they adopt a hybrid approach to coreference resolution, which combines the strengths of rule-based methods and learning-based methods, they official combined score over all three languages is 56.35. In particular, their score on the Chinese test set is the best among the participating teams. Habash [9] made an argument that is the many differences between Dialectal Arabic and Modern Standard Arabic (MSA) pose a challenge to the majority of Arabic natural language processing tools, which are designed for MSA, so in their paper retarget an existing state-of-the-art MSA morphological tagger to Egyptian Arabic (ARZ), their evaluation demonstrates that their ARZ morphology tagger outperforms its MSA variant on ARZ input in terms of accuracy in part-of-speech tagging, diacritization, lemmatization and tokenization; and in terms of utility for ARZ-to English statistical machine translation. Pasha [10] presented MADAMIRA, a system for morphological analysis and disambiguation of Arabic that combines some of the best aspects of two previously commonly used systems for Arabic processing, MADA (Habash and Rambow, 2005; Habash et al., 2009; Habash et al., 2013) and AMIRA (Diab et al., 2007). MADAMIRA improves upon the two systems with a more streamlined Java implementation that is more robust, portable, extensible and is faster than its ancestors by more than an order of magnitude.
3. **People Title Properties**

When a person name is mentioned in the Arabic text usually it is attached to a title. We have studied these titles and identified different properties for each one of them; we also identified some attributes and values for each property. We have identified six properties for each title with respect to gender, type, class, status, format, and existence of an entity. In this section, we explain each one of them and we show some examples in table 1.

**Gender**: in the Arabic language, there are two values for this property masculine and feminine. In Arabic language to form a feminine title from the masculine, you simply add “taa’ marbuta” which looks like (ة) to the end of the title, for example وزیر (he) Minister is a masculine and to form the feminine from it we add “taa’ marbuta” “ة” to the end of the title وزيرة (she) Minister.

**Type**: We have classified title into three Types:

- Occupational title that indicates a position or job of the person like مدير جمیر Minister and Consultant مستشار and Manager مستشار وزیر and Consultant
- A social title like Mr. السيد Mr, Ms. السيدة and Mrs.
- Professional title that refers to a certain profession like طبيب مهندس طبيب, physician, engineer and teacher

A person might hold two titles at the same time such as (Mr. and consultant), and (engineer and manager), etc. some titles could be used for two classes like Sheikh, could be used for social or occupational.

**Class**: based on job field that people they hold we have classified titles into different classes: politics, religion, education, sport, media, industry, military, etc. some titles could be used for two types like Sheikh, could be used for religion or politics, some titles like president and manager could be classified into different classes politics, education, sports, industry, etc. to identify the class for these cases we use the entities as we discuss later. Some titles do not have a class such as Mr. and Mrs.

**Status**: person title could be simple or compound, simple title has one word such as السيد Mr. وزیر Minister, compound title has two words like Official Speaker ولي العهد and Official Speaker

**Format**: there are two formats of the title either defined or indefinite. Arabic word starts with ال (the) to define it. Al-وزیر (ال) is the definite article in the Arabic language. For example, the word وزیر wazer "Minister" can be made definite by prefixing it with al-, resulting in الوزیر al-wazer "the Minister". Consequently, al- is typically translated as The in English. A defined title that starts with and the word followed is not a verb, adjective, nationality, or particle then the noun is most likely is a person name.

**Entity Existence**: Sometimes an entity comes between person title and person name, entity existence property has two values either Yes or No. Most likely if a person title is defined no entity to follow. More details about entities are discussed in the next section.
4. **Entities and Their Properties**

Sometimes entity or nationality comes between person title and person name. We have studied hundreds of cases and based on our study we identified four properties for each one of them with respect to type, class, gender, and status. See Table 2. In this paper we handle the following three most common scenarios:

- **Person Title + Entity (both title and name are mentioned) + Person Name**

  Example: رئيس الدولة تري كارداغان President of *Turkey State* Erdoğan

  In this example entity title (دولة State) and entity name (Erdoğan) are mentioned.

- **Person Title + Entity (title is omitted and name is mentioned) + Person Name**

  Example: مراسل الجزيرة إلياس كرام Report of *Aljazeera* Alysa Karram

  In this example entity title (خبارية قناة News Channel) is omitted and entity name (الجزيرة Aljazeera) is mentioned.

- **Person Title + Entity (title is mentioned and name/nationality is omitted) + Person Name**

  Example: الناطق الرسمي باسم الحكومة الدكتور محمد المومني Official speaker of the *government* Dr. Muhammad Almumani

  In this example entity title (الحكومة government) is mentioned and nationality (الأردنية Jordanian) is omitted.

In some cases it is difficult to identify the class property of person title like President and Manager where each one of them can be classified into different classes such as politic, sport, industry, but when an entity is maintained, and by using the value of the class of that entity, it helps to identify the class of the person title, for example when a company name is attached to person title President, the title is classified as industry category and when university name is attached to the same person title President president, the title this time is classified as education category.

In this paper, we identify nationality as an adjective and we identify four properties for it with respect to gender, type, format and country. Gender could be either masculine or feminine, the type has the value nationality, and the country has the value which country the nationality belongs to. When nationality comes directly after the رئيس president, we add a class property with a value politics.
### Table 1. Title Properties

<table>
<thead>
<tr>
<th>Gender</th>
<th>Format</th>
<th>Type</th>
<th>Class</th>
<th>Entity Existence</th>
</tr>
</thead>
<tbody>
<tr>
<td>ماصن</td>
<td>Indefinite</td>
<td>Occupational</td>
<td>Politics</td>
<td>Yes</td>
</tr>
<tr>
<td>The Minister</td>
<td>Defined</td>
<td>Social</td>
<td>Politics</td>
<td>No</td>
</tr>
<tr>
<td>لاعب</td>
<td>Indefinite</td>
<td>Occupational</td>
<td>Sport</td>
<td>Yes</td>
</tr>
<tr>
<td>The Player</td>
<td>Defined</td>
<td>Occupational</td>
<td>Sport</td>
<td>No</td>
</tr>
<tr>
<td>ولی</td>
<td>Indefinite</td>
<td>Occupational</td>
<td>Politics</td>
<td>No</td>
</tr>
<tr>
<td>Crown Prince</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>الناطق الرسمي</td>
<td>Defined</td>
<td>Occupational</td>
<td>Follows entityclass</td>
<td>Yes</td>
</tr>
<tr>
<td>Official Speaker</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>مدير</td>
<td>Indefinite</td>
<td>Occupationally</td>
<td>Follows entityclass</td>
<td>Yes</td>
</tr>
<tr>
<td>Manager</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>المدير</td>
<td>Defined</td>
<td>Social</td>
<td>Follows entityclass</td>
<td>No</td>
</tr>
<tr>
<td>The Manager</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>المحامي</td>
<td>Defined</td>
<td>Professional</td>
<td>Law</td>
<td>No</td>
</tr>
<tr>
<td>The Attorney</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>الشيخ</td>
<td>Defined</td>
<td>Social</td>
<td>Politic or Religion</td>
<td>No</td>
</tr>
<tr>
<td>The Sheikh</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>السيد</td>
<td>Defined</td>
<td>Social</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>Mr.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>المدرس</td>
<td>Indefinite</td>
<td>Professional</td>
<td>Education</td>
<td>No</td>
</tr>
<tr>
<td>The Teacher</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>مراصل</td>
<td>Indefinite</td>
<td>Media</td>
<td>Media</td>
<td>Yes</td>
</tr>
<tr>
<td>Reporter</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2. Entity Properties

<table>
<thead>
<tr>
<th>Title</th>
<th>Gender</th>
<th>Class</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>دولة</td>
<td>Feminine</td>
<td>Politics</td>
<td>Location</td>
</tr>
<tr>
<td>حكومة</td>
<td>Feminine</td>
<td>Politics</td>
<td>Location</td>
</tr>
<tr>
<td>مملكة</td>
<td>Feminine</td>
<td>Politics</td>
<td>Location</td>
</tr>
<tr>
<td>وزارة</td>
<td>Feminine</td>
<td>Politics</td>
<td>Organization</td>
</tr>
<tr>
<td>جمعية</td>
<td>Feminine</td>
<td>Social</td>
<td>Organization</td>
</tr>
<tr>
<td>نادي club</td>
<td>Masculine</td>
<td>Social / Sport</td>
<td>Organization</td>
</tr>
<tr>
<td>جامعة</td>
<td>Feminine</td>
<td>Education</td>
<td>Organization</td>
</tr>
<tr>
<td>كلية college</td>
<td>Feminine</td>
<td>Education</td>
<td>Organization</td>
</tr>
<tr>
<td>مدرسة school</td>
<td>Feminine</td>
<td>Education</td>
<td>Organization</td>
</tr>
<tr>
<td>ملعب stadium</td>
<td>Masculine</td>
<td>Sport</td>
<td>Location</td>
</tr>
<tr>
<td>قناة اخبارية News Channel</td>
<td>Feminine</td>
<td>Media</td>
<td>Organization</td>
</tr>
<tr>
<td>كنيسة church</td>
<td>Feminine</td>
<td>Religion</td>
<td>Organization</td>
</tr>
</tbody>
</table>
5. ANALYSIS

We use graphs to implement the concepts in this paper; we use nodes to represent person title, person name, nationality, entity and their properties, and we use edges to present inherited properties from parent node to child node, see fig 1. We tag person name phrases in the text, each phrase starts with a title and terminated with a person name, next we tag the elements in the tagged phrases with respect to titles, entities, and nationalities and then apply the concepts of this paper to identify the properties of each one of them. Properties are inherited from one node to another (parent to child) and once a child gets inherited property from the parent they also forward this information to next node in the graph. When reach the last node in the graph we process all information from all nodes and produce the results. Harmony of inherited information between nodes is also validated with respect to the properties of the titles, entities, and nationalities. There is should be a match between two titles belong to the same person, same gender and format values between adjective (nationality) and the element presided to it either a title or an entity.

