Computer Science & Information Technology 133

Natarajan Meghanathan,
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

8th International Conference on Computational Science and Engineering (CSE 2020),
December 12~13, 2020, Dubai, UAE.

AIRCC Publishing Corporation

Volume Editors

Natarajan Meghanathan,
Jackson State University, USA
E-mail: nmeghanathan@jsums.edu

Dhinaharan Nagamalai,
Wireilla Net Solutions, Australia
E-mail: dhinthia@yahoo.com

ISSN: 2231 - 5403
ISBN: 978-1-925953-31-2
DOI: 10.5121/csit.2020.101701- 10.5121/csit.2020.101706

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The 8th International Conference on Computational Science and Engineering (CSE 2020), December 12~13, 2020, Dubai, UAE, International Conference on Big Data, Machine Learning and IoT (BMLI 2020), 5th International Conference on Education (EDU 2020), 8th International Conference of Artificial Intelligence and Fuzzy Logic (AI & FL 2020) and 7th International Conference on Computer Networks & Communications (CCNET 2020) was collocated with 8th International Conference on Computational Science and Engineering (CSE 2020). 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 CSE 2020, BMLI 2020, EDU 2020, AI & FL 2020 and CCNET 2020 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.

In closing, CSE 2020, BMLI 2020, EDU 2020, AI & FL 2020 and CCNET 2020 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the CSE 2020, BMLI 2020, EDU 2020, AI & FL 2020 and CCNET 2020.

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,
Dhinaharan Nagamalai (Eds)

General Chair	Organization
Natarajan Meghanathan,	Jackson State University, USA
Dhinaharan Nagamalai,	Wireilla Net Solutions, Australia

Program Committee Members

Abdel-Badeeh M. Salem,	Ain Shams University, Egypt
Abdelhafid ZEROUAL,	University of Artois, France
Abdelkrim Khireddine,	University of Bejaia, Algeria
Abdulhamit Subasi,	Effat University, Saudi Arabia
Adrian Olaru,	University Politehnica of Bucharest, Romania
Ahmed EL Oualkadi,	Abdelmalek Essa?di University, Morocco
Ahmed Farouk AbdelGawad,	Zagazig University, Zagazig, Egypt
Ahmed Kadhim Hussein,	Babylon University, Iraq
Akhil Gupta,	Lovely Professional University, India
Alessio Ishizaka,	NEOMA Business School, France
Ali Ghasemi,	Isfahan University of Technology, Iran
Amel Boufrioua,	University of Mentouri Brothers, ALGERIA
Amina El murabet,	Abdelmalek Essaadi University, Morocco
Amir Hossein Niknamfar,	Qazvin Islamic Azad University, Iran
Amizah Malip,	University of Malaya, Malaysia
Ammar A. Aldair,	University of Basrah, Iraq
Anas M.R. Al Sobeh,	Yarmouk University, Jordan
Anazida Zainal,	Universiti Teknologi Malaysia, Malaysia
Anouar Abtoy,	Abdelmalek Essaadi University, Morocco
Antoni B. Chan,	City University of Hong Kong, Hong Kong
António Abreu,	ISEL, Portugal
Arjav Bavarva,	RK University, India
Assas Ouarda,	University of Batna 2, Algeria
Assia DJENOUHAT,	University Badji Mokhtar Annaba, Algeria
Atindra Dahal,	PhD, Nepal
Attila Kertesz,	University of Szeged, Hungary
Auxiliar,	University of Beira Interior, Portugal
Ayman A. Aly El-Naggar,	Taif University, Saudi Arabia
B.Krishna Kumar,	Anna University, India
Basanta Joshi,	Tribhuvan University, Nepal
bdullah,	Adigrat University, Ethiopia
BENYETTOU Mohammed,	Relizane University Center, Algeria
Bernard Cousin,	University of Rennes, France
Bilal H. Abed-alguni,	Yarmouk University, Jordan
Bilal Hisham Ghanem,	Technical university of valencia, spain
Binod Adhikari,	IOE,Nepal
Chandrasekar Vuppalapati,	San Jose State University, USA
Chandrashekhar Bhat,	Manipal Institute of Technology, India
Chi-Cheng Chang,	National Taiwan Normal University, Taiwan
Chitra Javali,	National University of Singapore (NUS)
Claudio Gallicchio,	University of Pisa, Italy
Claudio Schifanella,	University of Turin, Italy
D.Lakshmi Padmaja,	Anurag University, India
Dac-Nhuong Le,	Haiphong University, Vietnam

Dadmehr Rahbari,	University of Qom, Iran
Dhanya Jothimani,	Ryerson University, Canada
Ding Wang,	Nankai University, China
Domenico Ciuonzo,	University of Naples Federico II, Italy
Dongping Tian,	Baoji University of Arts and Sciences, China
Douglas Vieira,	CEO at ENACOM, Brazil
Edwin Lughofer,	Johannes Kepler University Linz, Austria
El-Sayed M. El-Horbaty,	Ain Shams University, Egypt
Elzbieta Macioszek,	Silesian University of Technology, Poland
Emad Eldin Mohamed,	Canadian University Dubai, UAE
Eng Islam Atef,	Alexandria University, Egypt
Erdal OZDOGAN,	Gazi University, Turkey
Farhan Masud,	Universiti Teknologi Malaysia, Malaysia
Farzin Piltan,	University of Ulsan, Korea
Fazle Baki,	University of Windsor, Canada
Fei HUI,	Chang'an University, P.R.China
Fei Yuan,	Ryerson University, Canada
Francesco Zirilli,	retired Sapienza Universita Roma, Italy
Francisco Martínez-Alvarez,	Pablo de Olavide University, Spain
Gayathri Devi S,	SASTRA Deemed University, India
Gazi Erkan Bostanci,	Ankara University, Turkey
Gerard Deepak,	National Institute of Technology, India
Ghassan Qas marrogy,	Cihan University, Iraq
Grigorios N. Beligiannis,	University of Patras, Greece
Grzegorz Sierpiński,	Silesian University of Technology, Poland
Hala Abukhalaf,	Palestine Polytechnic University, Palestine
Hamed Taherdoost,	Hamta Business Solution Sdn Bhd, Canada
Hamid Ali Abed AL-Asadi,	Basra University, Iraq
Hasnaoui Salem,	University Tunis El-Manar, Tunisia
Heba Elgazzar,	Morehead State University, USA
Heldon Jose,	Professor of Integrated Faculties of Patos, Brazil
HEMN BARZAN ABDALLA,	Wenzhou-Kean University, China
Hiromi Ban,	Nagaoka University of Technology, Japan
Hyunsung Kim,	Kyungil University, Korea
Ibrahim Hamzane,	Hassan II University, Morocco
Ibrahim Yakubu,	Abubakar Tafawa Balewa University, Nigeria
Ilango Velchamy,	CMR Institute of Technology, India
Ines Bayoudh Saadi,	Tunis University, Tunisia
Ishmanov Farruh,	Kwangwoon University, South Korea
Islam Atef,	Alexandria University, Egypt
Islam Tharwat Abdel Halim,	Misr International University (MIU), Egypt
Israa Shaker Tawfic,	Ministry of Science and Technology, Iraq
Jackelou S. Mapa,	Saint Joseph Institution of Technology
Javid Taheri,	Karlstad University, Sweden
Javier Gozalvez,	Universidad Miguel Hernandez de Elche, Spain
Jaymer Jayoma,	Caraga State University, Philippines
Joseph G. Davis,	University of Sydney, Australia
Jun Peng,	University Of Texas Rio Grande Valley, USA
Junath Naseer Ahamed,	Ibri College of Technology, Sultanate of Oman
Junmei Zhong,	Marchex Inc, USA
K.Suganthi,	Vellore Institute of Technology, India
Ka Chan,	University of Southern Queensland, Australia

Kamel Jemaï,	University of Gabès, Tunisia
Kavitha R,	SASTRA Deemed University, India
Ke-Lin Du,	Concordia University, Canada
Khaireddine Bacha,	University of Tunisia, Tunisia
Khalid M.O. Nahar,	Yarmouk University, Jordan
khaoula boutouhami,	southeast university, Nanjing china
Klenilmar L.Dias,	Federal Institute of Amapa-Macapa Campus
Kolla Bhanu Prakash,	Koneru Lakshmaiah Education Foundation
L. Alfonso Ureña-López,	Universidad de Jaén, Spain
LABRAOUI Nabila,	University of Tlemcen, Algeria
Lal Pratap Verma,	Moradabad Institute of Technology, Moradabad
Ling Xing,	Henan University of Science and Technology
Lorena Kodra,	Polytechnic University of Tirana, Albania
Lucian Lupu-Dima,	University of Petrosani, Romania
M. B. Mu'azu,	Ahmadu Bello University, Nigeria
M. Zakaria Kurdi,	University of Lynchburg, USA
Mabrouka Gmiden,	National Engineering School of Sfax, Tunisia
Maciej Kusy,	Rzeszow University of Technology, Poland
Maissa HAMOUDA,	SETIT & ISITCom, University of Sousse
Malka N. Halgamuge,	University of Melbourne, Australia
Manish Kumar Mishra,	University of Gondar, Ethiopia
María Hallo,	Escuela Politécnica Nacional, Ecuador
Mario Henrique Souza Pardo,	University of São Paulo, Brazil
Mario Versaci,	DICEAM - Univ. Mediterranea, Italy
Marzak Bouchra,	Hassan II University, Morocco
Medjahed Chahreddine,	University of Hassiba ben bouali chlef, Algeria
Meghna Sharma,	The NorthCap University, India
Mihai Lungu,	University of Craiova, Romania
Mohammad Jafarabad,	Qom University, Iran
Mohammad Siraj,	King Saud University, Kingdom of Saudi
Mohammed A.	Akour, Yarmouk University, Jordan
Mohammed Alchaita,	Syrian Virtual University, Syria
Mohammed Al-Mai'itah,	Al-Balqa applied university, Jordan
Morteza Alinia Ahandan,	University of Tabriz, Iran
Mostafa Ghomeishi,	University of Malaya, Malaysia
Muhammad Alhammami,	Multimedia University, Malaysia
Murtaza CİCİOĞLU,	Ministry of National Education, Turkey
Mustafa Abuzaraida,	Universiti Utara Malaysia, Malaysia
Nabila Labraoui,	University of Tlemcen, Algeria
Nadia Abd-Alsabour,	Cairo University, Egypt
Nidal M. Turab,	Al-AHliyya Amman University, Jordan
Nihar Athreyas,	Spero Devices Inc, USA
Nouh Elmitwall,	Jouf University, KSA
Odikwa Ndubuisi Henry,	Wellspring University, Nigeria
Omid Mahdi Ebadati E,	Kharazmi University, Tehran
Orhan Dagdeviren,	Ege University, Turkey
P. Bhattacharya,	Mody University Of Science & Technology
Paulo Vitor Campos Souza,	CEFET-MG and UNA Belo Horizonte, Brazil
Peter Plapper,	Université du Luxembourg, Luxembourg
Picky Butani,	Austin Energy - SRNL, USA
Prasan Kumar Sahoo,	Chang Gung University, Taiwan
Quang Hung Do,	University of Transport Technology, Vietnam

R.Arthi,	SRM Institute of Science and Technology, India
Rahul P. Deshmukh,	IIT Bombay, India
Rajdeep Chowdhury,	JIS College of Engineering, India
Rajeev Kanth,	Savonia University of Applied Sciences, Finland
Rakesh Kumar Mahendran,	Veltech Multitech Engineering College, India
Ramadan Elaiess,	University of Benghazi, Libya
Ramgopal Kashyap,	Amity University Chhattisgarh, India
Reguia Mahmoudi,	Yahia Fares University of Medea, Algeria
Richa Purohit,	D Y Patil International University, India
Rituparna Datta,	University of South Alabama, USA
Roberto De Virgilio,	Università Roma Tre, Italy
Rohola Zandie,	University of Denver, USA
Ruhaidah Samsudin,	Universiti Teknologi, Malaysia
Ruijuan Zheng,	Henan University of Science and Technology
Ruksar Fatima,	Khaja Bandanawaz University, India
S.Geetha,	VIT University, India
S.P.Vimal,	Sri Ramakrishna Engineering College, India
S.Senthil Kumar,	Universiti Sains Malaysia, Malaysia
S.SyedAmeer Abbas,	MepcoSchlenk Engineering College, India
S.Taruna,	JK Lakshmipat University, India
Sabyasachi Pramanik,	Haldia Institute of Technology, India
Sadaqat Ur Rehman,	Associated College of the University, UK
Said AGOUJIL,	University of Moulay Ismail Meknes, Morocco
Saida Bouakaz,	Claude Bernard University Lyon, France
Sajadin Sembiring,	Universitas Sumatera Utara, Indonesia
Sathyendra Bhat J,	St Joseph Engineering College, India
Satish Gajawada,	IIT Roorkee,India
Senthil Thilak,	National Institute of Technology Karnataka
Shahram Babaie,	Islamic Azad University, Iran
Siarry Patrick,	Universite Paris-Est Creteil, France
Singam Jayanthu,	National Institute of Technology, India
Smain Femmam,	UHA University, France
Solomiia Fedushko,	Lviv Polytechnic National University, Ukraine
Soodeh Hosseini,	Shahid Bahonar University of Kerman, Iran
Soo-Gil Park,	Chungbuk National University, South Korea
Stefano Michieletto,	University of Padova, Italy
Subhendu Kumar Pani,	Oec, Bput, India
Sudhakar Sengan,	Sree Sakthi Engineering College, India
Suhad Faisal Behadili,	University of Baghdad,Iraq
SUJATHA,	VIT University, India
Sun-yuan Hsieh,	National Cheng Kung University, Taiwan
Suyel Namasudra,	NIT patna, India
Sweta Srivastava,	Sharda University, India
Syed Umar Amin,	King Saud University, Saudi Arabia
Taha Ali,	Alzaim Azhari University, Sudan
Thafseela Koya Poolakkachalil,	National Institute of Technology, India
Tri Kurniawan Wijaya,	Technische Universitat Dresden, Germany

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Artificial Intelligence Community (AIC)

Soft Computing Community (SCC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

8th International Conference on Computational Science and Engineering (CSE 2020)

International Conference on Big Data, Machine Learning and IoT (BMLI 2020)

5th International Conference on Education (EDU 2020)

8th International Conference of Artificial Intelligence and Fuzzy Logic (AI & FL 2020)

7th International Conference on Computer Networks & Communications (CCNET 2020)

GENETIC ALGORITHM FOR EXAM TIMETABLING PROBLEM - A SPECIFIC CASE FOR JAPANESE UNIVERSITY FINAL PRESENTATION TIMETABLING

Jiawei LI and Tad Gonsalves

Department of Information & Communication Sciences
Faculty of Science and Technology, Sophia University, Tokyo, Japan

ABSTRACT

This paper presents a Genetic Algorithm approach to solve a specific examination timetabling problem which is common in Japanese Universities. The model is programmed in Excel VBA programming language, which can be run on the Microsoft Office Excel worksheets directly. The model uses direct chromosome representation. To satisfy hard and soft constraints, constraint-based initialization operation, constraint-based crossover operation and penalty points system are implemented. To further improve the result quality of the algorithm, this paper designed an improvement called initial population pre-training. The proposed model was tested by the real data from Sophia University, Tokyo, Japan. The model shows acceptable results, and the comparison of results proves that the initial population pre-training approach can improve the result quality.

KEYWORDS

Examination timetabling problem, Excel VBA, Direct chromosome representation, Genetic Algorithm Improvement.