In figure 2 we illustrate different scenarios about the same person and we show how the person inherits all properties from all parent nodes, the person name is Albright, Albright is a female, she is a politician, and her occupation is the minister of the USA foreign ministry. Example 1 contains two titles and one adjective (nationality), the first title and the adjective both have the same values of format property and gender property, both titles mentioned for the same person and they have the same value of gender property. Example 3 contains one entity which is foreign and one adjective American, since both of them have the same value of format property “defined”, the title وزيرة minister has a different format property value “indefinite”, then the nationality refers to the organization and not to the person title, the entity وزارة ministry is omitted, both titles and the nationality of this example belong to the same person they have the same value of gender property.

<table>
<thead>
<tr>
<th>Example 1</th>
<th>وزيرة The Minister</th>
<th>Title</th>
<th>Defined</th>
<th>Feminine*</th>
</tr>
</thead>
<tbody>
<tr>
<td>الأเมريكية American</td>
<td>Adjective</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>السيدة The Mrs.</td>
<td>Title</td>
<td></td>
<td>Feminine*</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 3</th>
<th>وزيرة Minister</th>
<th>Title</th>
<th>Indefinite</th>
<th>Feminine*</th>
</tr>
</thead>
<tbody>
<tr>
<td>الخارجية Foreign</td>
<td>Entity</td>
<td>Defined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>الأเมريكية American</td>
<td>Adjective</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>السيدة The Mrs.</td>
<td>Title</td>
<td></td>
<td>Feminine*</td>
<td></td>
</tr>
</tbody>
</table>

In figure 3 we show different scenarios where the value of the class property of person title is uncertain and we are going to use the value of the class property of the entity to identify it. In example 1, the value of the class property of the entity is education, and then we use it for the value of the class property of the title “وزيرة”. Example 2, the value of the class property of the entity is politics, and then the value of the class property of the title “وزيرة” is also politics, the second title is Sheik “شيخ” and could have two possibilities either politics or religion but since the first title has the class property value is politics we select the value politics as well for the second
title. In example 3, the value of the class property of the entity is industry, and then the value of the class property of the title “رئيس” is also industry.

6. CONCLUSION

In this paper we have studied person titles, entities and nationalities attached to them, identify the properties for each one of them and used this information to extract information about people mentioned in the Arabic text we also validated the inherited property values between the nodes in the graph. Our source of data is Ajazeera.net, further analysis to be done in the future.
Fig 2. Illustration A
REFERENCES


STRATEGY OF THE REMOVE AND EASY TBT IN GCC 6 COUNTRIES

Yong-Jae Kim

Department of Business Administration, Korea Polytechnic University, Prof. ph.D

ABSTRACT

The last technical barriers to trade(TBT) between countries are Non-Tariff Barriers(NTBs), meaning all trade barriers are possible other than Tariff Barriers. And the most typical examples are (TBT), which refer to measure Technical Regulation, Standards, Procedure for Conformity Assessment, Test & Certification etc. Therefore, in order to eliminate TBT, WTO has made all membership countries automatically enter into an agreement on TBT. In this study, the elimination strategy of TBT with aid of technical regulations or standards is excluded, and only the conformity assessment shall be considered as the strategic measure of eliminating TBT in GCC(Gulf Cooperation Council) 6 countries. The measure for every membership country to accord with the international standards corresponding to their technical regulations and standards, is only to present TBT related Specific Trade Concern(STC) to WTO. However, each of countries retains its own conformity assessment area, and measures to settle such differences are various as well. Therefore, it is likely required an appropriate level of harmonization in them to carry forward this scheme. KTC(Korea Testing Certification) written MRA with GCC test & certification company in 2015 years. So Korea exporting company can export to GCC goods with attached test & certification documents in Korea. To conclude, it is suggest MRA for the remove and reduce TBT to increase export and import among countries.

KEYWORDS

FTA, Standards, Conformity Assessment, TBT(Technical barrier to trade), MRA, WTO

1. INTRODUCTION

This paper purpose make to remove and to easy TBT of industrial products such as IT, S/W, IOT, BigData, Home network. Research methodology is review 2nd data analysis and focus group Interview Government officer, Professor and CEO.

This paper compare & analyze International rule & system as follow.

First, It is to compare & analyze the standard, technical regulation, Test & certification procedure and Inspection. Second, it is review electric/electronic Test, certification and calibration. Third, it is analyze MRA between Korea and GCC 6countries, SDoC, Mutual Acceptance of International certification such as ILAC(APLAC) and IEC CB scheme.

This paper intends to draw conclusion and make implication as follows.

First, we must promote FTA and MRA. Second, we make to remove and to easy TBT by MRA between Korea and GCC 6 countries. Although the MRA is a system where all parties that have concluded agreement enjoy the advantage, Korea is under a state of concluding only the stage 1 agreements (exchange test documents) with GCC 6 countries. Also, we must conclude MRA stage 2 agreements (exchange certification documents) with GCC 6 countries.

2. PREVIOUS STUDIES

2.1 STRATEGY TO REMOVE & EASE TBT IN OECD

SDoC has strengths in cost reduction, time saving and product information protection aspects compared to the certification system while having vulnerability in terms of product safety issue, etc. Therefore, an effective post market surveillance of the regulation authorities must be supported to be operated effectively. WTO’s TBT Committee has suggested that the SDoC is more effective TBT elimination method than the MRA (OECD, 2000).

2.2 STRATEGY TO REMOVE & EASE TBT IN APEC TEL MRA

Testing and certification are expensive procedures for exporters, importers and regulators that increase the cost to users and delays the availability of products in a large number of markets.

All stakeholders benefit from simplified procedures that can reduce these costs. At the same time, regulators need to have confidence in the quality of testing that provides the basis for certification of equipment.

In June 1998, the APEC1 Telecommunications and Information Ministers agreed to streamline APEC-wide processes for the testing and type-approval of telecommunications equipment.

This landmark arrangement, the Mutual Recognition Arrangement for Conformity Assessment of Telecommunications Equipment (APEC TEL MRA2), was the first multilateral agreement of its type in the world.

This Arrangement streamlines the Conformity Assessment Procedures for a wide range of telecommunications and telecommunications-related equipment and facilitates trade among the APEC member economies.

It reduces a significant barrier to what is projected to be a US$60 billion industry by 2010. Its scope includes all equipment subject to telecommunication regulations, including wire line and wireless, terrestrial and satellite equipment. For such equipment, the MRA covers electromagnetic compatibility (EMC), specific absorption rate (SAR) and electrical safety aspects as well as purely telecommunications aspects of the conformity assessment requirements.

3. STRATEGY TO REMOVE/EASE OF THE TBT

TBT is an abbreviation for ‘Technical Barriers to Trade’ while this stands for the various obstacles in terms of trade that hinder the free movement of goods and services as the trading
partner countries adopt and apply different Technical Regulations, Standards, Test & Certification Procedures and Inspection Systems, etc.

3.1 STRATEGY TO REMOVE/EASE OF THE TBT

TBT makes the countries to harmonize technical regulations, standards or conformity assessments with the international standards and does not occur in case of being transparent. However, the fact is that TBT occurs if a specific country does not comply with the principles above during legislation and amendment of the laws related to technical regulations, standards or conformity assessments while STC must be submitted to settle this TBT. In the conformity assessment of ICT section, various methods of solution exist on TBT depending on the issue other than filing a lawsuit to WTO if a specific country operates the conformity assessment section differently from TBT.

(1) Request for Introduction of SDoC System

SDoC system stands for the one to guarantee market autonomy and raise efficiency of restriction as a system for the supplier to guarantee by evaluating whether its own product is appropriate for the concerned standard by escaping from the compulsory certification system which requires certification in relation to the product manufacture. Since SDoC(Supplier's Declaration of Conformity) is a follow-up and legal system, it is the method of releasing new products under the manufacturer’s own responsibility to become responsible for various problems to follow.

(2) Strategy of MRA

The manufacturers of industrial products are able to export only after acquiring a compulsory standard certification mark. While MRA is concluded in order to save cost and time required for this, only the test report implemented at the exporting country is recognized if MRA stage 1(exchange test documents) is concluded while both the test report and the certification market may be implemented at the exporting country may be implemented if MRA stage 2(exchange certification documents) is concluded.

If both countries conclude the MRA such as FTA, it would be opening the homeland market to the manufacturer of partner country since it is customs-free.

3.2 DOMESTIC ELECTRIC & ELECTRONIC CERTIFICATION SYSTEM AND RELATED LAWS

- Conformity Assessment System of Korea

Supplier's Declaration of Conformity (SDoC) the one to guarantee market autonomy and raise efficiency of restriction as a system for the supplier to guarantee by evaluating whether its own product is appropriate for the concerned standard by escaping from the conventional compulsory certification system which requires certification in relation to the product manufacture.
(1) Acceptance of Internationally Certified Test Report

In addition to the method of concluding an MRA, various methods to recognize the test reports estimated at the partner country or a third country exist. Among them, the most widely used method is the one to accept test reports of the testing agencies that have been recognized by ILAC(APLAC) and CB Scheme. Test & Certification Based Infrastructure Setup Support

3.3 COMPARATIVE ANALYSIS AND STRATEGY OF REMOVE AND EASE TBT

The systems mentioned above have different characteristics from each other. If the comparative analysis is performed from the perspectives of scope of effect, intensity of effect and usage status in Korea, they can be summarized as follows.

[Figure 1] Comparative Analysis and strategy of the remove and ease TBT

<table>
<thead>
<tr>
<th>Scope of Effect</th>
<th>MRA</th>
<th>SDoC</th>
<th>Mutual Acceptance of International Certification</th>
<th>Infrastructure Setup Support Project</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Partner country of agreement</td>
<td>All countries</td>
<td>Participating countries</td>
<td>Participating countries</td>
</tr>
<tr>
<td>Intensity of Effect</td>
<td>In stages</td>
<td>Limited to the products that have applied the system</td>
<td>By accepted field</td>
<td>By accepted field</td>
</tr>
<tr>
<td>Current Status of Usage In Korea</td>
<td>Completed stage 1 conclusion with 5 countries and negotiating with a number of countries</td>
<td>Applied to the products with low level of harm</td>
<td>Field of private sector standards</td>
<td>EMC field is not used</td>
</tr>
<tr>
<td>Remarks</td>
<td>Need to promote upper stage with more countries</td>
<td>Necessity to extend applied products is low in a short run</td>
<td>Handle flexibly depending on the acceptance situation of foreign countries</td>
<td>Handle flexibly depending on the acceptance situation of foreign countries</td>
</tr>
</tbody>
</table>

If both countries conclude the MRA such as FTA, it would be opening the homeland market to the manufacturer of partner country since it is customs-free.
Among APEC members, time and cost required for preparing the copy of agreement can be saved if the MRA Guide prepared by this organization is used. Although Korea has concluded MRA stage 1 with the United States, Canada, Chile and Vietnam, etc., the effect is clearly shown only in the MRA with the United States.

4. CONCLUSION

This study intends to draw conclusions and make policy implications as follows.

First, we must promote a multi-track simultaneous agreements with the countries that have necessity of short-term promotion.

Second, the countries with necessity of short-term promotion on the preferential basis are China, Japan and USA, etc.

Third, it is necessary to conclude MRA agreement with the leading countries among the GCC 6 countries on the preferential basis. It is necessary to prepare negotiation on the preferential basis with GCC 6 countries.

Fourth, support on the countries that have not fully prepared the conformity assessment system needs to be gradually extended. However, the method of support on these countries also must vary depending on the country. KTC (Korea Testing Certification) written MRA GCC test & certification company in 2015 years. So Korea exporting company can export to GCC goods with attached test & certification documents in Korea. To conclude, it is suggest MRA for the remove and reduce TBT to increase export and import among countries.

REFERENCES


AUTHORS

ISO TC 68 Member
ISO TC 195 Member
ISO/IEC SC32 WG1 Vice convenor
PERFORMANCE ANALYSIS OF SYMMETRIC KEY CIPHERS IN LINEAR AND GRID BASED SENSOR NETWORKS

Kaushal Shah and Devesh C. Jinwala

Department of Computer Engineering,
S.V.National Institute of Technology, Surat, India.

ABSTRACT

The linear and grid based Wireless Sensor Networks (WSN) are formed by applications where objects being monitored are either placed in linear or grid based form. E.g. monitoring oil, water or gas pipelines; perimeter surveillance; monitoring traffic level of city streets, goods warehouse monitoring. The security of data is a critical issue for all such applications and as the devices used for the monitoring purpose have several resource constraints (bandwidth, storage capacity, battery life); it is significant to have a lightweight security solution. Therefore, we consider symmetric key based solutions proposed in the literature as asymmetric based solutions require more computation, energy and storage of keys. We analyse the symmetric ciphers with respect to the performance parameters: RAM, ROM consumption and number of CPU cycles. We perform this simulation analysis in Contiki Cooja by considering an example scenario on two different motes namely: Sky and Z1. The aim of this analysis is to come up with the best suited symmetric key based cipher for the linear and grid based WSN.

KEYWORDS

Linear and Grid Based Wireless Sensor Networks, Symmetric Key Based Ciphers, Performance Analysis, Contiki Cooja.

1. INTRODUCTION

The Wireless Sensor Networks (WSNs) are considered to be deployed in random or tree based fashion. However, there are applications of WSN that form specific topology like linear or grid based WSN. The applications where objects being monitored are distributed in either linear or square grid inherently form linear and grid based WSN. Examples of the same are:

i. Monitoring the traffic level of city streets.

ii. Monitoring pipelines carrying oil, water or gas.

iii. Monitoring goods in a warehouse.

iv. Perimeter surveillance.

The typical examples of nodes forming a linear network and square grid are as shown in Fig. 1 and 2 respectively. As shown in Fig. 1, there are 7 nodes deployed in a linear fashion and each node has a communication range of 2. Therefore, such network is known as (7, 2) linear network.
As shown in Fig. 2, there are 50 houses of a colony deployed in a square grid manner. Here, each house is considered to have a device that is used for monitoring the energy consumption of the house. These devices pass the aggregated data in a hop by hop manner to the aggregator node, which send the data to the BS. Based on the considered application, there can be multiple aggregator nodes. The same deployment of devices can be considered for the applications of grid based networks. When we consider a single row of houses, it forms a linear network of 10 houses. Therefore, the grid based WSN are formed through the combination of linear networks (when 5 rows are considered, it forms grid based WSN). As the data are passed in hop by hop manner for getting the advantage of aggregation as discussed in [1], the security of the same is a critical issue. The intermediate nodes can alter or passively monitor the data and use it for their own advantage as discussed in [2, 3]. Therefore, the data are required to be encrypted before passing to the next node and the same is discussed in [4-6].
The lightweight schemes for the same are discussed in [7, 8]. However, these schemes do not take the deployment knowledge of sensor nodes under consideration unlike the secure data aggregation scheme for linear WSN proposed in [9]. For protecting the privacy of data, there are schemes proposed in the literature [10-13]. The scheme discussed in [10] is based on the idea of pseudonym changing. The idea of anonymous communication is presented in [11]. The key establishment scheme for the networks monitoring roads is discussed in [14]. The energy efficient communication protocol based on the idea of clustering in road networks is presented in [15]. Using the public key cryptography without using the certificates, the authentication scheme is discussed in [16]. In this paper, we perform the simulation analysis to find out the number of CPU cycles and memory consumption required in providing data security for linear and grid WSN.

The devices (sensor nodes) used in the considered applications are constrained with regard to the computation capabilities, battery power, storage, bandwidth [17, 18]. Therefore, we perform an analysis on the symmetric key based ciphers (as asymmetric key based ciphers require more storage and computation) and analyse the same for the considered example scenario. The performance parameters used in the simulation analysis are: RAM, ROM consumption and number of CPU cycles. The keys are pre-distributed as proposed in [19]. Such analysis is already discussed in the literature [20-22]. However, they have not considered the knowledge of deployment of nodes and the specific topology formed by the application. The main objective of our work is to come up with the best suited cipher for the considered scenario of linear and grid based WSN. Moreover, we perform our analysis on two different motes: sky and z1. The z1 motes have more capability regarding the RAM and the off memory and the analysis results shows that they require lesser number of CPU cycles as compared to sky motes. In addition, the comparison of two operating systems contiki cooja and tinyOS for sensor nodes is discussed in the paper. We perform a detailed analysis for one of the most significant cryptographic primitives of WSNs: Symmetric Key Block Cipher. We consider the performance parameters, storage and energy for a set of candidate lightweight ciphers. Analysing the performance of the symmetric key based ciphers with respect to the performance parameters: number of CPU cycles and RAM, ROM in contiki cooja, is the major contribution of this paper.