1. INTRODUCTION

Examination Timetabling Problem (ETP) is a well-known NP-hard problem which tries to find the best examinations schedule for schools, colleges, and universities. As a discrete optimization algorithm, Genetic algorithm (GA) is naturally suitable to solve ETP. Moreover, compared to other search algorithms, GA is more robust in searching complex search spaces [1]. This paper focuses on a special case of Examination Timetabling Problem (ETP), which is common in Japanese universities or colleges. In Japanese universities or colleges, every final-year student must give a presentation to evaluate their academic research. Each student has three examiners, so the examination timetabling problem is related to allocating proper time and rooms to students and examiners. It is preferable to put all students from the same laboratory together in the same room for presentations in successive time slots which make a SESSION. Moreover, one session should be held on one day. These two constraints make the traditional common GA no longer effective because the traditional crossover operation can break the intact session with a high probability resulting in a large number of infeasible solutions. Therefore, one has to design a novel variant of GA to satisfy this special case.

This paper designed an automatic ETP algorithm by using a variant of GA where the constraint-based initialization and crossover operations are applied to satisfy the two constraints about the

Natarajan Meghanathan et al. (Eds): CSE, BMLI, EDU, AI & FL, CCNET - 2020
pp. 01-12, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101701

sessions. The penalty point system is also implemented to optimize other soft constraints. However, during the research, it is found that the constraint-based initialization and crossover operations usually cause premature convergence and then output an unacceptable result. To improve the result quality of the proposed model, an improvement which we call initial population pre-training is adopted.

The application is written in VBA language. VBA is embedded in the Microsoft office suite. Although VBA code execution is slower than C or Python, it has the advantage of using the familiar MS Excel worksheet interface with an unobtrusive macro embedded in it. Further, since all the timetabling data is in Excel, direct encoding in Excel VBA becomes very handy for the university staff who is responsible for creating and managing the timetable.

2. RELATED WORK

As Qu et al. [2] and Carter, & Laporte [3] summarized, the early research– on ETPs can be roughly divided into four types of approaches, which are Cluster, Sequential, Constraint Based and Generalized Search. Generalized Search has another well-known name, namely, meta-heuristic algorithms. In the early research on ETP, simulated annealing algorithm [4][5] and Tabu search [6][7] were the two main Generalized Search approaches.

In recent years, with the development of ETPs research, meta-heuristic algorithms have become some of the main approaches to solve ETPs, and many different kinds of meta-heuristic algorithms are implemented. In 2005, Azimi [8] applied Simulated Annealing (SA), Tabu Search (TS), Genetic Algorithm (GA) and Ant Colony System (ACS) to solve the ETP. They then proposed three novel hybrid combinations of those four algorithms, which are Sequential TS–ACS, Hybrid ACS/TS, and Sequential ACS–TS algorithms. By testing more than 10 different scenarios of the ETP, they demonstrated that all the three hybrid algorithms performed significantly better than the performance of four non-hybrid algorithms. In 2006, Eley [9] applied two ant colony approaches, Max-Min and ANTCOL approach for solving the ETP and compared these two approaches with other timetabling heuristics. In 2015, Mandal & Kahar [10] applied the great deluge algorithm to partial exam assignment. In their approach, the total exams are ordered in advance based on graph heuristic and then partial exams are improved by the great deluge algorithm one by one until all exams have been scheduled. Compared to the state-of-the-art approaches, this novel method shows a competitive performance. In 2018, Leite et al. [11] solved the ETP with the cellular memetic algorithm. The cellular memetic algorithm organizes the population in a cellular structure to provide a smooth actualization and improve the diversity of the population. The algorithm gives improvements on partial functions of incapacitated Toronto and capacitated ITC 2007 benchmark sets.

As one of the most popular optimization approaches, GA has also been widely used to solve ETPs. In 2017, Rozaimee, et al. [12] tried to use GA to construct the final exam timetable automatically for the UniSZA computer system, to save time for the university staff. In 2017, Shatnawi et al. [13] proposed a two-stage approach optimization algorithm by running Greedy Algorithm and GA in parallel, to help the Arab East College of High Education in Saudi Arabia solve the problem of scheduling exams. Their result shows that the required number of conflicts, exam days and available venues had been reduced successfully. In 2019, Dener [14] introduced a two-stage GA, where the first stage carries out the assignment of courses to sessions and the second stage assigns the students who participated in the test session to the examination room. The system was designed to allocate students and supervisors in a more efficient way to reduce the number of rooms and time consumption. In 2020, Gozali, et al. [15] attempted to solve the university course timetabling by using localized island GA with dual dynamic migration policy

(DM-LIMGA). The results show that the proposed algorithm can produce a feasible timetable in student sectioning problem with a better result than previous works.

3. Objective Problem

The proposed model is designed to solve the thesis presentation final exams of the Informatics Graduate School of Sophia University, Tokyo, Japan, which is different from the general Examination Timetabling Problems. At the end of the academic year, all final-year students in Science and Engineering School at Sophia University must give a presentation to explain and discuss their research. Each student's presentation is evaluated by three examiners, one of which is the student's supervisor, who is called the prime examiner, and the two other examiners are called deputy examiners, who are familiar with the student's research topic. The student's examiners are decided in advance which cannot be changed. The capacity of the room is limited. Each room can only have one presentation at a time. A room can hold a maximum of 10 presentations per day. The purpose of the proposed problem is to allocate all students to the proper timeslots and rooms, by considering the examiners' available time and some other constraints which will be explained below. The specific information about the proposed problem is shown in Table 1.1.

Table 1. Algorithm input, basic information of the proposed problem.

Presentation event dates	3 days
Presentation event start time of a day	9:00
Presentation event end time of a day	17:30
Length of presentation	40 minutes
Lunch period	12:00—13:00
Length of break	5 minutes
Room numbers of each day	2/3/3

For each timeslot, an examiner has three options, "O", "X" and "Δ". "O" means the examiner is available for that time period. An "X" means the examiner is not available for that time period. A "Δ" means the examiner is available during that time period, but that should be avoided if possible. "X" has higher priority than "Δ".

The proposed timetabling problem also has 3 hard constraints and 4 soft constraints as follows:

Hard constants

1. All students with the same supervisor should hold the presentation one after another, which is called a session.

2. A session should be held on a single day.

3. No student or examiner can be removed from the session.

Soft constraints

1. Examiners should be allocated to appropriate time depending on the examiner's own schedule.

2. Examiners should be assigned to contiguous time slots, if possible.

3. The timetable should be compact.

4. An examiner should not occur in two places at the same time. (Note that, although from the feasibility perspective this is a hard constraint, for fast convergence of our algorithm, it is classified as a soft constraint).

In the common ETP problem, each examination is independent, which means a single examination could be moved independently to wherever feasible. The common GA operations can be directly applied to solve the problem because the arbitrary chromosome cut and join does not break the feasibility of the solutions. However, in Japan, people prefer to make the students with the same supervisor together to conduct the presentation one after another, which is called a session. Moreover, a whole session should be held on the same day. These two constraints corresponds to hard constraint 1and 2. The two hard constraints make the common GA operation no longer work because the arbitrary chromosome cut and join does not break the session with a high probability, and it is hard to recombine the scattered session again by the arbitrary GA operation. To solve this problem, a constraint-based initialization and crossover operations are proposed, where these two operations will never break the three hard constraints and therefore, the three hard constraints can be satisfied automatically. Moreover, the four soft constraints are optimized by using a penalty points system. This is the reason why we classify the constraint “An examiner should not occur in two places at the same time” as a soft constraint, but not a hard constraint, although commonly people will regard this constraint as a hard constraint intuitively. In the proposed model, all hard constraints are satisfied automatically by the specific GA operations, and all soft constraints are satisfied by the penalty point systems.

4. PROPOSED VARIANT GA MODEL

4.1. Chromosome representation

The proposed model uses the direct chromosome as the encoding, where each chromosome corresponds to a specific arrangement to the generated blank position unit. The total chromosome length is equal to the total number of available timeslots. For example, there is a two-day presentation period, each day has 2 available rooms, each one can hold maximum 10 presentations each day, then the total chromosome length is equal to 2*2*10=40 timeslots. The students will be numbered from 1 to some maximum number. Each chromosome is a list of numbers which contain all numbers from 1 to the maximum number of students and numbers of 0, meaning a certain position unit does not have an arrangement for a student’s presentation.

For example, assume we have a two-day presentation event, each day has two available rooms, Room A and Room B, then the chromosome is as follows:

[5,6,7,0,0,0,0,0,0,0, 0,0,0,0,0,0,11,10,9,8, 15,16,0,1,2,3,4,12,13,14, 0,0,0,0,0,0,0,0,0,0]

In this chromosome string, the student 5 is allocated into timeslot 1, which corresponds to the first presentation in the Room A of the first day. Similarly, the student 11 is allocated into timeslot 17, which corresponds to the 7th presentation in the Room B of the first day, the student 16 is allocated into timeslot 22, which corresponds to the 2nd presentation in the Room A of the second day, while there are no students allocated into the Room B of the second days which corresponds to the timeslots from 31 to 40.

4.2. Constraint-based initialization operation

To satisfy the hard constraints 1 and 2, a constraint-based initialization operation is introduced. The operation is divided into two steps:

First, the students are grouped with the same supervisor to form a session. Then, the order of the sessions and the order of students within one session are re-ordered.

Secondly, the first session of the new order is allocated into the first available place in the first room. The second session of the new order is allocated into the first available place in the second room and so on. If all rooms have been allocated with a session, the operation goes back to the top and allocates the next session to an available timeslot of a random room. If the selected random room does not allow all students from the same laboratory to hold the presentation in the same day and same room, a new random target room is looked for. If there is no way to allocate all laboratories to the proper position, an error is reported to ask the staff to re-arrange the rooms or add some new rooms. In this way, all the initial possible solutions can satisfy the hard constraints automatically, where the students from the same session can be allocated together and a session can hold all presentations on the same day. Meanwhile the constrained initialization can maintain a certain degree of randomness to maintain the diversity of the initial population. Below is an example of an initial chromosome. In this example, each room corresponds to 10 time slots.

[6,7,5,0,0,0,0,0,0,0, 10,8,9,11,0,0,0,0,0,0, 15,16,0,0,0,0,0,0,0,0, 13,12,14,0,1,2,3,4,0,0].

4.3. Fitness evaluation and penalty system

The fitness evaluation of the proposed model uses the penalty points system to optimize the soft constraints in the proposed problem. In penalty points system, if the solution breaks any soft constraints, the penalty will be given. The proposed timetabling problem then becomes an optimization problem to find the solution with minimum penalty. Four soft constraints in the proposed problem are decomposed into 6 types of penalties. The soft constraints 1, “Examiner should be allocated to an appropriate time depending on examiner’s own schedule” corresponds to penalty 1 and penalty 2. The soft constraint 2 corresponds to penalty 3 and 4. The soft constraints 3 and 4 correspond to penalty 5 and 6, respectively. Table 1.4 shows the specific distribution of the penalty points.

Table 2. Penalty category and penalty points.

Penalty	Penalty points
1. Examiner is allocated into time period with X.	242
2. Examiner is allocated into time period with △.	60
3. Examiners are placed in contiguous slots in the same session.	10
4. Examiners are placed in contiguous slots in different sessions.	9
5. An examiner occurs in two places at the same time.	390
6. Session did not start during 1st period in one room OR, two sessions are not contiguous.	1 per timeslot

During the calculation, the penalties which have larger penalty points are more likely to be avoided during the evolution of GA. Depending on the actual situation of the priority of the penalties, each penalty is allocated with different penalty points. Since it is physically impossible for “an examiner occurring in two places at the same time”, penalty 5 is allotted the highest penalty points. The penalty 1, “examiner is allocated into timeslots with X” is then another important penalty we want to avoid, therefore it is allotted the second-highest penalty points, followed by penalties 2, 3, 4 and 6. Moreover, according to the real situation, penalties 5 and 2 are two situations which we want to avoid as far as possible, the other penalties, however, have some degree of tolerance. Therefore, the value of penalties 5 and 2 should be far greater than the value of other penalties.

4.4. Selection

Tournament selection is implemented, the tournament size is set as 2.

4.5. Crossover

In this paper, a constraint-based crossover operation is designed to satisfy the hard constraints. During the constrained crossover operation, the chromosomes have been decomposed into two parts. The first problem is the optimization of the examiner’s schedule problem, which corresponds to penalty 1 and 2 in Table 1.4. The second problem is to make sure the same examiner can attend the presentations contiguous in the same session and between two sessions to save the examiner’s time, which is called examiner’s time continuity problem, which corresponds to penalty 3, 4 and 5 in Table 1.4. Therefore, the constraint-based crossover must achieve the exploration for both, examiner’s schedule problem search region and examiner’s time continuity problem search region. In the proposed model, a variant multi-point crossover has been applied to achieve the information exchange for both examiner’s schedule problem and examiner’s time continuity problem, and meanwhile, maintain the feasibility for each chromosome.

In the first step, two parent chromosomes will be selected. Then, for each parent chromosome, two random sessions, session A and session B will be selected. Session A and session B could be either real sessions or zero sessions. A zero session means no session is to be placed in these timeslots. Then session A on the first parent chromosome and session B on the second parent chromosome will be swapped to get two new chromosomes. Similarly, session B on the first parent chromosome and session A on the second parent chromosome will be swapped. Since the different sessions could have different number of students, once a session with fewer students is swapped with a session with more students, the algorithm will check if there is enough blank space beside the shorter session to fill the position of longer session. If not, the model will then check if it is possible to move the adjacent session up or down to vacate enough space. However, the moving of the session should maintain this session within one single room and one single day. Secondly, if enough space cannot be vacated by moving the adjacent session up or down, the adjacent session will be moved to another random place, where this random place should also satisfy the constraints to maintain the session within one single room and one single day. In this way, the exchange of gene segments could be achieved, and meanwhile, the integrity of each session can be maintained and ensured that each session is held on the same day.

For example, there is a problem with 5 sessions, [1,2,3,4],[5,6,7],[8,9,10,11],[12,13,14] and [15,16], and there are two parent chromosomes,

[15,16,0,0,0,0,0,0,0,0, 9,8,10,11,0,0,0,0,0,0, 13,12,14,0,0,0,0,0,0,0, 5,6,7,0,1,2,3,4,0,0] and

[0,0,0,0,0,0,0,0,0,0, 8,9,10,11,0,1,2,3,4,0, 0,0,00,0,0,0,0,0,0, 7,6,5,0,0,12,13,14,15,16]

Assuming that session [8-11] and [15, 16] are selected. The session [8-11] from parent chromosome 1 and session [15, 16] from parent chromosome 2 will be swapped. Similarly, the session [15, 16] from parent chromosome 1 and session [8-11] from parent chromosome 2 will be swapped as well. However, since the length of session [15, 16] is shorter than that of session [8-11] and in parent chromosome 2, there are not enough blank timeslots beside the session [15, 16]. Therefore, the session [12-14] in the parent chromosome 2 will be moved two timeslots forward to vacate enough position for session [8-11]. And the offspring chromosomes after the crossover are:

[8,9,10,11,0,0,0,0,0,0, 15,16,0,0,0,0,0,0,0,0, 13,12,14,0,0,0,0,0,0,0, 5,6,7,0,1,2,3,4,0,0]

[0,0,0,0,0,0,0,0,0,0, 15,16,0,0,0,1,2,3,4,0, 0,0,00,0,0,0,0,0,0, 7,6,5,12,13,14,9,8,10,11]

Moreover, if the session [12-14] in the parent chromosome 2 could not vacate enough timeslot for the session [8-11] then the session [12,13,14] in the parent chromosome 2 will be moved to another random feasible position. This operation is to guarantee every session would have the same opportunity to be exchanged.

However, compared to the traditional crossover, the amount of information exchanged by this constraint-based crossover is relatively limited, where only two sessions of information can be exchanged during each crossover operation. Moreover, as we mentioned in section 4.3, the constraints with high priority are given much higher penalty points. The limited information exchange and uneven proportion of penalty points make the algorithm usually be stuck into the local optima solution with a high penalty point. To improve the result quality, an improvement approach called initial population pre-training is applied to the proposed algorithm, which will be described in section 4.8.