1.1 Organization of the Paper

The rest of the paper is organized as follows: In Section 2, brief introduction about the examined symmetric key based ciphers is discussed. In Section 3, we look at the simulation setup and the methodology for the evaluation of the ciphers. Moreover, the comparison of the operating systems contiki cooja and tinyOS is discussed in this section. In Section 4, we show the simulation results and discuss the same. We summarize our work with conclusions in Section 5.

2. THE SYMMETRIC KEY BASED CIPHERS: EXAMINED

In this section, we discuss block ciphers that are lightweight in nature. Different ciphers are based on different structures like Substitution Permutation Network (SPN), Feistel or Lai-Massey. AES [23]; KLEIN [24]; LED [25]; PRESENT [26] are based on SPN structure. HIGHT [27]; LBlock [28]; MIBS [29]; PICCOLO [30]; SEA [31]; SIMON [32]; TWINE [33] are based on Feistel structure. SPECK [32] is based on ARX (Add-Rotate-Xor) structure. IDEA [34] is based on Lai-Massey structure. We select these ciphers for the analysis as the applications we considered require lightweight solution and with the analysis of the paper we come up with the best suited cipher from the considered lightweight ciphers. We provide an overview of the examined ciphers and the attacks that are possible on each of them. We do not go in the designing details of the ciphers as our main focus is to analyse the performance in terms of CPU cycles and RAM, ROM consumption.
2.1. Substitution Permutation Network (SPN) Structure and Related Ciphers

SPN structure [35] takes a plaintext block and a key as inputs, and performs exchanging "rounds" of substitution (S) and permutation boxes (P-boxes) respectively to deliver the ciphertext block. An S-box substitutes a block of bits given as input by another block of bits as the output. This substitution must guarantee invertibility. A P-box is a permutation of all the bits: it takes input from the outputs of all the S-boxes of previous round, applies permutation of bits, and augments them into the S-boxes of the following round. The key is combined using some group operation like XOR at each round. The S-boxes and P-boxes transformations are efficient to perform in device (like sensors), E.g. exclusive or (XOR) and bitwise rotation.

The following ciphers are based on SPN structure:

**Advanced Encryption Standard (AES).** We analyse two different implementations of Advanced Encryption Standard (AES) ciphers. One is publicly available and other is designed by contiki cooja developers. Contiki has LLSec (Link Layer Security) layer. This layer is hardware independent, as it uses generic AES driver API instead of directly accessing the hardware. There are multiple AES drivers implemented in Contiki - software-only version and a couple of hardware accelerated ones, including for CC2420 (the radio chip on Sky mote). Authors of [36] show a possible attack on AES, known as biclique cryptanalysis. It uses the concept of exhaustive search on the key with an improvement by linking the keys through key schedule. This attack takes a time complexity of $2^{126.2}$ AES encryptions on the data amount $2^{88}$. The other possible attack is meet-in-the-middle [37] that takes less than $2^{100}$ data/time/memory complexity.

**KLEIN.** We analyse two different implementations of this cipher: KLEIN64 and KLEIN96. Both the implementations take 64 bits block size. Key lengths are 80 and 96 bits respectively. The number of rounds can either be 12, 16, or 20. The possible attack on this cipher is chosen plaintext key recovery as discussed in [38].

**LED.** We analyse two different implementations of this cipher: LED64 and LED128. Both the implementations take 64 bits block size. Key lengths are 64 and 128 bits respectively. The number of rounds is 32 and 48 respectively. This cipher does not use key schedule and this is the main difference from other ciphers. The XORing of key is done after every four rounds instead of key schedule. The number of rounds of this cipher is more as compared to other ciphers for compensating the key schedule. The differential cryptanalysis results on this cipher are discussed in [39]. The attacks on LED64 can be reduced to 12 and 16 rounds is described by the authors. The other possible attack is meet-in-the-middle as discussed in [40]. The complexity of the attack on 8 rounds of LED64 and 16 rounds of LED128 is lesser as compared to exhaustive key search.

**PRESENT.** It is the most popular cipher among all lightweight block ciphers. We analyse two different implementations of this cipher: PRESENT Size and PRESENT Speed. Both implementations take 64 bits block size. Key lengths are either 80 or 128 bits with number of rounds as 31. There many cryptanalysis results as discussed in [41-43]. Authors in [44] discuss about two bicliques possible on two implementation of PRESENT.

2.2. Feistel Structure and Related Ciphers

Feistel structure [45] takes plaintext block as input and divides it into two halves, L (left) and R (right). R half is given as input to a feistel function along with the round key. It is also used as an L half for the next round. Output of the feistel function is XORed with L half and used as an R half for the next round. The same process is repeated till last round. The advantage with this structure is, just by reversing the key schedule decryption can be done.
The following ciphers are based on feistel structure (or a modified feistel structure):

**HIGHT.** This cipher uses block size of 64 and key length of 128 bits. It uses 32 rounds and sometimes uses modular addition instead of XOR operation. The biclique attack against HIGHT is proposed in [46]. Moreover, the differential cryptanalysis attack is described in [47].

**LBLOCK.** This cipher uses block size of 64 and key length of 80 bits. The number of rounds is 32 and the usage of 8 S-boxes and permutation of 4 bits are applied. The biclique attack against LBLOCK is proposed in [48]. The authors also discuss the prevention of this attack with the help of modified key schedule algorithm.

**MIBS.** We analyse two different implementations of this cipher: MIBS64 and MIBS80. Both implementations take 64 bits block size. Key lengths are 64 and 80 bits respectively. The linear attacks on MIBS are discussed in [49]. The authors show the differential cryptanalysis on 14 rounds, ciphertext only attacks on 13 rounds and an impossible differential attack on 12 rounds of MIBS.

**PICCOLO.** We analyse two different implementations of this cipher: PICCOLO80 and PICCOLO128. Both implementations take 64 bits block size. Key lengths are 80 and 128 bits respectively. It uses two feistel functions. It requires less than 1000 gates when implemented on hardware. Authors of [44] discuss the biclique attacks on both the implementations of PICCOLO.

**SEA.** This cipher uses \( n \) bits block size. The value of \( n \) can be 48, 96, or 144 bits. It uses a two branch feistel structure as modified feistel structure. The security analysis is discussed in [31].

**SIMON.** This cipher uses different block sizes like 32, 48, 64, 96, 128 bits. It is based on a balanced feistel network. The key size can be 64, 72, 96, 128, 144, 192, 256 bits. It is optimized for the hardware implementations. The differential cryptanalysis is possible on this cipher as discussed in [50, 51].

**TWINE.** We analyse two different implementations of this cipher: TWINE80 and TWINE128. Both the implementations take 64 bits block size. Key lengths are 80 and 128 bits respectively. The number of rounds is 36 in both the implementations. The feistel function uses a single Sbox and subkey addition. This function is repeated 8 times in each round. Two biclique attacks on two implementations are discussed in [52].

### 2.3. Add-Rotate-Xor Structure and Related Cipher

This structure involves 3 operations:

1. Modular Addition
2. Rotation with fixed rotation amounts
3. XOR

These ARX operations are immune to timing attacks because they run in defined constant time. As these operations are fast and cheap in hardware and software, the ciphers based on ARX operations are popular.

**SPECK.** This cipher uses different block sizes like 32, 48, 64, 96, 128 bits. It is based on an Add-Rotate-Xor (ARX) structure. The key size can be 64, 72, 96, 128, 144, 192, 256 bits. It is
optimized for the software implementations. The differential cryptanalysis is possible on this cipher as discussed in [50, 51].