4.6. Mutation

During the constrained mutation operation, one session in the chromosome is selected first and then two random students in this session are swapped.

4.7. Elitism

The research of [16] [17] shows that the elitism of GA can improve the convergence speed. However, the increased number of remained elitism individuals can increase the evolution pressure which may cause premature convergence [18]. To preserve the diversity of the population, in the proposed model, only the first best solution in the proposed model can be retained to the next iteration.

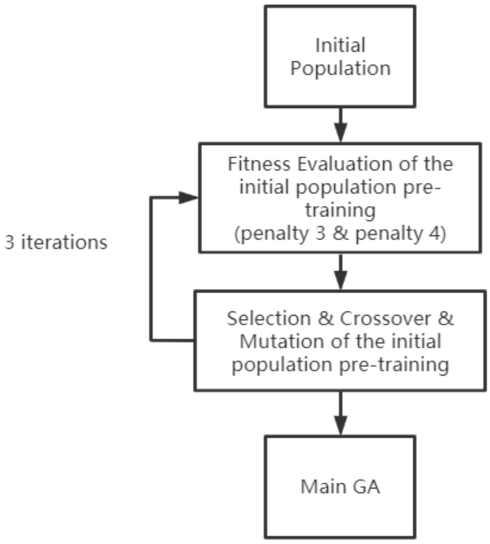
4.8. Initial population pre-training

As we mentioned in crossover part, the proposed problem can be decomposed into two optimization problems: examiner's schedule problem and examiner's time continuity problem. However, it is usually hard for the proposed algorithm to both the objective problems at the same time. If relatively larger penalty points are applied to the examiners' schedule problem, the examiners' time continuity problem will be stuck in a local optimum at a higher opportunity. Similarly, larger penalty points on the examiner's time continuity problem could cause a bad result on examiners' schedule problem. In the proposed model, depending on the priority of the penalties in the real situation, examiners' schedule problem is given relatively higher penalty points. Moreover, in the proposed algorithm, the constraint-based crossover makes a concession on search ability to satisfy the hard constraints. Therefore, this constraint-based crossover operation is not able to make enough information exchange and cannot keep the diversity of the population. Therefore, during the research of the variant GA algorithm, it is found that the algorithm is usually stuck into the local optimal solution with high penalty points on penalty 3 and penalty 4. For the sake of keeping balance between the examiners' schedule problem and the examiners' time continuity problem, and to further improve the diversity of the population, we propose an improvement approach, namely, initial population pre-training which has been proved to be effective in enhancing the result quality.

In 1993, Schoenauer and Xanthakis [19] introduced a genetic optimization based on the Behaviour Memory Paradigm. The method first only considers only one constraint; when sufficient number of feasible individuals satisfy this constraint, the algorithm will then consider next constraint and eventually, all constraints can be satisfied.

The initial population pre-training operation in the proposed model refer to this idea. The initial population pre-training operation is conducted on the population before the main iterations of GA. During pre-training operation, the populations are evaluated by only the examiner's time continuity problem for serval iterations. In this way, some good solutions on the examiners' time continuity problem can be generated in advance to reduce the search pressure on examiner's time continuity problem. However, the iteration number of pre-training cannot be so high to avoid the homogeneity of the initial population. In the proposed model, the pre-training operation runs only 3 iterations. The test result shows the initial population pre-training can improve the result quality. Figure 1 shows the flowchart of the initial population pre-training operation.

Figure 1. Flowchart of the initial population pre-training operation


Initial Population
Fitness Evaluation of the initial population pre-training (penalty 3 & penalty 4)
3 iterations
Selection & Crossover & Mutation of the initial population pre-training
Main GA

4.9. Flowchart of the whole model

Figure 2 shows the flowchart of the whole proposed model.

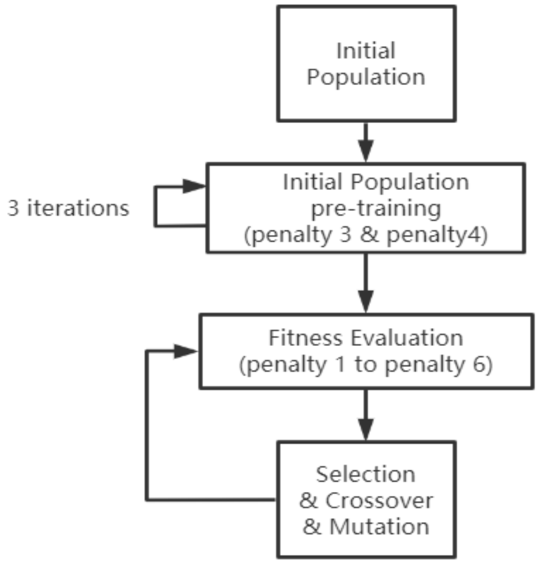

Initial Population
Initial Population pre-training (penalty 3 & penalty4)
3 iterations
Fitness Evaluation (penalty 1 to penalty 6)
Selection & Crossover & Mutation

Figure 2. Flowchart of the whole proposed model

5. TESTING

The test uses the real data from Sophia University, 2020 Winter Presentation event for Informatics Graduate School, which contains 31 students and 25 examiners. Below, two different variants of the proposed model are tested, where model 1 is the proposed model and model 2 is the control group. Each model is tested 10 times to get an average value. The total population size for each model is equal to 120. The initial population pre-training runs 3 iterations. The toleration for the stop condition is 30 iterations. The crossover ratio is set as 0.7, and the mutation ratio is set as 0.1.

Model 1: Proposed GA with initial population pre-training

Model 2: Proposed GA only

Table 3 and Table 4 show the testing result of model 1 and model 2 respectively, where the penalty 1 to penalty 6 in the tables means the quantity of penalty imposed on the solution for constraint violations.

For model 1, the average penalty points are 116.8, which breaks down to an average 0.1 times for penalty 1, 5.0 times for penalty 3, 4.2 times for penalty 4 and 4.8 times for penalty 6. Examiners can be allocated into their available time for most of the time. The model shows an acceptable performance as verified by the university staff.

The comparison between model 1 and model 2 shows that, by adding the pre-training, the average penalty points of penalty 1 slightly decrease from 0.2 to 0.1, the average penalty points of penalty 2 decreases from 0.8 to 0.0, and the average penalty points of penalty 3 decreases from 7.1 to 5. On the contrary, the average penalty points of penalty 5 slightly increase from 4.0 to 4.2, the average penalty points of penalty 6 increase from 3.4 to 4.8. However, penalty 4 and penalty

6 have the lowest priority in the problem, therefore, we think the increase of the penalty points on penalty 4 and penalty 6 are acceptable. In total, model 1 obtained lower total average penalty points than that of model 2, which is 116.8 compared to 206.8. The result shows that the adding of initial population pre-training can improve the result quality. Moreover, the average calculation iteration of model 1 is 104.1 plus 3 iterations of initial population pre-training, compared to the 103.7 average iterations of model 2, the calculation speed of the improved model does not decrease so much.

Table 3. Test result for Model 1.

test	**penalty points**	**iteration**	**penalty 1**	**penalty 2**	**penalty 3**	**penalty 4**	**penalty 5**	**penalty 6**
1	135	59	0	0	6	8	0	3
2	352	91	1	0	7	4	0	4
3	92	127	0	0	5	4	0	6
4	124	117	0	0	7	5	0	9
5	91	82	0	0	4	5	0	6
6	70	90	0	0	3	4	0	4
7	61	54	0	0	5	1	0	2
8	97	182	0	0	4	6	0	3
9	72	195	0	0	4	3	0	5
10	74	44	0	0	5	2	0	6
Average	**116.8**	**104.1**	**0.1**	**0**	**5**	**4.2**	**0**	**4.8**

Table 4. Test result for Model 2.

test	**penalty points**	**iteration**	**penalty 1**	**penalty 2**	**penalty 3**	**penalty 4**	**penalty 5**	**penalty 6**
1	160	162	0	1	9	1	0	1
2	180	55	0	1	10	2	0	2
3	198	54	0	1	9	5	0	3
4	167	96	0	1	5	6	0	3
5	225	67	0	1	8	9	0	4
6	740	50	2	2	8	5	0	11
7	131	119	0	1	5	2	0	3
8	68	153	0	0	3	4	0	2
9	107	188	0	0	7	4	0	1
10	92	93	0	0	7	2	0	4
Average	**206.8**	**103.7**	**0.2**	**0.8**	**7.1**	**4**	**0**	**3.4**

6. CONCLUSION

This paper focuses on a specific examination timetabling problem which commonly occurs in Japanese universities and proposed a variant genetic algorithm approach to solve the exam timetabling problem. The proposed model is written in VBA programming language, which is easy for the university staff to use. Constraint-based initialization and crossover operations are used to satisfy the hard constraints, and a penalty system is implemented to optimize the soft constraints. To improve the result quality, the initial population pre-training, which optimizes partial objective problem in advance for several iterations, is applied. The proposed model is compared with the model without pre-training, by using the real exam data from Informatics Graduate school from Sophia University in 2020. The positive comparison results support the idea that the initial population pre-training is an effective approach to improve the result quality of the proposed model.

However, to compromise on the feasibility of the problem constraints, the constraint-based crossover operation cannot reach wide-enough search space. Therefore, the solution often got stuck into the local-optima solution. Even though the initial population pre-training can solve this problem in a way, the research on a better crossover approach for this specific examination timetabling problem is still needed. In the proposed case, the discrete search space of ETP is further scattered because of the hard constraints 1 and 2. The local search approaches have the advantage of exploitation, while the population-based approaches have the stronger ability of exploration. Therefore, applying other types of population-based algorithms is one direction of further research. Another research direction is improving the exploration ability and the result quality of the GA. A lot of researches show that parallel GA can improve the diversity of the population to give a better result compared to the normal GA [20] [21]. Therefore, as for the future direction of this research, we plan to use the GA island model for faster convergence and better results.

REFERENCES

[1] Song, Y., Wang, F., & Chen, X. (2019). An improved genetic algorithm for numerical function optimization. Applied Intelligence, 49(5), 1880-1902.

[2] Qu, R., Burke, E. K., McCollum, B., Merlot, L. T., & Lee, S. Y. (2009). A survey of search methodologies and automated system development for examination timetabling. Journal of scheduling, 12(1), 55-89.

[3] Carter, M. W., & Laporte, G. (1995, August). Recent developments in practical examination timetabling. In International Conference on the Practice and Theory of Automated Timetabling (pp. 1-21). Springer, Berlin, Heidelberg.

[4] Thompson, J. M., & Dowsland, K. A. (1996). Variants of simulated annealing for the examination timetabling problem. Annals of Operations research, 63(1), 105-128.

[5] Thompson, J. M., & Dowsland, K. A. (1998). A robust simulated annealing based examination timetabling system. Computers & Operations Research, 25(7-8), 637-648.

[6] Di Gaspero, L., & Schaerf, A. (2000, August). Tabu search techniques for examination timetabling. In International Conference on the Practice and Theory of Automated Timetabling (pp. 104-117). Springer, Berlin, Heidelberg.

[7] Abdullah, S., Ahmadi, S., Burke, E. K., Dror, M., & McCollum, B. (2007). A tabu-based large neighbourhood search methodology for the capacitated examination timetabling problem. Journal of the Operational Research Society, 58(11), 1494-1502.

[8] Azimi, Z. N. (2005). Hybrid heuristics for examination timetabling problem. Applied Mathematics and Computation, 163(2), 705-733.

[9] Eley, M. (2006, August). Ant algorithms for the exam timetabling problem. In International Conference on the Practice and Theory of Automated Timetabling (pp. 364-382). Springer, Berlin, Heidelberg.

[10] Mandal, A. K., & Kahar, M. N. M. (2015, April). Solving examination timetabling problem using partial exam assignment with great deluge algorithm. In 2015 International Conference on Computer, Communications, and Control Technology (I4CT) (pp. 530-534). IEEE.

[11] Leite, N., Fernandes, C. M., Melicio, F., & Rosa, A. C. (2018). A cellular memetic algorithm for the examination timetabling problem. Computers & Operations Research, 94, 118-138.

[12] Rozaimee, A., Shafee, A. N., Hadi, N. A. A., & Mohamed, M. A. (2017). A framework for university’s final exam timetable allocation using genetic algorithm. World Applied Sciences Journal, 35(7), 1210-1215.

[13] Shatnawi, A., Fraiwan, M., & Al-Qahtani, H. S. (2017, February). Exam scheduling: A case study. In 2017 Ninth International Conference on Advanced Computational Intelligence (ICACI) (pp. 137-142). IEEE.

[14] Dener, M., & Calp, M. H. (2019). Solving the exam scheduling problems in central exams with genetic algorithms. arXiv preprint arXiv:1902.01360.

[15] Gozali, A. A., Kurniawan, B., Weng, W., & Fujimura, S. (2020). Solving university course timetabling problem using localized island model genetic algorithm with dual dynamic migration policy. IEEJ Transactions on Electrical and Electronic Engineering, 15(3), 389-400.

[16] E. Zitzler, K. Deb, and L. Thiele, “Comparison of multiobjective evolutionary algorithms: Empirical results,” Evol. Comput., vol. 8, no. 2, pp. 173–195, Summer 2000.

[17] G. Rudolph, “Evolutionary search under partially ordered sets,” Dept. Comput. Sci./LS11, Univ. Dortmund, Dortmund, Germany, Tech. Rep. CI-67/99, 1999.

[18] Ahn, C. W., & Ramakrishna, R. S. (2003). Elitism-based compact genetic algorithms. IEEE Transactions on Evolutionary Computation, 7(4), 367-385.

[19] M. Schoenauer and S. Xanthakis, “Constrained GA optimization,” in Proc. Int. Conf. Genetic Algorithms, vol. 5, 1993, pp. 573–580.

[20] Gozali, A. A., & Fujimura, S. (2019). DM-LIMGA: dual migration localized island model genetic algorithm—a better diversity preserver island model. Evolutionary Intelligence, 12(4), 527-539.

[21] Sun, X., Chou, P., Wu, C. C., & Chen, L. R. (2019). Quality-oriented study on mapping island model genetic algorithm onto CUDA GPU. Symmetry, 11(3), 318.

AUTHOR

Jiawei Li is a PhD student in the Department of Information & Communication Sciences. Faculty of Science and Technology, Sophia University, Tokyo, Japan. His research interests are Evolutionary Computation and Neuro-Evolution.

A STUDY INTO MATH DOCUMENT CLASSIFICATION USING DEEP LEARNING

Fatimah Alshamari[1, 2] and Abdou Youssef [1]

[1]Department of Computer Science,
The George Washington University, Washington D.C, USA
[2]Department of Computer Science, Taibah University, Medina, KSA

ABSTRACT

Document classification is a fundamental task for many applications, including document annotation, document understanding, and knowledge discovery. This is especially true in STEM fields where the growth rate of scientific publications is exponential, and where the need for document processing and understanding is essential to technological advancement. Classifying a new publication into a specific domain based on the content of the document is an expensive process in terms of cost and time. Therefore, there is a high demand for a reliable document classification system. In this paper, we focus on classification of mathematics documents, which consist of English text and mathematics formulas and symbols. The paper addresses two key questions. The first question is whether math-document classification performance is impacted by math expressions and symbols, either alone or in conjunction with the text contents of documents. Our investigations show that Text-Only embedding produces better classification results. The second question we address is the optimization of a deep learning (DL) model, the LSTM combined with one dimension CNN, for math document classification. We examine the model with several input representations, key design parameters and decision choices, and choices of the best input representation for math documents classification.