2.4. Lai-Massey Structure and Related Cipher

As shown in Fig. 3, Lai-Massey Structure [53] divides the plaintext in two equal halves $L_0$ and $R_0$ as input. Two round functions are used; $H$ and $F$. Keys are used with function $F$. The output of function $H$ is given as input to function $F$.

$$(L_{i+1}^\prime, R_{i+1}^\prime) = H(L_i^\prime + T_i, R_i^\prime + T_i)$$

where $$T_i = F(L_i^\prime - R_i^\prime, K_i)$$ and $$(L_0^\prime, R_0^\prime) = H(L_0, R_0)$$

Fig. 3. Lai-Massey Structure

**IDEA.** This cipher uses 64 bits block size and 128 bits key. It is composed of 8.5 rounds. A final round that is "half round", comes after eight rounds and used for the output transformation (the swap of the centre two values counterbalances the swap toward the end of the last round, so that there is no net swap). It is included in the package PGP (Pretty Good Privacy). There is six rounds attack that exploits key schedule of IDEA with linear cryptanalysis as discussed in [54]. The biclique framework is used by the authors of [55] to speed up the key recovery.
The summary of all the examined symmetric key based block ciphers with regard to the structure they are built on, block size, key size, and the attacks that are possible on each of them is shown in Table 1. From all the examined ciphers, SIMON and SPECK are considered to be the block ciphers for IoT (Internet of Things) environment as discussed in [56].

Table 1. Summary of Block Ciphers Examined.

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Cipher</th>
<th>Reference</th>
<th>Block Size (bits)</th>
<th>Key Size (bits)</th>
<th>Structure</th>
<th>Possible Attacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AES</td>
<td>[23]</td>
<td>128</td>
<td>128</td>
<td>SPN</td>
<td>- Biclique</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cryptanalysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[36]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Meet-in-the-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>middle [37]</td>
</tr>
<tr>
<td>2</td>
<td>HIGHT</td>
<td>[27]</td>
<td>64</td>
<td>128</td>
<td>Fiестel</td>
<td>- Differential</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cryptanalysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[47]</td>
</tr>
<tr>
<td>3</td>
<td>IDEA</td>
<td>[34]</td>
<td>64</td>
<td>128</td>
<td>Laи-Mасsey</td>
<td>- Linear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cryptanalysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[54]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Biclique</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[55]</td>
</tr>
<tr>
<td>4</td>
<td>KLEIN</td>
<td>[24]</td>
<td>64</td>
<td>64,96</td>
<td>SPN</td>
<td>- Chosen</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>plaintext [38]</td>
</tr>
<tr>
<td>5</td>
<td>LBLOCK</td>
<td>[28]</td>
<td>64</td>
<td>80</td>
<td>Fiестel</td>
<td>- Biclique</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[48]</td>
</tr>
<tr>
<td>6</td>
<td>LED</td>
<td>[25]</td>
<td>64</td>
<td>64,128</td>
<td>SPN</td>
<td>- Differential</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cryptanalysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[39]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Meet-in-the-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>middle [40]</td>
</tr>
<tr>
<td>7</td>
<td>MIBS</td>
<td>[29]</td>
<td>64</td>
<td>64,80</td>
<td>Fiестel</td>
<td>- Differential</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cryptanalysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[49]</td>
</tr>
<tr>
<td>8</td>
<td>PRESENT</td>
<td>[26]</td>
<td>64</td>
<td>80</td>
<td>SPN</td>
<td>- Cryptanalysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[41-43]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Biclique</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[44]</td>
</tr>
<tr>
<td>9</td>
<td>PICCOLO</td>
<td>[30]</td>
<td>64</td>
<td>80,128</td>
<td>Fiестel</td>
<td>- Biclique</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[44]</td>
</tr>
<tr>
<td>10</td>
<td>SEА</td>
<td>[31]</td>
<td>96</td>
<td>96</td>
<td>Fiестel</td>
<td>- Cryptanalysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[31]</td>
</tr>
<tr>
<td>11</td>
<td>SIMON</td>
<td>[32]</td>
<td>32,48,64, 96,128</td>
<td>64,72,96,128,</td>
<td>Fiестel</td>
<td>- Differential</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cryptanalysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[50, 51]</td>
</tr>
<tr>
<td>12</td>
<td>SPECK</td>
<td>[32]</td>
<td>32,48,64, 96,128</td>
<td>64,72,96,128,</td>
<td>ARX</td>
<td>- Differential</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cryptanalysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[50, 51]</td>
</tr>
<tr>
<td>13</td>
<td>TWINE</td>
<td>[33]</td>
<td>64</td>
<td>80,128</td>
<td>Fiестel</td>
<td>- Biclique</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[52]</td>
</tr>
</tbody>
</table>

3. Simulation Setup and The Methodology For Evaluation

We analyse the number of CPU cycles and RAM, ROM consumption in achieving data security for the secure data aggregation scheme proposed in [9]. The scheme uses the pre distributed keys (proposed in [19]) for the purpose of encryption. We work with the nodes that are constrained regarding:

- Storage
- Communication range (It is assumed that nodes can communicate at least till one hop)
- Battery life
Contiki Cooja is a simulator specifically designed for IoT devices that are having the constraints as described. It is also used as an emulator because the code to be executed by the node is the exact same firmware one may upload to physical nodes [57].

### 3.1. Simulation Setup

This section discusses the simulation results for the considered example scenario. We focus on the data security as it is crucial when designing a data aggregation scheme. Passively acquired data can be used for malicious purpose if confidentiality of data is not taken care of. Our criteria are to measure the number of CPU cycles and RAM, ROM consumption for providing data security. In contiki cooja, there are several options regarding the selection of devices for which one wants to emulate. E.g. Cooja mote, MicaZ mote, CC430 mote, Z1 mote, Sky mote, etc. The comparison between the operating systems TinyOS and Contiki is discussed in [58]. In TinyOS, the application has to be replaced totally when the code is changed. However, the contiki OS is better when it comes to updating the deployed application as it can dynamically replace the changed programs. The protocols discussed in [9, 59-60] require the code to be updated every time the value of N (total number of nodes in the network) or k (number of consecutive nodes) is changed. Moreover, Contiki supports dynamic loading and unloading of the code and multithreading. Contiki is an event driven OS and event handlers cannot pre-empt each other. However, interrupts can pre-empt the current running process.

We use Sky motes and Z1 motes of contiki cooja for the purpose of simulation. Sky mote features a 16-bit MSP430 MCU, 10 kB RAM, 48 kB ROM, a cc2420 802.15.4 radio transceiver, an external Flash memory, and temperature, humidity and brightness sensor [17]. Z1 mote has higher configuration and uses MSP430F2617 MCU [18]. The specifications considered for Sky and Z1 motes are as shown in Table 2 and Table 3 respectively.

<table>
<thead>
<tr>
<th>Flash Memory</th>
<th>48 KB</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAM</td>
<td>10 KB</td>
</tr>
<tr>
<td>Current Consumption</td>
<td>20 mA</td>
</tr>
<tr>
<td>Operating Voltage</td>
<td>3 V</td>
</tr>
<tr>
<td>Micro-controller</td>
<td>MSP430</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Flash Memory</th>
<th>92 KB</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAM</td>
<td>8 KB</td>
</tr>
<tr>
<td>Current Consumption</td>
<td>19.7 mA</td>
</tr>
<tr>
<td>Operating Voltage</td>
<td>3 V</td>
</tr>
<tr>
<td>Micro-controller</td>
<td>MSP430F2617</td>
</tr>
</tbody>
</table>

### 3.1. Methodology for Evaluation of Ciphers

In this section, we analyse the symmetric key based block ciphers with regard to the number of CPU cycles, energy and memory they consume if applied on Sky and Z1 Motes. The energy is calculated through following steps:
- Add the header file `#include “energest.h”` in a `.c` file of the considered cipher.

- To get CPU cycles involved in the different ciphers, add `“printf(“energy cpu: %lu”, energest_type_time(ENERGEST_TYPE_CPU));”` line in `PROCESS_THREAD` of the `.c` file.

- To get the power consumption, the formula is: \[ \text{Power(mW)} = \frac{r_{ON}}{CPU + LPM} \times 20mA \times 3V \]

Sky and Z1 motes have 20mA current value and 3V voltage. Therefore, CPU cycles received from running the code on Sky or Z1 motes, if multiplied with 60 will give the power consumption in Watts. When this value is multiplied with simulation time, it gives energy consumption in joules. For getting the number of CPU cycles involved in ciphers, we run each cipher separately and follow the steps as discussed. Fig. 4 shows the simulation result, when we run a TWINE cipher on a grid of 20 Sky motes.

In order to obtain the memory (RAM, ROM) consumption, we use “size” command. Fig. 5 shows the use of size command. Here, `.text` column refers to the ROM consumption by different ciphers in bytes. The `.data` and `.BSS` columns show the RAM consumption. We run all the ciphers in the same manner on both Sky and Z1 motes respectively and the results are as shown in Tables 4 and 5.
4. SIMULATION RESULTS

We can see from the Table 4 that, each cipher requires different amount of CPU cycles and RAM, ROM consumption in the grid network of 20 sky motes. The one that requires maximum number of CPU cycles is LED128 (11117) and the one that requires minimum number of CPU cycles is KLEIN64 (1401). The AES designed by contiki developers specifically for sky motes uses hardware acceleration that helps in reducing number of CPU cycles compared to publicly defined AES (from Table 4, we can see AES (Contiki) requires 1503 whereas AES (Public) requires 1582 number of CPU cycles). SPECK (128 bits block and key size) cipher is designed specifically for resource constrained environments requires 1403 number of CPU cycles.