KEYWORDS

Math, document, classification, deep learning, LSTM

1. INTRODUCTION

Recently, online libraries have been increasing exponentially in the number of publications [1]. These documents are required to be manually annotated by an expert in order to be indexed in online libraries. The annotation, if done manually, is an expensive process in terms of cost and time. For this reason, some libraries ask the authors to help with this process by adding a subject code before submitting the document [1]. However, this does not appear to solve the problem adequately, because the final subject for indexing a document should be reviewed by an expert. Accordingly, many studies have been conducted in automated text classification to tackle the problem. Most of these studies aim to find a reliable classification approach that could replace the expert role [2].

Thus, different classification approaches have been investigated in text classification across several areas. These approaches could include traditional machine learning algorithms or deep learning classification models. However, mathematics documents differ from documents in other areas because the former contain mathematical symbols, formulas, and expressions. Inasmuch as they are indispensable for a reader to understand the mathematics content, using these mathematical expressions as features in any classifiers model could improve the classification

Natarajan Meghanathan et al. (Eds): CSE, BMLI, EDU, AI & FL, CCNET - 2020
pp. 13-22, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101702

process. In document classification, features are crucial as they can improve the accuracy and efficiency of the classifier or hurt them. In this study, we investigated the impact of math features either alone or in conjunction with the text contents of documents on the math document classification performance. Therefore, we conducted comparative experiments by examining a number of models, several key design parameters, and decision choices to choose the best deep learning math documents classifier design.

This paper is organized as follows. Section 2 presents an overview of related work. Section 3 discusses the dataset we used and its different set-ups. Section 4 illustrates our approach. In Section 5, we describe our experiments, and discuss our results. Our conclusions and future work are provided in Section 6.

2. RELATED WORK

In the field of document classification, many studies have been conducted in different areas aiming to find a reliable model [2], which include classical machine learning and deep neural networks models.

Harish et al presented a general technical review that studies the text classification and conducted a performance comparison of a set of known machine learning classification models [3]. Their study concludes that under different criteria, such as algorithm parameters and time complexity, each model behaves differently. There is no optimum combination between text representation techniques with machine learning models that could produce the best result.

In a recent study, Scharpf, et al investigated the impact of selecting and combining different representation techniques on the classification accuracy and cluster purity on scientific documents, sections and abstracts [4]. They applied different encodings techniques such as tfidf, and doc2vec on text and math content separately. Their study showed that text encoding outperforms standalone math encoding, and combining text and math encodings does not improve the classification performance. They observed a low correlation between text and math similarity, which suggests that text and math should better be treated as separate features.

Dalal and Zaveri [2] presented some issues with text classification that could affect the performance of any model. The pre-proccessing and feature selection techniques should be applied carefully because they have a critical impact on performance.

Our work is aligned with [5], [6], and [1], where they focused on mathematics document classification. Barthel et al presented a large-scale experiment by applying simple pre-proccessing. Although they used a large dataset, about three million documents, they limited the experiments to metadata only. Despite the fact they presented good results, it would not be possible to classify papers that have short or no abstract using their model. In addition, if a formula is too simple, during the pre-processing stage, they added it to the only-text features; this enrichment of the only text features could have some impact on performance that was not investigated in that paper. In our paper, we are curious to comprise a large set of formula features beyond what is revealed in the metadata only. Therefore, we adopt the full-text dataset including the metadata.

In a recent paper [6], Suzuki and Fujii investigated math document classification, but they used MATHML that produces the math features in a tree structure, whereas we use LLaMaPUn [7], and our math features are linear rather than tree-structured.

In terms of deep neural networks [8], [9] and [10] studied different deep neural network models on the text classification task. Semberecki and Maciejewski applied long short-term memory (LSTM) model in documents classification to study different representations approaches [8]. Their evaluation showed that the vector representation approach outperformed a standard bag-of-word approach based on the LSTM model in the document classification task.In [9] a recurrent Convolutional Neural Network (CNN) model was proposed for text classification. They applied a recurrent structure to capture the contextual information in order to construct the representation of the text. In addition they applied max-pooling to capture the main tokens in the text. Their experiments were conducted on four different datasets, and showed that their proposed model outperformed CNN and Recursive Neural Networks (RNN).A hybrid architecture consisting of an LSTM model and CNN model was proposed for different text classification tasks in [10]. Their experiments showed that combining these two models in one architecture outperforms both CNN and LSTM alone.

3. DATASET

3.1. Dataset Collection

In order to evaluate our models, we obtained one of the HTML5 dataset subsets [11]. It is a large scale dataset recently published by the arXiveLive project [12]. It consists of over a million scientific documents from the arXiv.org [13], converted from the original LaTeX format to HTML format using the LaTeXML conversion tool [14], then grouped based on the conversion success level into three subsets. Since not all the math documents converted fully (without errors and warnings) using LaTexML, we selected only the documents that had no conversion issues. This resulted in a smaller subset, but large enough, which we used for this work. This subset consists of 705,095 multi-label documents assigned to one or more subjects from seven major areas: Physics, Computer Science, Mathematics, Quantitative Biology, Quantitative Finance, Statistics, Electrical Engineering and Systems Science, and Economics. For the Mathematics area, ArXiV uses a set of 32 subjects to label math documents inclusively. The first subject that is assigned to a document reflects the principal contribution of that document, and all other subjects, if any, are considered a secondary principal contribution.

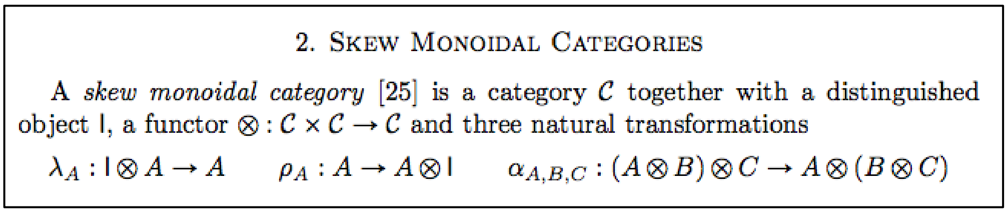
Our focus in this study is document classification in the Mathematics areas; therefore, from the used subset we retrieved a collection of about 171,000 full-text English-language documents with math subject as a first label. The number of documents in each subject varies from 372 to 20522. The popularity of math paper submissions varies from subject to subject, causing a bias between the numbers of documents in the different subject areas. Therefore, we will consider weighted accuracy in future work. In addition, the size or length of the documents varies; some documents are short papers while others are dissertations with a significantly large number of pages. For this reason, the input length is one of the parameters that we considered in the training process. Table 1 provides detailed information regarding the subject distribution in the dataset.

Table 1. Dataset Statistics based on the number of documents in each subject.

Subject	#of Document	Subject	#of Document	Subject	#of Document
mathGT	3075	mathST	3155	mathCO	14116
mathGR	4902	mathMP	20627	mathCT	327
mathKT	416	mathAG	10577	mathDS	7895
mathMG	1335	mathAT	1697	mathGM	819
mathNT	7698	mathAP	18145	mathHO	480
mathNA	4869	mathCA	6947	mathIT	6413
mathOA	2076	mathAC	4370	mathLO	2317
mathOC	4321	mathCV	4819	mathPR	11523
mathQA	3040	mathDG	9724	mathRT	1817
mathRA	2703	mathFA	8470	mathSG	542
mathSP	786	mathGN	1164		
Total # of Documents		171165			

3.2. Dataset Preparation

One of the advantages of the HTML format is preserving all the properties of formulas and mathematical expressions. However, NLP tools require text data; therefore, an additional step of converting the HTML format has been done. In order to get the raw text, we applied the LLaMaPUn [7] toolkit. The raw text is segmented into three different sets: (1) Text and Math, (2) Text, and (3) Math. The Text-and-Math set is the original output of the LLaMaPUn, Figure 1 shows the input for this set. While the Text set consists of a clean text without any math formulas or expressions, as shown in Figure 2, the Math set consists of only math formulas and expressions, as shown in Figure 3.

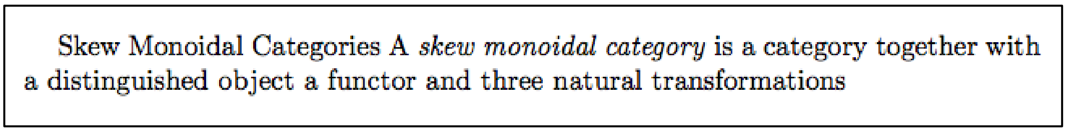


2. Skew Monoidal Categories

A *skew monoidal category* [25] is a category $\mathcal{C}$ together with a distinguished object I, a functor $\otimes : \mathcal{C} \times \mathcal{C} \to \mathcal{C}$ and three natural transformations

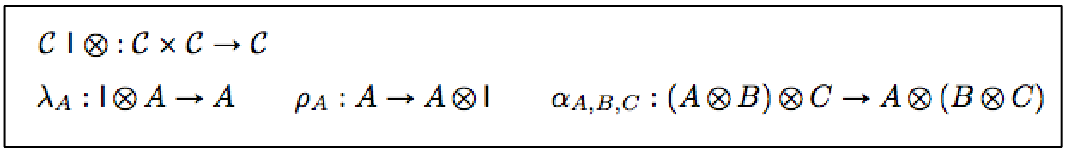
$$\lambda_A : \mathsf{I} \otimes A \to A \qquad \rho_A : A \to A \otimes \mathsf{I} \qquad \alpha_{A,B,C} : (A \otimes B) \otimes C \to A \otimes (B \otimes C)$$

Figure 1. The input for Text and Math set



Skew Monoidal Categories A *skew monoidal category* is a category together with a distinguished object a functor and three natural transformations

Figure 2. The input for Text set



$$\mathcal{C}\ \mathsf{I}\ \otimes : \mathcal{C} \times \mathcal{C} \to \mathcal{C}$$

$$\lambda_A : \mathsf{I} \otimes A \to A \qquad \rho_A : A \to A \otimes \mathsf{I} \qquad \alpha_{A,B,C} : (A \otimes B) \otimes C \to A \otimes (B \otimes C)$$

Figure 3. The input for Math set

3.3. Dataset Pre-processing

Data pre-processing is an essential step for any text-based task. For each set resulting from the segmentation step, we applied different pre-proccessing. For the Text set, we applied the following processing: text normalization and text removal. Specifically, in the text normalization step, we replace all punctuation with a keyword PUN, and all numbers with keyword NUM. In the text removal step, we used the Natural Language Toolkit (NLTK) [15], to omit stop-words since they carry less significance in classification [2]. Also, we discarded all non-English words and words with length more than 25. In addition, as math documents contain theories' names and other named entities, which are case-sensitive, we preserved case-sensitivity. For the Math set, math formulas and expressions are case-sensitive, hence, we used LLaMaPUn to preserve case-sensitivity, font styles, weights and faces [16]. Finally, a combination of the set-specific (i.e. Text set and Math set) pre-processing steps has been applied on the Text-and-Math set.

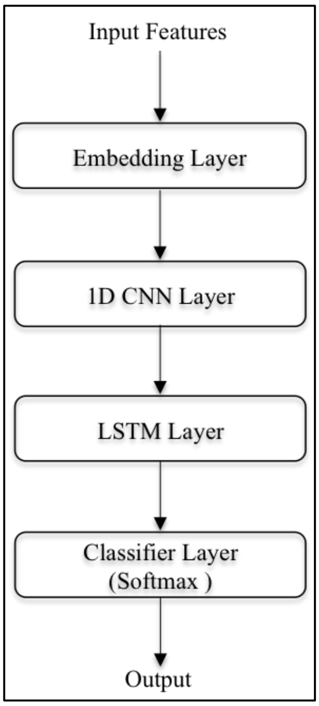


Figure 4. The Approach for Classifying Mathematics Document

4. THE DEEP LEARNING APPROACH

Our approach consists of (1) examining a deep learning model architecture with several key design parameters; (2) evaluating the model on document classification, and (3) investigating the best feature set based on the performance. Recently, deep learning models have shown significant improvements in different NLP tasks such as Question Answering [17], Information Extraction [18], and Text Classification [8], [9] and [10]. Different architectures including RNN [19] and its variants LSTM [20], BiLSTM [21], CNN [22], and Attention [23] have been widely studied for different Neural Network Language Processing (NLP) tasks.

While there is no rule that defines which model is the right one for a specific task, the task objective and the input characteristics play an important role in narrowing the set of models candidates. For example, the RNN has two limitations which are exploding gradient and vanishing gradient [24]. The vanishing gradient could accrue with a long sequence input, which is the case in document classification. Fortunately, the LSTM model tackles the gradient problems and is able to capture long-term dependencies in any length sequences. In addition, combining CNN with LSTM in one architecture outperforms LSTM model in text classification [10].

Therefore, our focus in this paper is to study LSTM with one CNN layer on the math documents classification with different setups. Our model, shown in Figure 4, consists of four layers: an embedding layer, a 1D CNN layer, an LSTM layer, and a softmax layer. The model starts with feeding the document as raw input features into the embedding layer, which passes its embeddings to the CNN layer; the CNN layer derives important features and passes them to the LSTM layer [24]. Finally, the output of the LSTM layer is fed to the softmax layer that determines the class of the input document.

As we are dealing with math documents which contain text with math formulas and expressions, we define three different sets of input representations: Text-Only and Math-Only, and Text-Math as detailed in section 2.2.

5. EXPERIMENTAL RESULTS AND ANALYSIS

5.1. Experiments

Our model is built using the deep learning platform Keras [25] with Tensorflow [26] as the backend framework. In order to train and evaluate the classifier, we divided the dataset randomly into 80\% training and 20\% testing split. For all setups we report and compare the accuracy on the test set. The categorical-crossentropy is the loss function used, and for optimizing all models Adam is used [27]. All Out-Of-Vocabulary (OOV) words are randomly initialized.

The main parameters we focused on are the documents length and the word embeddings. We examined different input length (1000, 1500, 2000), where an input length of *N* means that from each document, only the first *N* tokens are taken while the remaining tokens are ignored. This is commonly done in deep learning for memory and speed considerations. In the embedding setups, we evaluated our model based on the word embeddings types using two main methods: Randomly Initialized embeddings, and GloVe embeddings. For the GloVe embeddings we have experimented with the type of the data used to train the embeddings, and the size of embeddings dimensions. Specifically, we used pre-trained embeddings from [11] which are trained on global data in different subject areas including math; we refer to the pre-trained GloVe embedding as GloVe-PR. GloVe-PR came in one dimension size 300 and two flavors based on its training data: Text only excluding math symbols and expressions, and text-and-math which includes both the text and the math symbols and expressions.

In addition to the pre-trained embeddings, we trained our own GloVe on our dataset, which includes math subjects only. We follow similar setups and parameters used in [11] with additional dimension sizes in (100, 200, 300); we refer to this mode of embedding as GloVe-M. For all pre-trained embedding experiments, GloVe-PR and GloVe-M we evaluate both tuning the word embeddings and freezing the word embeddings layer. For the baseline, we trained the LSTM model using randomly initialized embeddings.

5.2. Results

We report in Tables 2, 3 and 4 the performance of the models based on the different setups. Table 2 illustrates the accuracy performance corresponding to embedding dimension of 100 in all models and all different input lengths. In this table Text-Only outperforms both the Text-Math and Math-Only, where the highest results across all models are given by GloVe-M+Tuned model. We also observe that increasing the input length provides some improvement, though the gains are very modest.The results of 200 embedding dimensions are shown in Table 3. Overall, we observed similar trends to the ones we observed in Table 2, except for the input length impact on the results. In particular, adding more input data resulted in sustainable improvement in the performance.

Table 2. Classification Experimental Results of the 100 Embedding Dimension. The numbers represent the classification accuracy on the test set.