Table 4. Sky Mote: No. of CPU Cycles and RAM, ROM

<table>
<thead>
<tr>
<th>Cipher</th>
<th>CPU Cycles</th>
<th>RAM,ROM (bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AES (Contiki)</td>
<td>1503</td>
<td>49973</td>
</tr>
<tr>
<td>AES (Public)</td>
<td>1582</td>
<td>51329</td>
</tr>
<tr>
<td>HIGHT</td>
<td>1461</td>
<td>50093</td>
</tr>
<tr>
<td>IDEA</td>
<td>2375</td>
<td>50151</td>
</tr>
<tr>
<td>KLEIN64</td>
<td>1401</td>
<td>50719</td>
</tr>
<tr>
<td>KLEIN96</td>
<td>1562</td>
<td>50763</td>
</tr>
<tr>
<td>LBLOCK</td>
<td>1404</td>
<td>50555</td>
</tr>
<tr>
<td>LED64</td>
<td>7819</td>
<td>50149</td>
</tr>
<tr>
<td>LED128</td>
<td>11117</td>
<td>50133</td>
</tr>
<tr>
<td>MIBS64</td>
<td>1559</td>
<td>50381</td>
</tr>
<tr>
<td>MIBS80</td>
<td>1632</td>
<td>50947</td>
</tr>
<tr>
<td>PRESENT_Size</td>
<td>4224</td>
<td>51187</td>
</tr>
<tr>
<td>PRESENT_Speed</td>
<td>3715</td>
<td>51251</td>
</tr>
<tr>
<td>PICCOLO80</td>
<td>1487</td>
<td>50055</td>
</tr>
<tr>
<td>PICCOLO128</td>
<td>1512</td>
<td>50111</td>
</tr>
<tr>
<td>SEA</td>
<td>1665</td>
<td>49923</td>
</tr>
<tr>
<td>SIMON128</td>
<td>1808</td>
<td>50731</td>
</tr>
<tr>
<td>SPECK128</td>
<td>1403</td>
<td>49995</td>
</tr>
<tr>
<td>TWINE80</td>
<td>1668</td>
<td>49943</td>
</tr>
<tr>
<td>TWINE128</td>
<td>1716</td>
<td>50137</td>
</tr>
</tbody>
</table>
Table 5. Z1 Mote: No. of CPU Cycles and RAM, ROM

<table>
<thead>
<tr>
<th>Cipher</th>
<th>CPU Cycles</th>
<th>RAM,ROM (bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AES (Contiki)</td>
<td>633</td>
<td>49279</td>
</tr>
<tr>
<td>AES (Public)</td>
<td>285</td>
<td>49131</td>
</tr>
<tr>
<td>HIGHT</td>
<td>167</td>
<td>48409</td>
</tr>
<tr>
<td>IDEA</td>
<td>643</td>
<td>48453</td>
</tr>
<tr>
<td>KLEIN64</td>
<td>184</td>
<td>48995</td>
</tr>
<tr>
<td>KLEIN96</td>
<td>262</td>
<td>49043</td>
</tr>
<tr>
<td>LBLOCK</td>
<td>176</td>
<td>48479</td>
</tr>
<tr>
<td>LED64</td>
<td>2899</td>
<td>48499</td>
</tr>
<tr>
<td>LED128</td>
<td>4342</td>
<td>48483</td>
</tr>
<tr>
<td>MIBS64</td>
<td>242</td>
<td>48421</td>
</tr>
<tr>
<td>MIBS80</td>
<td>277</td>
<td>48631</td>
</tr>
<tr>
<td>PRESENT_Size</td>
<td>1300</td>
<td>49021</td>
</tr>
<tr>
<td>PRESENT_Speed</td>
<td>896</td>
<td>49043</td>
</tr>
<tr>
<td>PICCOLO80</td>
<td>237</td>
<td>48383</td>
</tr>
<tr>
<td>PICCOLO128</td>
<td>251</td>
<td>48451</td>
</tr>
<tr>
<td>SEA</td>
<td>310</td>
<td>48319</td>
</tr>
<tr>
<td>SIMON128</td>
<td>373</td>
<td>49079</td>
</tr>
<tr>
<td>SPECK128</td>
<td>128</td>
<td>48347</td>
</tr>
<tr>
<td>TWINE80</td>
<td>307</td>
<td>48333</td>
</tr>
<tr>
<td>TWINE128</td>
<td>336</td>
<td>48531</td>
</tr>
</tbody>
</table>

When we run all the ciphers on the Z1 mote, it takes a lesser number of CPU cycles as we can see from Table 5 (AES (pub) on sky mote takes 1582 CPU cycles, whereas on Z1 mote it takes 285 CPU cycles). The AES code provided by contiki cooja developers is specifically designed for Sky mote by using hardware acceleration. Therefore, the RAM consumption of the same is lesser as shown in Table 4. The number of CPU cycles for Z1 motes is always lesser compared to Sky motes, as the configuration of Z1 mote is superior concerning the flash memory that can be used as either RAM or ROM. Therefore, the number of CPU cycles for running different ciphers is lesser for Z1 motes compared to Sky motes as shown in Tables 4 and 5.

![Fig 6. CPU Cycles](image-url)
The comparison of the CPU cycles required by both the motes is as shown in Fig. 6. It shows that Z1 mote takes lesser number of CPU cycles as compared to Sky mote for all the ciphers. The comparison in terms of RAM and ROM consumption of both the motes is as shown in Fig. 7. It shows that Z1 mote takes lesser amount of RAM, ROM consumption as compared to Sky mote in running all the ciphers.

Fig 7. RAM, ROM

5. CONCLUSIONS

We analysed symmetric key based block ciphers on two different motes (sky and z1) in contiki cooja. This analysis concerning the number of CPU cycles and RAM, ROM consumption helps in deciding which cipher can be used for the secure data aggregation scheme in different scenarios. In the constrained scenario of sensor nodes, it is better to use lightweight ciphers such as HIGHT; KLEIN; PICCOLO; SIMON; SPECK or TWINE. The hardware accelerated AES cipher by contiki cooja uses minimum number of CPU cycles when we take Sky mote under consideration. Since speed is correlated with energy consumption, SPECK 128/128 is a better choice in energy critical applications as it produces energy efficient solution with an encryption cost of 1403 cycles on Sky mote, or 128 cycles on Z1 mote. We have given a detailed analysis for one of the most significant cryptographic primitives for WSNs: Symmetric Key Block Cipher, by considering the performance parameters, storage and energy for a set of candidate lightweight ciphers. We are working on optimizing the AES cipher regarding speed (reducing the number of CPU cycles) and size (reducing the RAM, ROM consumption). This optimized version of AES will be hardware independent, i.e. it will not depend on the mote under consideration (E.g. sky or z1) and produce the optimal results.

REFERENCES


AUTHORS

Kaushal Shah is a PhD research scholar in Computer Engineering at the Department of Computer Engineering, S. V. National Institute of Technology, Surat, India. He has received his M.E. degree in Computer Science and Engineering from Government Engineering College, Modasa, India. His research interests broadly include Information Security, Wireless Sensor Networks and Protocol designing.

Devesh Jinwala has been working as a Professor in Computer Engineering at the Department of Computer Engineering, S. V. National Institute of Technology, Surat, India since 1991. His principal research areas of interest are broadly Security, Cryptography, Algorithms and Software Engineering. Specifically his work focuses on Security and Privacy Issues in Resource constrained environments (Wireless Sensor Networks) and Data Mining, Attribute-based Encryption techniques, Requirements Specification, and Ontologies in Software Engineering. He has been/is the Principal Investigator of several sponsored research projects funded by ISRO, GUJCOST, Govt of Gujarat and DiETY-MCIT-Govt of India.
A PROPOSED HSV-BASED PSEUDO-COLORING SCHEME FOR ENHANCING MEDICAL IMAGES

Noura A. Semary

Faculty of Computers and Information, Menoufia University, Egypt

ABSTRACT

Medical imaging is one of the most attractive topics of image processing and understanding research fields due to the similarity between the captured body organs colors. Most medical images come in grayscale with low contrast gray values; which makes it a challenge to discriminate between the region of interest (ROI) and the background (BG) parts. Pseudo-coloring is one of the solutions to enhance the visual appeal of medical images, most literature works suggest RGB-base color palettes. In this paper, pseudo-coloring methods of different medical imaging works are investigated and a highly discriminative colorization method is proposed. The proposed colorization method employs HSV/HSI color models to generate the desired color scale. Experiments have been performed on different medical images and different assessment methods have been utilized. The results show that the proposed methodology could clearly discriminate between near grayscale organs especially in case of tumor existence. Comparisons with other literary works were performed and the results are promising.