Input length	Model	Text-Math	Text-Only	Math-Only
1000	Baseline	57.4	64.08	37.7
	Random-Embedding	56.78	63.68	39.42
	GloVe-M	52.27	60.77	22.1
	GloVe-M+Tunned	57.98	**65.1**	35.43
1500	Baseline	59	65.94	38.94
	Random-Embedding	58.07	65.91	41.53
	GloVe-M	54.33	62.68	12.74
	GloVe-M+Tunned	58.5	**67.08**	37.04
2000	Baseline	59.43	65.86	39.85
	Random-Embedding	59.2	66.33	42.59
	GloVe-M	54.37	63.73	23.84
	GloVe-M+Tunned	60.19	**67.87**	38.18

Table 3. Classification Experimental Results of the 200 Embedding Dimension. The numbers represent the classification accuracy on the test set.

Input length	Model	Text-Math	Text-Only	Math-Only
1000	Baseline	58.07	64.55	37.38
	Random-Embedding	56.8	64.4	38.68
	GloVe-M	52.78	71.74	10.45
	GloVe-M+Tunned	58.41	**65.38**	34.76
1500	Baseline	58.87	65.87	38.93
	Random-Embedding	58.14	65.9	41.34
	GloVe-M	55.31	63.49	26.65
	GloVe-M+Tunned	59.37	**66.41**	36.89
2000	Baseline	59.82	66.2	46.71
	Random-Embedding	59	67.21	43.65
	GloVe-M	56.7	**68.07**	20.48
	GloVe-M+Tunned	60.33	68.04	38.56

Table 4 shows the results using 300 dimension embedding with different input lengths. In general, the Text-Only yields the best performance compared to both the Math-Only and Text-and-Math across all setups. Specifically, in the GloVe-PR+Tuned model with the 2000 input length, Text-Only outperforms the Math-Only and Text-Math with 24.13\% and 7.7\% gain respectively. For embedding types, both GloVe-PR+Tuned and GloVe-M+Tuned produced the best results across different input length.

Table 4. Classification Experimental Results of the 300 Embedding Dimension. The numbers represent the classification accuracy on the test set.

Input length	Model	Text-Math	Text-Only	Math-Only
1000	Baseline	58.04	64.6	36.71
	Random-Embedding	57.15	63.5	40.04
	GloVe-PR	52.8	62.36	39
	GloVePR+Tunned	58.48	**65.52**	40.22
	GloVe-M	54.03	62.18	26.52
	GloVe-M+Tunned	58.29	65.1	31.11
1500	Baseline	59.31	65.94	38.61
	Random-Embedding	58.48	66.08	41.58
	GloVe-PR	55.81	63.67	40.38
	GloVePR+Tunned	59.56	67.06	41.67
	GloVe-M	55.84	63.66	28.79
	GloVe-M+Tunned	59.73	**67.08**	31.17
2000	Baseline	60.12	65.86	39.91
	Random-Embedding	55.95	66.24	42.96
	GloVe-PR	56.94	65.37	42.62
	GloVePR+Tunned	60.54	**68.24**	44.11
	GloVe-M	56.32	64.39	14.42
	GloVe-M+Tunned	60.24	67.87	37.63

Interestingly, the result shows that the classifiers performance degraded when including the math formulas and expressions. The best Text-Only model performed significantly better than the other models that contain math information. Thus, considerable improvement can be achieved when omitting all the math information from the text.

We believe the reason that the Text-Only model performs better is twofold. First, unlike text tokens, the math symbols in equations and expressions are abstract and often generic, thus lacking any differentiation power. Second, when we include math tokens and fix the number of tokens taken to represent a document, the number of text tokens in the representation in the Text-and-Math model is fewer than in the Text-Only model, thus ending up with fewer differentiating features in the Text-and-Math model than in the Text-Only model, resulting naturally in lesser accuracy.

With respect to the input length and embeddings size, all models seem to benefit from increasing both factors. However, this improvement is slight with increasing the embeddings size as shown in Table 5.

Table 5. Best Performance in Text Only among different Dimensions and Input-Length.

Model	Dimension	Input-Length	Accuracy
GloVePR+Tunned	300	2000	**68.24**
GloVe-M	200	2000	**68.07**
GloVe-M+Tunned	100	2000	**67.87**

6. CONCLUSIONS

Mathematics document classification is an important problem to address due to the growing demand for automatic models to handle such task. Document classification in general and the Mathematics document classification in particular are still an open challenge given the large set of subjects associated with a document.

Thus, in this study, we carried out comparative experiments using well-known deep learning models, LSTM with one dimension CNN. One main goal of this study is to investigate the impact of inclusion/exclusion of the math features on the classification output. The other main goal was to identify optimal models, optimal feature representations, and optimal design choices.

The results showed that overall the model performance improved with excluding all math information from the document, using the Text-only representations. Furthermore, the result of testing the impact of input length and embeddings size showed that increasing both factors impacted the performance positively. Specifically, larger improvement was achieved with increasing the input length and smaller improvement achieved with increasing the embedding size.

Of all the models and variations in parameters and design choices considered, the best choice was determined to be: 300D embedding with input length of 2000, and with post-trained (i.e. tuned) GloVe embedding that was pretrained on global data.

This study can be improved in several ways. For instance, in our future work, we are planning to use the full text from a document for further experiments and testing. We also plan to investigate other embedding representations such as Word2Vec/Fasttext and contextualized embeddings (i.e. ELMO and BERT), and test other deep neural architectures.

REFERENCES

[1] Řehůřek, Radim, & PetrSojka, (2008) "Automated classification and categorization of mathematical knowledge", International Conference on Intelligent Computer Mathematics, pp543-557.

[2] Dalal, Mita K., & Mukesh A. Zaveri, (2011) "Automatic text classification: a technical review", International Journal of Computer Applications, Vol. 10, No. 5, pp37-40

[3] Harish, Bhat S., Devanur S. Guru, & ShantharamuManjunath,(2010) "Representation and classification of text documents: A brief review", IJCA, Special Issue on RTIPPR (2), pp110-119.

[4] P. Scharpf, M. Schubotz, A. Youssef, F. Hamborg, N. Meuschke, andB.Gipp,(2020) "Classification and Clustering of arXiv Documents, Sections, and Abstracts, Com-paring Encodings of Natural and Mathematical Language", Proceedings of the JCDL Conference 2020, doi: 10.1145/3383583.3398529.

[5] Barthel, Simon, SaschaTönnies, & Wolf-TiloBalke, (2013) "Large-Scale Experiments for Mathematical Document Classification", International Conference on Asian Digital Libraries, pp83-92.

[6] Suzuki, Tokinori, & Atsushi Fujii, (2017) "Mathematical document categorization with structure of mathematical expressions", ACM/IEEE Joint Conference on Digital Libraries, pp1-10.

[7] DeyanGinev& Jan Frederik Schaefer, (2019) "LLaMaPUn: common language and mathematics processing algorithms", https://github.com/dginev/llamapun/

[8] Semberecki, Piotr & HenrykMaciejewski, (2017) "Deep learning methods for subject text classification of articles", Federated Conference on Computer Science and Information Systems, pp1357-360.

[9] Lai, Siwei, et al., (2015) "Recurrent convolutional neural networks for text classification", Twenty-ninth AAAI conference on artificial intelligence

[10] Zhou, Chunting, et al., (2015) "A C-LSTM neural network for text classification", arXiv:1511.08630

[11] DeyanGinev, (2018) "an HTML5 conversion of arXiv.org", https://sigmathling.kwarc.info/resources/arxmliv-dataset-082018/

[12] Stamerjohanns, Heinrich, Michael Kohlhase, DeyanGinev, Catalin David, & Bruce Miller, (2010) "Transforming large collections of scientific publications to XML", Mathematics in Computer Science,Vol. 3, No. 3, pp299-307

[13] Arxiv E-print archive, http://arxiv.org

[14] Miller, Bruce, (2010) "LaTeXML: ALatex to xml converter",http://dlmf. nist. gov/LaTeXML/

[15] Loper, Edward, & Steven Bird,(2002) "NLTK: the natural language toolkit",arXiv preprint cs/0205028

[16] Ginev, Deyan, & Bruce R. Miller, (2019) "Scientific Statement Classification over arXiv.org.",arXiv:1908.10993
[17] Yu, Lei, et al., (2014) "Deep learning for answer sentence selection.",arXiv:1412.1632
[18] Narasimhan, Karthik, Adam Yala, & Regina Barzilay, (2016) "Improving information extraction by acquiring external evidence with reinforcement learning", arXiv:1603.07954
[19] Rumelhart, David E., Geoffrey E. Hinton, & Ronald J. Williams, (1986) "Learning representations by back-propagating errors", nature Vol. 323, No. 6088, pp533-536
[20] Hochreiter, Sepp, & Jürgen Schmidhuber, (1997) "Long short-term memory", Neural computation, Vol. 9, No. 8, pp1735-1780
[21] Schuster, Mike, & Kuldip K. Paliwal, (1997) "Bidirectional recurrent neural networks", IEEETransactions on Signal Processing, Vol. 45, No. 11, pp2673-2681
[22] Fukushima, Kunihiko, & Sei Miyake, (1982) "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition", In Competition and cooperation in neural nets, Berlin, Heidelberg, pp267-285.
[23] Wang, Yequan, et al., (2016) "Attention-based LSTM for aspect-level sentiment classification", Empirical methods in natural language processing, pp606-615.
[24] Hallac, Ibrahim R., Betul Ay, & GalipAydin, (2018) "Experiments on Fine Tuning Deep Learning Models With News Data For Tweet Classification", IEEE International Conference on Artificial Intelligence and Data Processing, pp1-5.
[25] Franc¸ois Cholletet al., (2015) "Keras: The python deep learning library", https://keras.io
[26] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... &Ghemawat, S, (2016) "Tensorflow: Large-scale machine learning on heterogeneous distributed systems", arXiv:1603.04467
[27] Kingma, Diederik P., and Jimmy Ba., (2014) "Adam: A method for stochastic optimization", arXiv:1412.6980

AUTHORS

Fatimah Alshamari is a Ph.D research student. She received B.C.Sc (Bachelor of Computer Science) degree in 2007, and M.C.Sc (Master of Computer Science) degree in 2014. She is now Assistant Lecturer of Taibah University. Her research interests include Natural Language Processing and Mathematics Language Processing.

Abdou Youssef has 30 years of research and teaching experience in the field of computer science. He is currently a tenured Professor at The George Washington University, Washington, D.C, which he joined as Assistant Professor in Fall of 1987. His current research interests are applied data science, math search and math language processing, audio-visual data processing, pattern recognition, theory and algorithms. He has published over 125 papers in those areas, and co-edited the book Interconnection Networks for High-Performance Parallel Computers, published by IEEE Computer Society Press in 1994. His research has been funded by NSF, NSA, and NIST. Currently, he is developing novel techniques for part-of-math tagging, math semantics extraction and question answering, and big-data applications such as fraud detection in the retail business, next-generation recommendation systems, and more.

AN EXAMINATION OF RELATIONSHIP BETWEEN CAREER MATURITY AND MULTIPLE FACTORS BY FEATURE SELECTION

Shuxing Zhang[1] and Qinneng Xu[2]

[1]Shenzhen College of International Education, Shenzhen, China
[2]Shenzhen Liangyi Information Technology Co., Ltd, Shenzhen, China

ABSTRACT

The purpose of this study is to investigate the relationship between career maturity and a branch of factors among senior school students. The sample data were collected from a total of 189 students. The linear relationship between career maturity and 72 factors were tested by using feature selection methods. LASSO and forward stepwise were compared based on cross-validation. The results showed that LASSO was a feasible method to select the significant factors, and 12 of the total 72 factors were found to be important in predicting career maturity.

KEYWORDS

Career maturity, Feature selection, LASSO, Stepwise.

1. INTRODUCTION

Since the 20th century, career development has been considered as a crucial topic. An individual's career development can be promoted by accelerating maturity, ability, skills, talent, and interests [1]; this viewpoint is supported by previous studies [2, 3]. Emphasis of this kind of promotion in education is necessary in order to enhance career development, and there is research suggesting that career preparation should be carried out early at school [4]. Career guidance at school is important in helping students to develop the knowledge and skills needed for making appropriate choices, managing transition in learning and moving into the workplace [5]. An early career development is also very significant for students, especially for those at high school, in the consideration that they are going to face the significant choice for colleges and major, which will have great effect on their lifetime career direction [6].

The concept of vocational maturity was first introduced by Super in 1955 [7], and then slowly replaced by the phrase "career maturity" [8], defined as the level of progress on career development tasks of an individual [9]. The concept of career maturity emphasizes that career is a continuous and dynamic process, and it can represent the degree of preparation an individual has, to face the tasks in career development [10]. In this case, career maturity can be a standard to measure how well a student has idea about his/her future career direction.

In previous studies, the relationship between career maturity and other influential factors has been widely investigated. Park [11] used correlation analysis and hierarchical regression to investigate the relationships between proactive personality, career indecision, and career maturity. Emerson [12] analyze the association between career maturity and decision-making self-efficacy based on correlation analysis and linear regression. Rahim et al. [13] focused on the

Natarajan Meghanathan et al. (Eds): CSE, BMLI, EDU, AI & FL, CCNET - 2020
pp. 23-28, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101703

affect of school type on students' career maturity via ANOVA, Spearman's correlation and t-test. MANOVA is the main approach used in the research on the effect of career education module and gender on career maturity by Talib and other scholars [1]. In Katoch's study about career maturity of secondary school students, hypothesis testings, means, standard deviation and t-tests are mainly used to find the patterns.

However, these researches only focus on one or two factors, and the existing methodology may not be appropriate in the case where there are a large amount of factors. This research gap motivates our study, and we investigate the relationships between career maturity and a series of factors in the students' daily lives which may have potential effects on their career maturity. In this study, we collect data from senior high school students in Shenzhen, and analyze the 48 potential indicators. Several feature selection methods are employed, and we obtain 12 significant factors which may be related to career maturity.

2. METHODS

2.1. Study Design

In our study, 189 students from secondary schools in Shenzhen and Gansu were invited to participate in our survey. The survey included two parts. In the first parts, we evaluated their career maturity based on their responses to the survey designed by Liu [10]. In the second part, we collected students' answers to other questions that may influence career maturity used in previous researches, including estimations of students' career exploration, big five personality, family cohesion and adaptability etc.[14,15].For the questions including several choices, we normalized the score ranging from 0 (i.e., strongly disagree) to 5 (i.e., strongly agree). For those true-false items, their answers were turned into binary variables. In general, this part contains 40 quantity factors and 32 binary variables.

Following the previous studies [12, 16] that career maturity has linear relationships with several factors., we also used the linear regression to investigate the relationships between career maturity and other factors. However, since there were 72 variables and 189 observations, overfitting could not be avoided. In this study, we implemented two feature selection methods, forward stepwise regression and least absolute shrinkage and selection operator (LASSO), to identify the most significant factors. In order to compare the performance of the two methods, we used a 10-fold cross validation to evaluate their prediction accuracy and root mean square error (RMSE) was the evaluation criterion. We then used the better feature selection method to fit the linear model to investigate the relationship between career maturity and those factors. Though the statistical analysis of the results, we could find the significant factors as well as how they influenced the career maturity.

2.2. Statistical Models

2.2.1. Forward Stepwise Regression

Stepwise regression[17] with forward selection approach is a way to carry out linear regressions and prevent overfitting at the same time. It starts with an empty model, tests the improvements of the model after addition of each variable and adds the variable that its inclusion gives the most statistically significant improvement of the fit until there is no variables that can be added.

2.2.2. LASSO

LASSO, or L1 norm penalized method [18], is one of the penalized regression methods, which is a kind of linear regression method that can prevent overfitting with measures to carry out variable selection. It carries out both variable selection and regularization in order to improve the accuracy of the model.

3. RESULTS

The RMSE of cross validation of linear regression, forward stepwise regression, and LASSO were 12.5, 12.4, and 11.7, respectively. Therefore, we applied LASSO to fit the whole data, and Table 1 shows the standardized regression coefficient (β) and the lower bound (0.025) and upper bound (0.975) of the 95% confidence interval of the coefficient for each predictor variables with non-zero coefficient in LASSO. Based on the p-value of the F test, 12 of the total 72 predictor variables were found to be significant to predict career maturity.

Table 1. Regression analysis for LASSO. Only the predictors with non-zero coefficients were listed

Predictor	***β***	**0.025**	**0.975**	**Predictor**	***β***	**0.025**	**0.975**
#2	-1.8738**	-3.552	-0.196	#26	-1.0551	-2.768	0.658
#3	1.0245	-0.663	2.712	#27	0.9840	-1.006	2.974
#4	-0.4258	-1.933	1.081	#28	-0.6012	-2.350	1.148
#5	-1.8816**	-3.757	-0.006	#29	1.1968	-0.232	2.626
#6	-1.1212*	-2.448	0.206	#30	0.7832	-0.331	1.898
#7	-0.8366	-2.541	0.868	#31	1.1671	-0.525	2.859
#8	0.0399	-1.304	1.384	#32	0.4635	-0.872	1.799
#9	-0.2375	-1.578	1.103	#33	0.3065	-1.510	2.123
#10	-1.5136**	-2.679	-0.348	#34	1.6729**	0.005	3.341
#11	0.8357	-0.998	2.669	#36	0.3898	-0.953	1.732
#12	0.2722	-1.349	1.894	#37	-0.0141	-1.457	1.429
#14	-0.5752	-1.845	0.695	#38	1.4364*	-0.078	2.951
#15	-0.4410	-1.903	1.021	#39	0.8561	-0.951	2.663
#16	0.9304	-0.279	2.139	#40	-1.2466	-2.986	0.493
#17	-1.5808*	-3.218	0.056	#41	6.2343**	1.255	11.214
#18	-1.1705	-2.876	0.535	#43	-13.7584**	-20.869	-6.648
#19	0.8706	-0.644	2.385	#45	8.9205**	3.496	14.345
#20	-0.7723	-3.349	1.804	#46	7.0938**	3.265	10.923
#21	-0.9461	-2.600	0.708	#48	-4.7518	-11.033	1.529
#22	-0.4540	-1.929	1.021	#51	-5.5741**	-9.479	-1.670
#23	0.5553	-0.853	1.964	#70	-3.3662	-8.715	1.983
#25	-1.0740	-2.427	0.279	#72	4.4005	-1.065	9.866

Note. The predictors were numbered from 1 to 72. **p<.05, *p<.1

4. DISCUSSION

In this section, we discussed the 12 important factors obtained by LASSO, and analyzed how they impacted the career maturity among students.

Factor 1: Educational level of fathers

The educational level of students' fathers was very important to predict their career maturity. Without accepting a undergraduate or graduate level of education, fathers are unlikely to provide a comprehensive information of all the careers, especially for the elite jobs. So it is reasonable that the career maturity of a student who has a father with low educational level is relatively low.

Factor 2: Frequency of mothers' companion

Similar with Factor 1, the frequency of mothers' companion was also strongly related to students' career maturity. The more frequently a student stays with his/her mother, the more potentially he/she can acquire enough experienced guidance of the career.

Factor 3: Academic performance at school

The students who achieve higher grade point average (GPA) were investigated to have higher career maturity. In most cases, GPA is related with the ability to learn, the students with better learning ability is believed to have higher career maturity.

Factor 4: Level of involvement of social activities (i.e. extracurricular clubs, student union, religious groups etc.)

According to the results, the level of involvement of social activities is not positively related with career maturity. One possible explanation for this phenomenon can be that social activities is not necessarily helpful for gaining career maturity. Although leadership and group-cooperation skills can be trained during group activities, these might be not as significantly related to career maturity.

Factor 5: The level of agreement of the statement that GPA and the efforts are highly correlated

This factor is significant to predict the career maturity, but whether they are positively or negatively related is not confirmed as the confidence interval of the coefficient can be either positive or negative. This might be explained by the fact that students can misunderstand their efforts since it is too subjective.

Factor 6: The level of agreement of the statement that tough experience in lives and studying can gain the courage when facing difficulties

The result demonstrates that students who strongly agree that they have more courage when facing difficulties after experiencing hardships tend to have high career maturity. The students who strongly agree with the statement are more likely to gain experiences after going through some difficulties in the past. These characteristics of self learning and self reflection can help gain career maturity.

Factor 7: The level of agreement of the statement that I will insist my opinion even if others are opposed to it

Inferred from the results, the level of agreement of this statement is likely to have positive relationship with career maturity. Students who strongly agree with this statement are usually insisting on their decisions and opinions of future career. It is reasonable that those who easily change their opinion after other people's persuasion can appear fluid, and it is more difficult for them to really make a firm decision on career.

Factor 8 - 11: Frequently used channels for obtaining information

These 4 factors are binary variables, and generate from the same question about the students' frequently used channel for gathering information. The results of the four factors show that if students frequently search for information via publications (i.e. books, newspaper, magazines), portal sites (i.e. Sina, NetEase, Sohu, Tencent), or search engine (i.e. Baidu, Google), they tend to own higher career maturity. However, if students always gain knowledge through person-to-person communication, their career maturity seems to be low. The results show that objective channels are more reliable than subjective channels for students to acquire information and knowledge.

Factor 12: Whether news is an important way to acquire information

This factor is a binary variable, and the result demonstrates that students tend to have a lower career maturity when they have a strong need for daily news, including politics, economics, sports, and entertainments. A possible explanation for this result is that some fields of the news are irrelevant to foster career maturity.

5. Conclusion

In this study, we developed a reasonable method to evaluate the relationship of career maturity with a branch of potential factors. The results show that LASSO is a reliable model to select essential factors. We also found that several factors which were not studied in previous studies were important to predict career maturity. However, this model has some limitations: first, the model only considered the linear relationship and non-linear relationship could also exist between career maturity and certain factors; second, the design of questionnaire might have deficiency and more factors should be included; third, the sample size is too small and more students in all the grades from different regions should participate. Those are all further directions in this research field.

References

[1] Jasmi, Talib, A. , Amla, Salleh, Salleh, & Amat, et al. (2015). Effect of career education module on career development of community college students. International Journal for Educational & Vocational Guidance.

[2] Ballout, H. I. (2009). Career commitment and career success: Moderating role of self-efficacy. Career Development International, 14, 655–670.

[3] Roaten, G. (2004). The effects of a career development intervention on the career decision making skills of high school students. Unpublished Doctoral Dissertation, Texas A&M University, College Station, TX.

[4] Rohany, N. (2008). Career counseling, shift from konvensionalisme, Prime Professor Lecture Series UKM. Bangi: Penerbit Universiti Kebangsaan Malaysia.

[5] Guide on Life Planning Education and Career Guidance for Secondary Schools, 1st edition, Career Guidance Section, School Development Division, Education Bureau(May 2014)

[6] Does the College Major You Choose Affect Your Career Potential? https://www.moneycrashers.com/college-major-career-potential/. Accessed August 9, 2020.

[7] Donald E. Super. (2005). Dimensions and measurements of vocational maturity. Teachers College Record.

[8] Crites, J. O. . (1975). A comprehensive model of career development in early adulthood. Journal of Vocational Behavior, 9(1), 105-118.

[9] Patton, W. , & Creed, P. A. . (2011). Developmental issues in career maturity and career decision status. Career Development Quarterly, 49.

[10] Linhui, L. , & Jiajia, L. . (2011). Current situation and relationship study on career anchor and career maturity of college students. Science of Social Psychology.

[11] In-Jo, P. . (2015). The role of affect spin in the relationships between proactive personality, career indecision, and career maturity. Frontiers in Psychology.

[12] Allen, Dara Ware. (2008). Career maturity and college persistence: a longitudinal study of first-year students. Dissertations & Theses - Gradworks.

[13] Abdul Rahim, Nor Syazila and Mohd Noah, Sidek and Wan Jaafar, Wan Marzuki (2018) Career maturity among students from three different types of school. International Journal of Education, Psychology and Counseling, 3 (7). pp. 8-17. ISSN 0128-164X

[14] Nadya A. Fouad, & Timothy J. Keeley. (2011). The relationship between attitudinal and behavioral aspects of career maturity. The Career Development Quarterly.

[15] Mcauliffe, G. J. . (1992). Assessing and changing career decision-making self-efficacy expectations. Journal of Career Development,19(1), 25-36.

[16] Houle, J. L. W. , & Kluck, A. S. . (2015). An examination of the relationship between athletic identity and career maturity in student-athletes. Journal of Clinical Sport Psychology, 1538(1), 119-128.

[17] Efroymson,M. A. (1960) "Multiple regression analysis," Mathematical Methods for Digital Computers, Ralston A. and Wilf,H. S., (eds.), Wiley, New York.

[18] Bowles, M. . (2015). Penalized Linear Regression. Machine Learning in Python®. John Wiley & Sons, Ltd.

AUTHORS

Shuxing Zhang is a high school student in Grade 12, and her research interests are data mining and education. While having a background in Python programming and being the 3rd author in two published Python algorithm books, she is also concerned about the problem of enhancing Chinese high school students' career maturity.

Qinneng Xu got his Phd degree in the School of Data Science, City University of Hong Kong, Hong Kong SAR, China. His research interests are data mining and bioinformatics, and he has published several papers in the fields of infectious diseases and public health.

How to Engage Followers: Classifying Fashion Brands According to their Instagram Profiles, Posts and Comments

Stefanie Scholz[1] and Christian Winkler[2]

[1]Department of Social Economy,
Wilhem Loehe University of Applied Sciences, Fuerth, Germany
[2]datanizing GmbH, Schwarzenbruck, Germany

Abstract

In this article we show how fashion brands communicate with their follower on Instagram. We use a continuously update dataset of 68 brands, more than 300,000 posts and more than 40,000,000 comments.

Starting with descriptive statistics, we uncover different behavior and success of the various brands. It turns out that there are patterns specific to luxury, mass-market and sportswear brands. Posting volume is extremely brand dependent as is the number of comments and the engagement of the community.

Having understood the statistics, we turn to machine learning techniques to measure the response of the community via comments. Topic models help us understand the structure of their respective community and uncover insights regarding the response to campaigns.

Having up-to-date content is essential for this kind of analysis, as the market is highly volatile. Furthermore, automatic data analysis is crucial to measure the success of campaigns and adjust them accordingly for maximum effect.

Keywords

Instagram, Fashion Brands, Data Extraction, Marketing, Analysis, Artificial Intelligence, Netnography, Descriptive Statistics, Visualization, Community Engagement, Artificial Intelligence, Unsupervised Learning, Topic Modelling

1. Introduction

Social media provide us with valuable information about users' behaviour, not only regarding (prospective) customers, but also in terms of brands' or organizations' online activities. Understanding the relationships between actions taken by companies and the corresponding reactions of the community is crucial for marketers to efficiently influence users' behaviour in an intended way.

Data is readily available and often consists of very large corpora –it is now decisive for market research to take advantage of this big data by handling it via AI-based approaches. However, care has to be taken on data quality. Often, AI projects fail because the underlying data is biased,

Natarajan Meghanathan et al. (Eds): CSE, BMLI, EDU, AI & FL, CCNET - 2020
pp. 29-50, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101704

incomplete or not recent. It is therefore crucial to use statistics first to analyse the (meta) data of the corpus before starting with AI-based analysis. Additionally, interesting insights can be derived from the statistical information itself.

So far, marketers over the last decade increasingly used a netnographic approach [1]. By using quantitative methods, netnography supports qualitative analysis and navigates the researcher through the large corpus of data. Netnography is acknowledged as a useful research tool for collecting and analyzing online customer information [2], [3], [4]. Originally developed as a response to customers' increasing internet use [2], netnography is based on an ethnographic research approach to studying and understanding consumption-related aspects of customers' lives online [5]. In today's environment of digitalization, netnography is more relevant than ever before [6], [7], [8].

Now it is the time to merge the advantages of netnography and AI-based analyses to gain valuable and valid insights in large UGC datasets.

As the main content in netnography is text, text analysis is in focus of computational support [9]. In this paper we focus on topic modelling, sentiment analyses, text complexity based on deep learning models.

Instagram is one of the most popular social networks in the world [10]. Particularly among young people, there is no other platform which attracts as many daily active users (DAU) as Instagram. Moreover, most of the content on Instagram is fashion-related with fashion being the most popular hashtag [11] with over 850 million posts matching the corresponding hashtag. Thus, it has also become one of the premier marketing platforms for fashion brands. On that platform brands and users post and share outfits that are then commented and voted on by other users, thus also serving as an inspiration for the brands themselves. Furthermore, Instagram has strong visual components that connect perfectly with fashion brands, such as images, videos, boomerangs, layout, hyperlapse, stories, live [12] and since 2020 also reels; therefore, Roncha and Radclyffe-Thomas [13] argue that it is ideal for companies to market their products. In our study, we focus on top fashion brands as their marketing budgets tend to be highest and they work as a first mover in the market. However, all insights are directly applicable to other fashion brands and with some transfer also to other brands in common.

However, communication on Instagram is not exclusively outbound. Users can also send feedback via "likes", comments, or shares and can follow their favourite profiles and brands, which is ideal for brands to generate engagement with users [14]. Furthermore, this offers excellent measurement capabilities for the brands to quantify the success of their marketing campaigns.

There are different definitions of digital engagement; most of them have in common that it presumes some form of active online behaviour, which is characterized by high personal involvement with the content, organization, brand, or cause presented in an online public space like Instagram [15], [16], [17].

2. DATA COLLECTION

2.1. Identifying Top Fashion Brands

We identified 68 fashion brands and analysed those brands using Netnography methods. Our list was combined by using several sources from the Internet [18]. We analysed the corresponding

Instagram profiles with respect to followers, popularity etc. Afterwards, we categorized the brands according to luxury brands, mass-market brands and sportswear brands. As we will see later, this turns out to be an interesting classification as the different classes have a distinctively diferent behaviour.

For each fashion brand, Table 1 shows the number of followers, the number of profiles followed by those brands, their number of posts, and their corresponding fashion category. The data about posts and followers included profile activities like posts and comment for each brand since their respectively first post on Instagram until October 2020. At first sight, it is striking how much the number of profiles followed by the brands differs. Carhartt follows 2,273 other Instagram users compared to Balenciaga which does not follow any. Obviously, community engagement works entirely differently (or at least is managed in very different ways).