KEYWORDS

Medical imaging, pseudo-coloring, colorization, HSV

1. INTRODUCTION

Medical images not only represent structural appearance information, they are also capable of examining complex and sophisticated internal biological processes. Medical imaging contributes to many disease diagnoses and also plays an important role in understanding the human anatomy which guides surgical assistance during the procedure.

Medical Imaging has various modalities and applications. X-ray, CT, MRI, Ultrasound, Mammogram, Nuclear Medicine, PET, Ultrasound, and Thermal imaging are some of the famous imaging technologies [1, 2]. Each of which has its suitable applications and features and exports gray shades images which usually come in low contrast intensities.

Digital image can be represented in different formats; 1) Grayscale (8-bits/pixel), 2) True Color (24 bits/pixel) and 3) Indexed (8-bits/pixel index image + Color Map) [3]. Medical images usually come in grayscale which has only 256 gray shades variations. While most image understanding researchers prefer to deal with color images instead of grayscale, as the color variations exceed 16 million degrees of colors.
For some imaging technologies, a pseudo-coloring system may be embedded in the imaging device. Pseudo-Coloring gives non-real colors to the grayscale image by converting it to an indexed image with a fixed color map. Generating the color map is the main contribution in this field of research.

There are two approaches for medical images colorization in the literature; semi-automatic (interactive) pseudo-coloring [1] and automatic (non-interactive) pseudo-coloring [4-10].

In this paper, we are concerned with automatic medical images pseudo-coloring. The rest of this paper is organized as follows, Section 2 presents automatic medical colorization literature review. The proposed colorization method is presented in Section 3 in details. Section 4 presents the experimental results and a comparative study of the proposed method and other different methods. Finally, Section 5 concludes this paper.

2. LITERATURE REVIEW

L. H. Juang and M. N. Wub [4] used color-converted segmentation with K-means clustering technique for Brain MRI tumor objects colorization and tracking. Starting with a gray image, the original image was segmented by k-means and mapped colors to the segments by using color-convert which starts by R, G, B then mapped to a single index value. Compare between Otsu segmentation results and claim 96% accuracy and 10 minutes processing time!

M. d. C. V. Hernández et. al. [5] used image fusion between T2W and FLAIR images for differentiating normal and abnormal brain tissue, including white matter lesions (WMLs). They modulated two 1.5T MR sequences in the red/green color space and calculated the tissue volumes using minimum variance quantization. M. Attique et. al. [6] also used image fusion for brain MRI images. They utilized Single slice of T2-weighted (T2) brain MR images using two methods; (i) A novel colorization method to underscore the variability in brain MR images, indicative of the underlying physical density of bio-tissue, (ii) A segmentation method to characterize gray brain MR images.

M. Martinez et. al. [7] proposed their color map for CT Liver images, the ROI is selected manually. Each grayscale pixel was assigned a color value (R, G, B) based on a generated color map. A color scheme was developed where the lowest tissue density value was colored red, blending towards green as the tissue density value increases and continued to blend from green to blue for the next range of increasing tissue densities. An associated segmentation process is then tailored to utilize this color data. It is shown that colorization significantly decreases segmentation time. M. E. Tavakol et. al [8] applied their proposed colorization system on Thermal Infrared Breast Images. Lab color model is considered. Two color segmentation techniques, K-means and fuzzy c-means for color segmentation of infrared (IR) breast images are modelled and compared.

Z. Zahedi et. al [9] proposed their colorization system for breast thermal images as a nonlinear function transforms for pseudo-coloring of infrared breast images based on physiological properties of the human eye. N. S. Aghdam et. al [10] proposed four pseudo-coloring algorithms for breast thermal images. The first two algorithms are in HSI color space and the other two are in CIE L*a*b*.
3. PROPOSED COLORIZATION SCHEME

The proposed colorization system is based on HSV/HSI color models where any RGB color triple value can be expressed in terms of Hue (H) Saturation (S) and Intensity (I or value V). Since the grayscale image has only intensity component and has no hue or saturation values [3], our system is based on generating suitable hue and saturation values for each intensity level. Figure 1 presents the main block diagram of the proposed system.

3.1. Intensity Component Generation

In order to generate a carefully designed pseudo-color coding which could preserve all the information of grayscale images and does not generate any distortion in the image, the grayscale image is loaded and saved as the intensity component for the output image.

\[ I = G \]  

(1)

where L is the Intensity component in HSV/HIS image and G is the gray image.

3.2. Hue Component Generation

Since the gray image has different gray shades which reflect different medical meanings, it’s suggested to use the same varieties for the domain color of each region. Human vision can discriminate between the main Rainbow colors; red, orange, yellow, green, cyan, blue and magenta, which can be generated by Hue component. Hue component reflects the dominant color of the pixel. Here, we studied 3 strategies of generation the Hue component.

A. Equal to Intensity (EI)

In this strategy, Hue is set to the same value of Intensity (2). That makes the shades vary from red to magenta for the gray shades from black to white respectively. For some medical images, the ROI gray shades are in light areas with very close values. By using this strategy the colors of ROI may come in red and magenta which may not be discriminative enough.

\[ H_{EI} = G \]  

(2)

B. Complement of Intensity (CI)

In this strategy, Hue is set to the Intensity complement value (3). That makes the shades vary from red to magenta for the gray shades from white to black respectively

\[ H_{CI} = 1 - G \]  

(3)

C. Stretched Inverted Intensity (SI)

Since the gray levels of medical images are very close, a stretched grayscale image has been generated by contrast stretching function (4). The aim of stretching function is to generate wider color space than the limited gray levels in the source gray image.
\[ H_{SI} = 1 - \frac{(b-a)}{(d-c)} (G - c) + a \] (4)

where \( G \) is the input gray image. The parameters \( a \) and \( b \) are the minimum gray level and the maximum gray level in the input image, while \( c \) and \( d \) are the minimum gray level and the maximum gray level in the desired image respectively.

3.3. Saturation Component Generation

For the Saturation component, the original gray image is used again to generate the suitable saturation. Since it is important to see the organs in full clear color, while neglecting the background, an adaptive thresholding [3] has been used to generate a binary image of black background pixels and white foreground pixels (5).

\[
S = \begin{cases} 
1 & I \geq th \\
0 & I < th 
\end{cases} \] (5)

Figure 1. Proposed colorization system
The complete process can be illustrated mathematically by (6).

\[
HSL = \begin{cases} 
    H_{SL} = 1 - \frac{(b - a)}{(d - c)} (G - c) + a \\
    S = 1 & I \geq th \\
    S = 0 & I < th \\
    L = G 
\end{cases}
\]  

(6)

There is a little difference between HSV and HSI color models; where the pure colors of full saturation are at the top of HSV while they are in the middle at the HSI. In our proposed system, both models are used considering the same transformations (6). Finally, the final RGB color image is generated after the well-known HSV/HSL to RGB conversion [3].

4. EXPERIMENTAL RESULTS

The proposed system has been implemented in MatlabR2013 on 8G-RAM 64bit-OS Windows 8 machine and different medical imaging modalities have been used for testing. A comparison between the proposed coloring system and other systems has been made with regard to different assessment methods.

4.1. Subjective Assessment

Subjective assessment methods are utilized to perform the task of assessing visual quality to the human subjects. This section presents the visual results of the proposed system as well as different literature methods [7-10]. Figure 2.a presents a grayscale image that presents the 256 shades of gray and the color palettes for the methods of [7-10] compared to the proposed method with 6 strategies; (3 for HSV and 3 for HSL). In this example, the saturation threshold has been set to 0.1 for illustration. Figure 3 presents more visual results for different types of medical imaging; Brain MRI, Breast Thermal, Liver CT, and Bone MRI. Generally, it seems that our proposed system could successfully discriminate between the background and the foreground, since it gave colors only to the desired imaged organs. It appears that it gives more details to the images by giving a lot of color varieties from hot (red) to cold (blue).