Table 1: Profile names and statistics for the top fashion brands

Brand profile name	**Number of followers**	**Number of profiles followed**	**Number of posts**	**Category**
adidas	25,925,972	202	799	Sportswear
American Eagle	3,626,506	1,502	1,507	Mass-market
Arc'teryx	868,659	681	2,706	Sportswear
Emporio Armani	16,824,301	8	7,356	Luxury
ASICS	895,053	459	2,563	Sportswear
Balenciaga	11,563,751	0	1,895	Luxury
A BATHING APE®	4,760,152	37	4,442	Mass-market
Bershka	9,327,305	24	3,548	Mass-market
Billabong	2,100,467	723	4,299	Sportswear
Bottega Veneta	2,445,828	0	186	Luxury
BVLGARI	9,416,411	27	3,633	Luxury
Burberry	17,588,966	2	5,250	Luxury
Calvin Klein	21,010,077	387	5,806	Luxury
Carhartt	863,448	2,273	1,376	Sportswear
Cartier Official	9,996,947	2	2,072	Luxury
Coach	4,845,644	380	5,219	Luxury
Comme des Garcons	2,122,301	0	446	Luxury
Converse	10,183,541	18	472	Sportswear
Dior Official	32,885,690	259	7,060	Luxury
Salvatore Ferragamo	5,456,683	105	5,206	Luxury
Gap	3,028,782	66	221	Mass-market
Gucci Official	41,396,760	249	7,205	Luxury
Gymshark	4,866,198	129	3,589	Sportswear
Hermès	10,477,323	0	2,013	Luxury
H&M	36,114,187	497	5,869	Mass-market
HUGO	1,827,704	191	2,322	Luxury
hurley	1,998,128	440	3,249	Sportswear
Jordan	21,246,134	134	307	Sportswear
Levi's	7,258,373	1,700	3,245	Mass-market
Louis Vuitton	39,925,873	5	4,630	Luxury
lululemon	3,447,659	175	3,004	Sportswear
Valentino	13,983,770	2	8,221	Luxury
Massimo Dutti	2,628,668	280	4,388	Mass-market
Michael Kors	15,992,659	301	5,456	Luxury
Moncler	3,277,777	52	3,486	Mass-market
Nautica	582,411	622	3,023	Sportswear
Nike	122,358,832	137	768	Sportswear

Off-White™	10,832,684	35	7,163	Mass-market
oldnavy	2,420,428	691	3,793	Sportswear
OMEGA	2,992,910	45	1,948	Luxury
On	518,764	1,057	1,471	Sportswear
PALACE	1,715,894	432	1,731	Mass-market
Patagonia	4,594,390	638	2,689	Sportswear
Prada	24,536,567	6	5,968	Luxury
Primark	8,670,829	802	12,408	Mass-market
PUMA	11,541,062	9	3,007	Sportswear
Ralph Lauren	12,061,735	7	5,453	Sportswear
Ray-Ban	5,281,356	78	102	Mass-market
Reebok	2,481,959	732	2,657	Sportswear
Reef	626,987	248	4,002	Sportswear
ROLEX	11,164,320	168	1,021	Luxury
SKECHERS USA, Inc.	544,567	320	975	Sportswear
Stüssy	4,032,604	1	2,596	Mass-market
Supreme	13,896,829	113	1,573	Sportswear
Swatch	1,279,593	36	1,011	Mass-market
TAG Heuer	2,488,827	136	5,442	Luxury
The North Face	4,714,665	413	348	Sportswear
Tiffany & Co.	11,636,286	175	1,591	Luxury
Timberland	3,034,485	64	3,562	Sportswear
Tommy Hilfiger	13,692,548	299	2,834	Luxury
Under Armour	8,086,500	433	3,227	Sportswear
UNIQLO Global	2,361,115	409	1,975	Mass-market
vans	17,851,394	299	3,991	Sportswear
VEJA	521,242	776	4,398	Mass-market
Victoria's Secret	68,756,526	249	1,223	Mass-market
Volcom	1,486,942	409	1,267	Sportswear
SAINT LAURENT	8,226,590	1	1,446	Luxury
ZARA Official	41,237,203	57	3,137	Mass-market

2.2. Collect Post Data Via Instaloader

To analyse metadata and content of posts and comments, we need to get the data. Unfortunately, Instagram does not offer a public API for that, maybe due to the data scandal of its parent company Facebook in 2018 [19]. Therefore, we had to find different ways how to acquire the content.

We use a customized version of Instaloader [20] for downloading the Instagram posts. Unfortunately, Instaloader stopped working on cloud servers as Instagram blocks the corresponding IP addresses.

To overcome this limitation, we set up an infrastructure of distributed proxy servers using Squid [21] on Raspberry PIs [22] which is distributed among different home networks. For easier handling, they connect to a central OpenVPN [23] hub which hosts another Squid proxy working as client with the Raspberry PIs as parents.

This distributed infrastructure has proved to work very well. New exit nodes can be setup very easily and connect automatically to the VPN extending the capacity of the network.

2.3. Collect Profile Data Via HTML Profile Page Download

Unfortunately, Instaloader was not able to download all profile information which we needed. Therefore, we implemented a proprietary solution for downloading the Instagram profiles. It works by downloading the profile page and parsing the JSON contained on the page itself. This makes the process stable against frequently occurring layout changes by Instagram.

As the profile pages suffer from the same limitation as the post pages, the data is collected via our private proxy network.

2.4. Update Regularly

The data set is updated regularly. Profiles are downloaded daily to have an always up-to-date number of followers and follow information. This information is kept to have a history and allow reports regarding successful marketing campaigns.

Posts and comments are downloaded weekly going back four weeks since the download date. This means that comments which go to older posts will not be downloaded to save Instagram requests.

2.5. Data Formats

All data is saved in its original format. The HTML profile pages are kept for safety reasons (format might have changed or additional information might be available). The posts, comments etc. are saved in an intermediate JSON [24].

After each successful download, data is transferred to an SQLite [25] database for easier access.

2.6. Data Volume

Our data was collected over a large time interval. The first post was created at Feb 9, 2011 (shortly after the start of Instagram) by Michael Kors [26]. The latest post was from Oct 20, 2020; in total we downloaded 312,131 posts over this time interval.

Apart from posts, we also collected the corresponding comments which in total amount to 41,399,076.

3. ANALYSIS PROCESS

3.1. Clean Data

Bots are a very common phenomenon in Instagram. Bots are mainly used for creating visibility in their own account or promoting third party accounts.

For most of our analysis, bots are not really relevant as their impact vanishes statistically. However, for some KPIs like seconds to first comment, it is essential to remove them.

For this, we searched for all profiles which have ever posted comments in less than 60 seconds after the post itself. Additionally, we calculated the frequency of these events and the average number of seconds it took them to write a comment. Combining these values, we could eliminate almost all of the bots.

3.2. Data Analysis

We performed our data analysis in Jupyter notebooks [27] using the Python [28] programming language and several libraries like Pandas [29], SciPy [30] and Scikit Learn [31]. These are standard libraries which have been tested extensively.

The Jupyter notebook connects directly to the SQLite database.

All analysis run automatically and generate (SVG) diagrams as artefacts. This is essential in our process as data changes continuously. We will keep an updated version of the diagrams on Github.

Our process completely avoids Excel for intermediate data and therefore is completely reproducible and safe against regressions.

4. RESULTS

4.1. Descriptive Data Analyses

To start the exploration of the big data set the following analyses are descriptive focusing on the number of posts in total, comments and average number of posts per day. In order to clearly visualize differences between fashion categories (see Table 1), each category was assigned to a colour.

4.1.1. Number of Posts and Average Comments

As shown in Figure 1 Primark has by far the most posts. As shown in Figure 1, other mass market-brands are evenly distributed regarding the number of posts (mean value = 3469, standard deviation = 2920). Whereas most of the luxury brands are located in the upper field of posts (mean value = 3980, standard deviation = 2333). Concerning sportswear brands the number of posts has an average of 2289 (standard deviation = 1275), which is significantly lower ($p = 0.03$) than the two other categories.

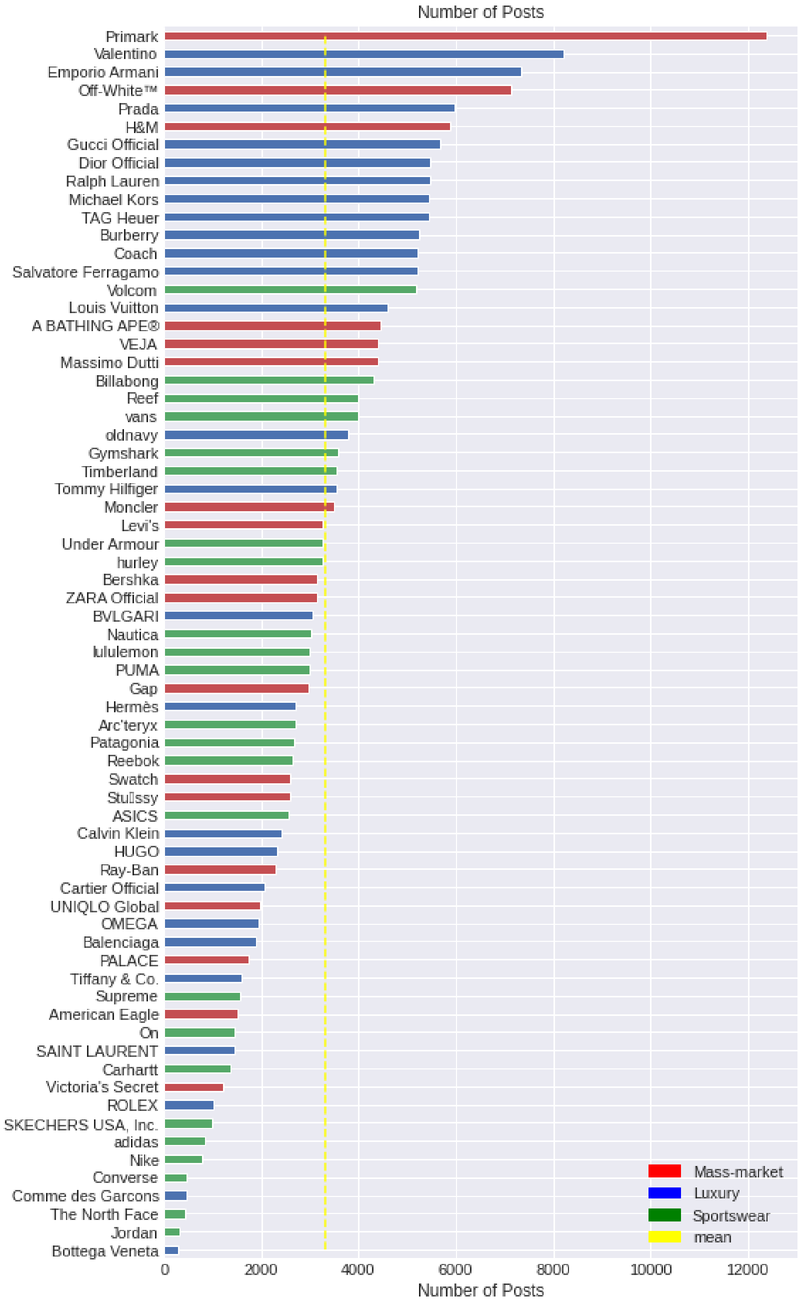


Figure 1: Average number of posts for each brand

Of course, the reason for this could be that Primark is using Instagram much longer than the other brands. This turns out not to be true given the number of posts per day (Figure 2). The distribution is similar to the number of posts in total. The three categories differ significantly (p = 0.004), with sports brands posting mainly below average per day and in total (mean = 0.9, standard deviation = 0.3) and with great distance to luxury (mean = 1.5, standard deviation = 0.9) and mass-market brands (mean = 1.6, standard deviation = 1.1).

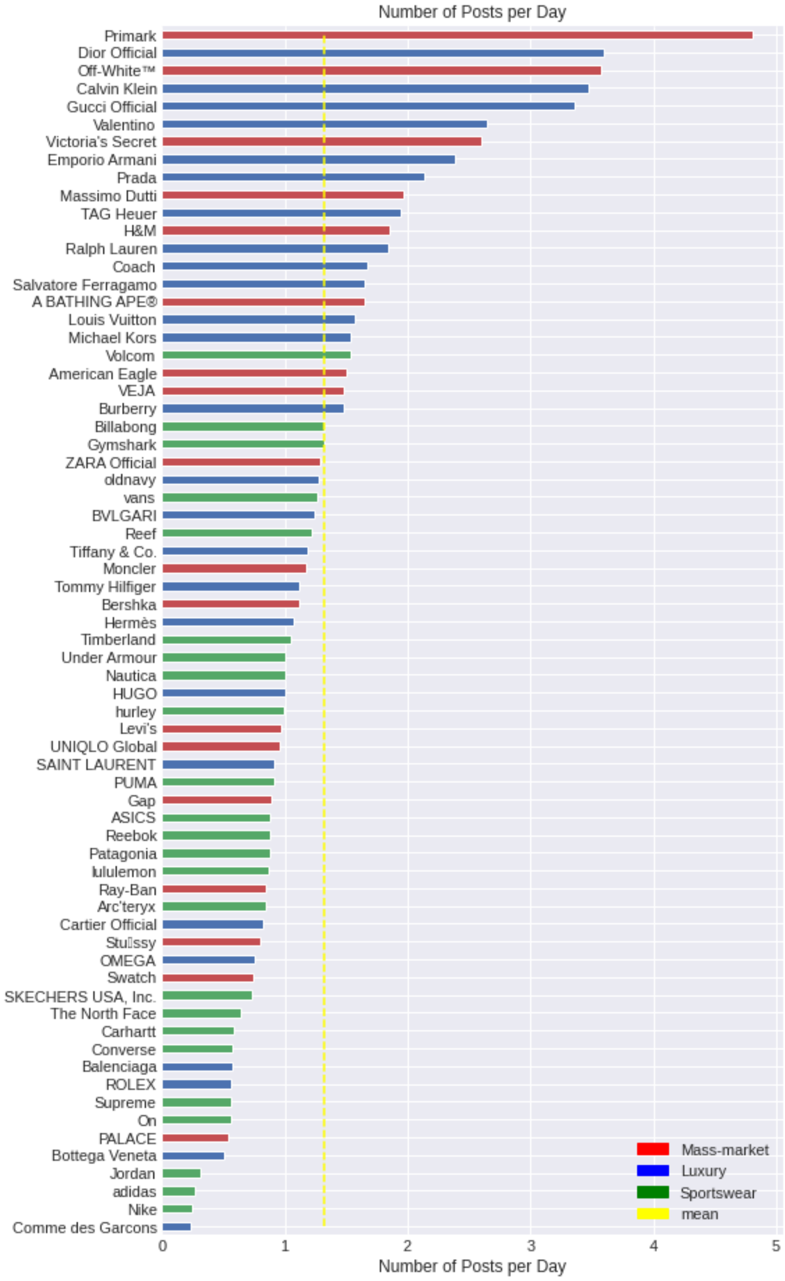


Figure 2: Average number of posts per day for each brand

Primark is a very active poster with almost five posts per day. Compared to the top sports brands with roughly one post each five days, which is more than 20 times the post volume. It will be interesting to see the effects of these extremes on community engagement.