Mean Opinion Score (MOS) [11] is a subjective measure that can be presented in numbers. It measures the users’ satisfaction with the visual results. It is calculated by averaging the results of a set of subjective test that gives a score to the results image from 1 (bad) to 5 (Excellent). The mean rate of a group of observers who join the evaluation is usually computed by the following equation:

\[
MOS = \frac{1}{N} \sum_{g=1}^{5} g p(g)
\]

(7)

, where g is grade and p(g) is grade probability. We have performed the MOS test using an online test. 60 volunteers have rated the different 12 methods we presented in this paper by giving them rates from 1 to 5. The obtained MOS is presented in Figure 4.
Figure 2. (a) 256 shaded gray palette (b-g) Color Palettes of [7-10] methods and (h-m) the proposed color palette with 6 strategies.
d. Aghdam HSI [10]
e. Aghdam Lab [10]
f. Proposed HSV-EI
g. Proposed HSV-CI
h. Proposed HSV-SI
i. Proposed HSL-EI
j. Proposed HSL-CI
4.2. Objective Assessment

It’s difficult to find the best objective metric to evaluate gray to color images conversion since no reference color image is provided. MSE (8) and PSNR (9) are widely known objective quality assessment methods for image enhancement. Since they measure the difference between two images in the same color space, it makes no meaning to get the MSE between a gray image and a colored one. Some works consider the MSE as an objective assessment method to their colorization methods by either getting the MSE between the original color image and the recolored one or to convert the gray image into RGB image with equal R, G and B values and calculate the difference between the gray RGB image and the colored one. In the latter case, it seems the higher the MSE, the higher the distance between color and gray values, the best the colorization result is.

\[
MSE = \frac{1}{3MN} \sum_{i=1}^{M} \sum_{j=1}^{N} [G(i,j) - Cl(i,j)]^2
\]  \hspace{1cm} (8)

\[
PSNR = 10 \log \left( \frac{255^2}{MSE} \right)
\]  \hspace{1cm} (9)

, where G is the gray image with 3 channels; R, G and B, and Cl is the colored image in RGB space.
Normalized Color Difference (NCD) [3] [12] is another objective measure that was used by some researchers. NCD represents the distances between colors in a given color space (Usually Lab color space). The larger the NCD, the worse the image quality is. The NCD indicator is calculated using the following formula:

\[
NCD = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \sqrt{\sum_{q=A,B} [G_q(i,j) - Cl_q(i,j)]^2}}{\sum_{i=1}^{N} \sum_{j=1}^{N} \sqrt{\sum_{q=A,B} Cl_q(i,j)^2}}
\]

, were \(G_q\) and \(Cl_q\) are the gray and the colored images with q channel (in Lab space).

Structure Similarity Metric (SSIM) [13, 14] is another objective measure that was proposed based on the human visual system. SSIM reflects how the colorization process affects the structure of the image. Popowicz [15] improved the SSIM by adding a color comparison to the criteria of the grayscale SSIM. Since the SSIM is defined only for grayscale images, it can be adapted for color image colorization by calculating the SSIM for every single color channel independently, then calculating the mean. When taking into consideration decorrelated color spaces, where the channels are orthogonal, we may assess the quality with the root of the sum of squared SSIM results obtained for each channel. The detailed Mean SSIM (MSSIM) could be found in [15].

In this work, MSE, PSNR, NCD, and SSIM measures have been calculated for works in [7-10] as well as the proposed method. Table 1 presents the objectives measures calculated for the proposed palette as well as the methods in [7-10]. As discussed before, MSE and PSNR give different meaning when it comes to colorization. As it is extremely sensitive to the offsets in color or luminance channels. It is possible that, although the visual impression of the image does not change, the PSNR value may be much lower, indicating the poor quality. From the table, it seems the proposed palettes have high MSE and low PSNR. Since NCD is the most suitable measure for colorization procedures assessment [15] it is considered for assessing the 12 methods. The lower the NCD, the better the image quality. Since our proposed system keeps the luminance channel untouched, it’s expected to have high MSSIM index. As the MSSIM could be presented as a grayscale image, Figure 5 presents the MSSIM for the Brain MRI image from our test set (Figure 3).

5. CONCLUSION

In this paper, we propose HSV/HSL pseudo-coloring schema for medical images. We have proposed 6 strategies based on the basic proposed structure. The proposed strategies vary in colors range (from red to blue and magenta or vice versa). Using HSV/HSL models enable giving rainbow colors to the gray shades either in the same order or with an inverse order. We recommend using the same gray image as Intensity for keeping the same structure of the image as well as the original intensity information. The recommended binary Saturation channel gives clear pure colors to the captured organs while neglecting the black background here what enhances the visual appeal of the images. Also, it increases the color difference between the original grayscale image (where saturation is set to 0) and the colored image (where the saturation is set to 1). For Hue component generation, three strategies have been suggested; Equal to Intensity (EI), Complement of Intensity (CI) and Stretched Inverted Intensity (SI). Hue contrast stretching has been recommended for giving more colors to a smaller range of gray shades. Subjective and objective testing has been performed showing that the proposed system has high
MOS and SSIM, and low NCD. We can conclude that using HSV/HSB color models in medical images is a good choice subjectively and objectively.

Table 1. Objective measures for literature palettes [7-10] and the proposed palettes

<table>
<thead>
<tr>
<th>Color Space</th>
<th>MSE</th>
<th>PSNR</th>
<th>NCD</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Martinez [7]</td>
<td>RGB</td>
<td>1.19E+04</td>
<td>16.9628</td>
<td>0.5497</td>
</tr>
<tr>
<td>Zahedi [9]</td>
<td>RGB</td>
<td>5.09E+03</td>
<td>25.4766</td>
<td>0.4488</td>
</tr>
<tr>
<td>Aghdam [10]</td>
<td>HSV_1</td>
<td>6.90E+03</td>
<td>22.4306</td>
<td>0.4803</td>
</tr>
<tr>
<td></td>
<td>HSV_2</td>
<td>5.30E+03</td>
<td>25.0627</td>
<td>0.4303</td>
</tr>
<tr>
<td>Aghdam [10]</td>
<td>Lab_1</td>
<td>4.62E+03</td>
<td>26.4341</td>
<td>0.3713</td>
</tr>
<tr>
<td></td>
<td>Lab_2</td>
<td>5.99E+03</td>
<td>23.8435</td>
<td>0.3854</td>
</tr>
<tr>
<td></td>
<td>HSV-U</td>
<td>1.06E+04</td>
<td>18.178</td>
<td>0.2933</td>
</tr>
<tr>
<td></td>
<td>HSV-SI</td>
<td>1.05E+04</td>
<td>18.2176</td>
<td>0.297</td>
</tr>
<tr>
<td></td>
<td>HSV-I</td>
<td>1.06E+04</td>
<td>18.178</td>
<td>0.3366</td>
</tr>
<tr>
<td></td>
<td>HSL-U</td>
<td>8.88E+03</td>
<td>19.9045</td>
<td>0.3302</td>
</tr>
<tr>
<td></td>
<td>HSL-SI</td>
<td>8.86E+03</td>
<td>19.9339</td>
<td>0.3247</td>
</tr>
<tr>
<td></td>
<td>HSL-I</td>
<td>8.88E+03</td>
<td>19.9045</td>
<td>0.3907</td>
</tr>
</tbody>
</table>

Figure 5. (a-f) MSSIM for Color Palettes of [7-10] methods and (g-l) MSSIM for the proposed color palettes.

REFERENCES


AUTHORS

Noura A. Semary works as an Assistant Professor in Faculty of Computers and Information, Menoufia University, Egypt. She has B.Sc. in 2001 from Cairo University, Faculty of Computers and Information. She worked as a Staff Member in the Faculty of Computers and Information, Menoufia University, Egypt in 2003. In 2007 and 2011 she has obtained her Master and Ph.D. degrees respectively in Information Technology from Computers and Information Faculty at Menoufia University. Her main research interests are in Image Processing, Computer Vision,
Data Compression, Data Hiding, Virtual Reality and Assistive Technology fields. In 2009 she won the first rank in “Made in Egypt” and “Made in Arab World” competitions for her project “Black and White Movies Colorizer”. She has about 40 publications in international conferences and journals.
AUTHOR INDEX

Devesh C. Jinwala  235
Douglas Lautner  127
Eric Nyiri  83
Farhad Foroughi  205
Ferdinand A. Katsriku  15
Ghada Badr  183
Hewei Yu  165
Jaewon Lim  63
Jamal-Deen Abdulai  15
James Dux  47
Jang Hyun Kim  63
Jennifer Priestley  97
JiHye Han  35
JinTao Huang  35
Joris Guérin  83
Kalpdrum Passi  111
KangNyeon Yi  35
Kaushal Shah  235
Keith Abney  73
Khalid Alsamara  219
Kingsley K. Ofosu  15
Lili Zhang  97
Martin Ciupa  73
Meiling Zhou  165
Niravkumar Pandey  111
Noor Al Madeed  143
Noura A. Semary  251
Olivier Gibaru  83
Peter Luksch  205
Saleem Abuleil  219
Sanghyeok Lim  01
Sara Al-Osimi  183
Scott DeBates  127
Shangping Ren  127
Sitalakshmi Venkatraman  175
Somaya Al Madeed  143
Stéphane Thiery  83
Taeho Cho  01
Vidhu Mitha  157
Wenhong Tian  29
Xiyu Hua  127
Xuelei Ni  97
Yong-Jae Kim  229
Zainab Awan  143