Does the number of posts also have direct implications on the number of comments? In other words, is the engagement of the community stronger if they can read more posts? This could well be possible as users tend to look more often at profiles with very frequent changes

To abstract this from the number of posts, we now take a look at the average number of comments per post:

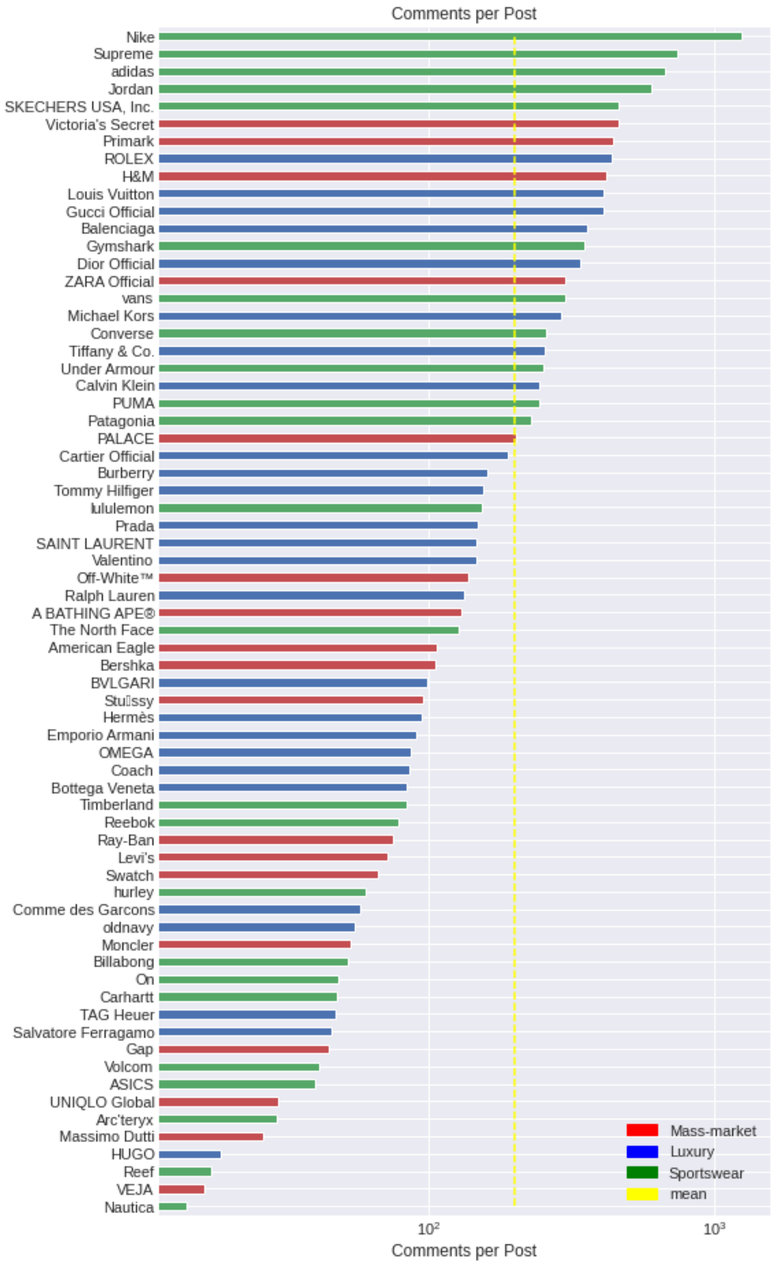


Figure 3: Average comments per Post for eacht brand

The results show a completely different ranking with Nike on the top generating 1253 comments per post on average

.

Contrary to the results concerning the posts of brands themselves, concerning community engagement sportswear brands have the most active users with 257 comments per Post on average (standard deviation = 302). Posts of luxury brands trigger 177 comments on average (standard deviation = 127), whereas mass-market brands follow with 155 comments per post (standard deviation = 149).

Still, Primark has a solid number of comments per post. So, the inverse is also not true; the community does not get bored by many posts.

4.1.2. Time to First Comment

In the next step, we considered the time it takes each profile to attract the first comment as some brands take this as an indicator for their community’s engagement. This turns out to be a difficult approach as there are many bots on Instagram which automatically comment on popular posts to gain popularity and increase their reach.

As shown in Figure 4 comments appear fastest for luxury brands with about 1442 seconds (24 minutes) in average (standard deviation = 2364 sec. – 39 minutes), followed by mass-market (mean = 1742 sec. – 29 minutes, standard deviation = 2175 sec. – 45 minutes). It seems that the sportswear community is slowest in terms of commenting on posts (mean 1908 sec. – 32 minutes, standard deviation = 1890 sec. – 32 minutes).

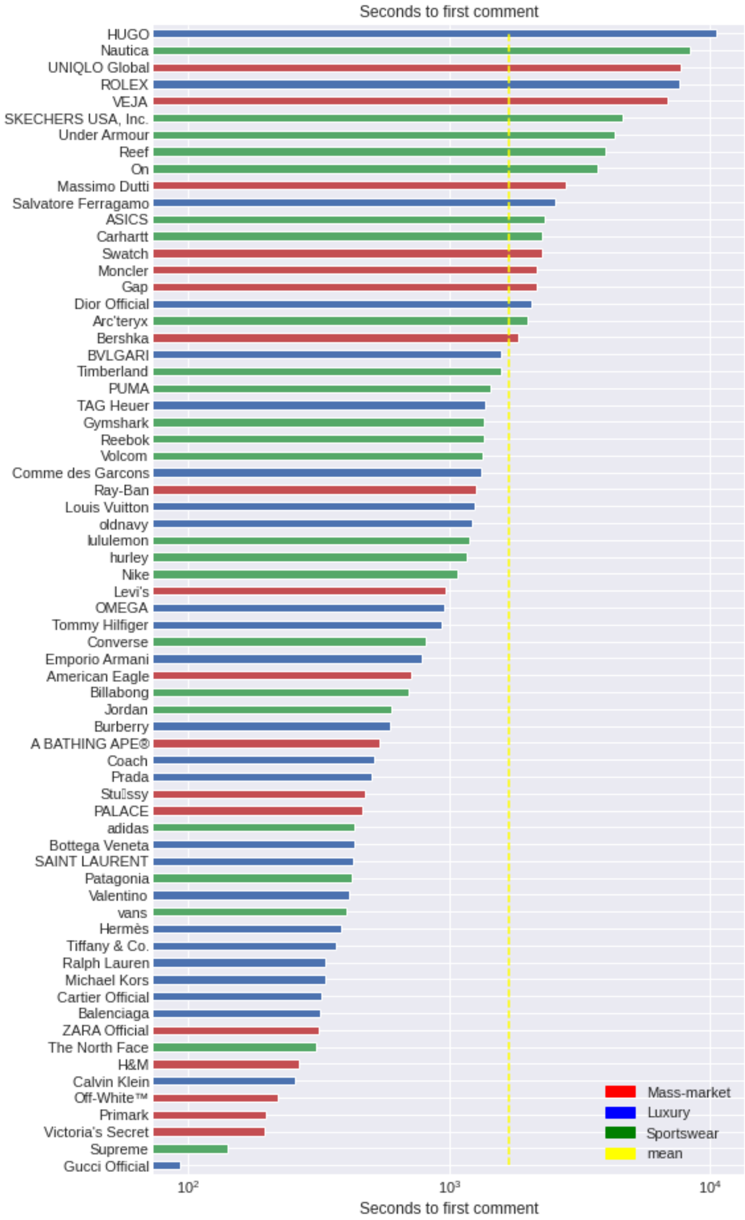


Figure 4: Seconds to first comment after post for each brand

4.1.3. Number of Likes Per Post and Follower

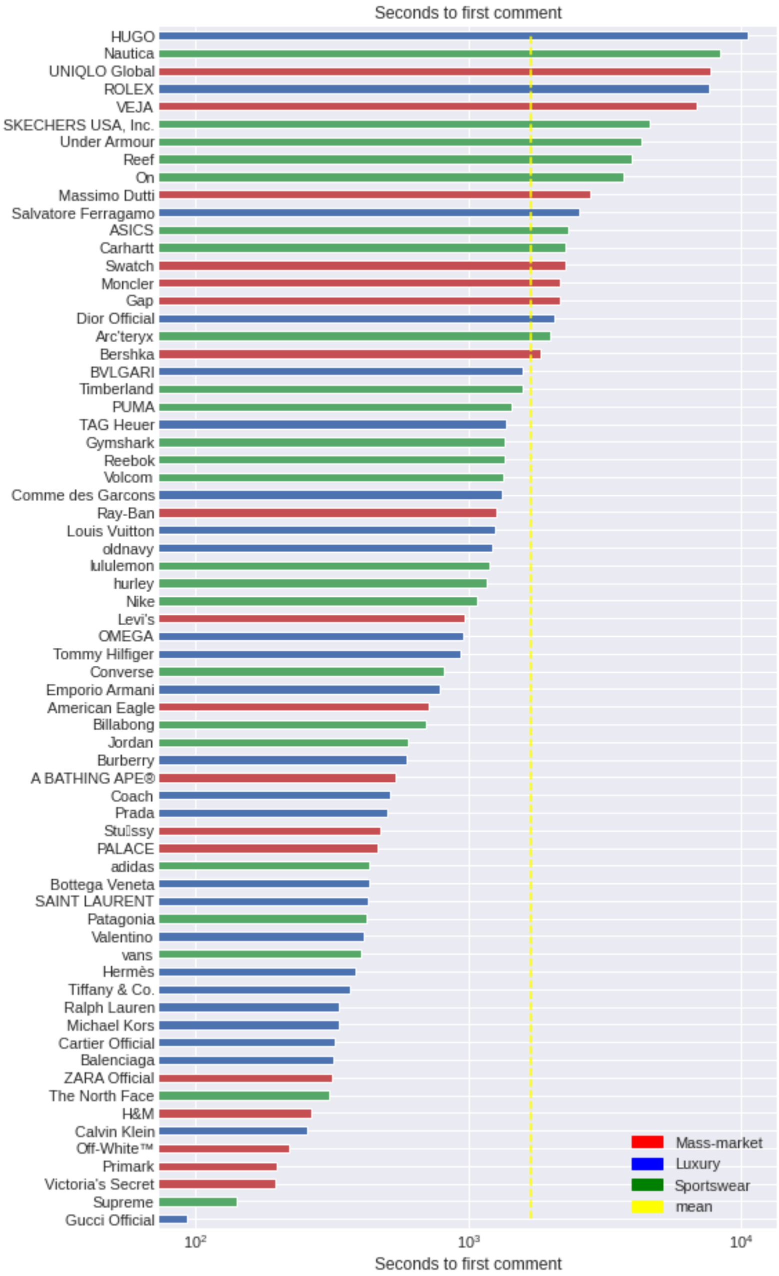


Figure 5: Average likes per post and follower for each brand

The last descriptive analysis refers to a normalized indicator by putting likes in relation to posts and followers for each brand. Figure 5 shows clearly that sportswear brands have the most active community (mean = 0.005, standard deviation = 0.002), followed by mass-market brands (mean = 0.004, standard deviation = 0.002) and luxury brands (mean = 0.003, standard deviation = 0.001).

Finally over all descriptive analyses the high variances limit the informative value concerning average indicators. Therefore we proceed with a deep dive regarding the Instagram corpus of our selected brands by analysing correlations and AI-based text analyses.

4.2. Correlation Analyses

4.2.1. Relation between Number of Followers and Time to First Comment

So, time to first comment must somehow depend on the community and engagement itself, not just on its size? It must also be taken into account that the historical time to the first of course depends on the number of followers at that time which is not available to us.

Figure 6 shows the correlation between the number of followers and the time to the first comment for each brand which is r = -0.24 (p = 0.05). That indicates that there is a significantly weak negative relationship between the both variables.

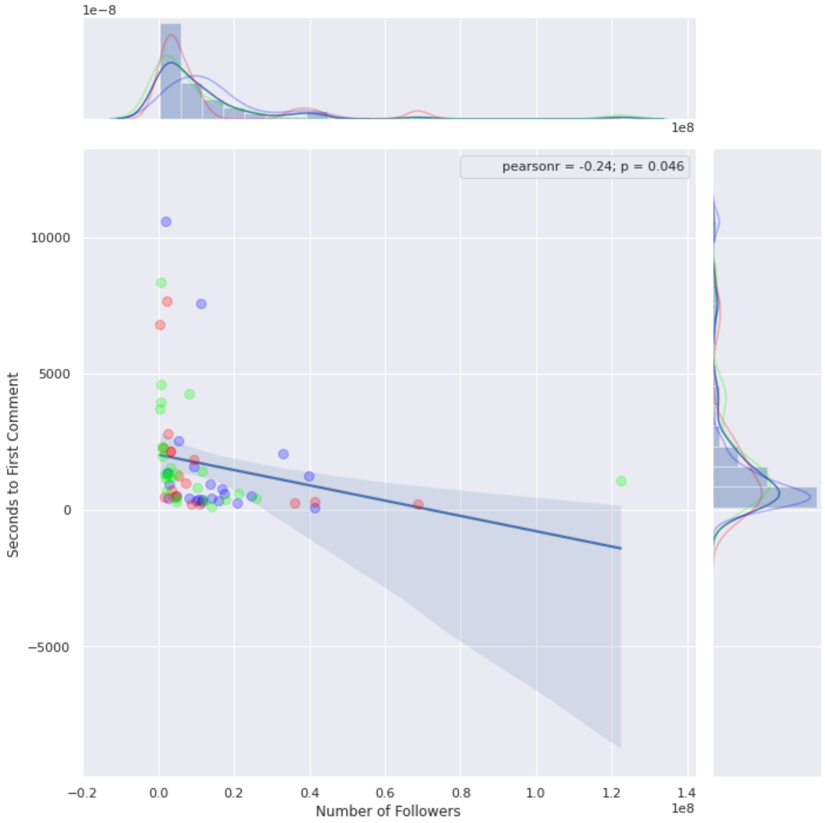


Figure 6: Jointplot correlation between number of followers and seconds to first comment

4.2.2. Relation between Likes and Comments

The correlation between the number of comments per post and the number of likes per post is strongly positive and highly significant (r = 0.86, p = 3.3e-21), which is quite intuitive: the more likes a post generates, the more comments will be made and vice versa.

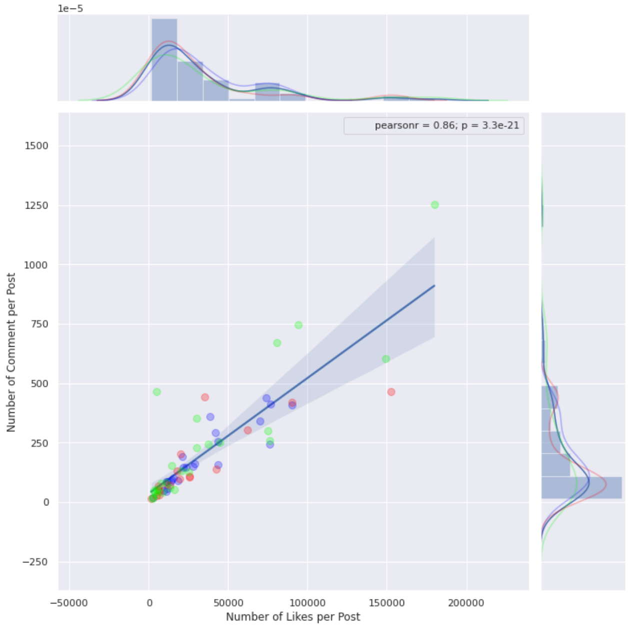


Figure 7: Jointplot of correlation between number of comments per post and numbers of likes per post

4.2.3. Relation between Number of Followers and Comments and Likes

As comments and likes are more or less linearly related, it suffices to consider comments. Likes are the same.

The strong positive correlation (r = 0.77, p = 9.3e-15) between number of followers and number of comments per post is also quite intuitive: The more followers a brand has the more comments can be expected per post.

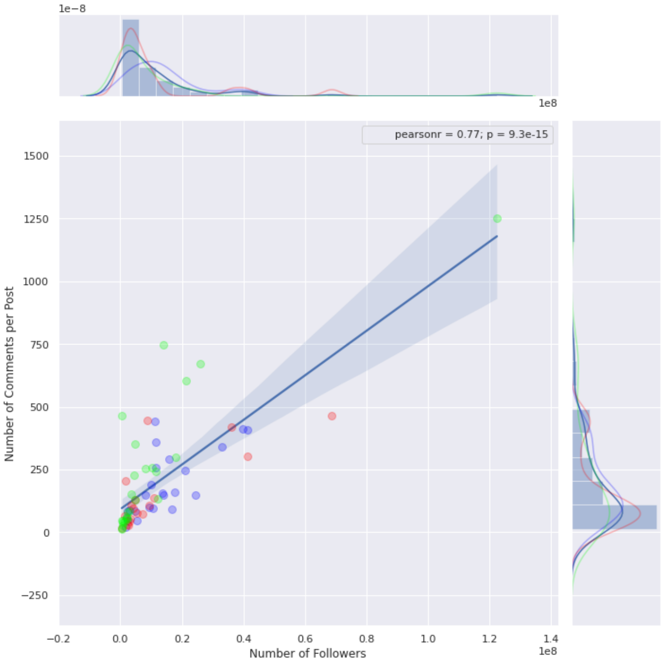


Figure 8: Jointplot of correlation between number of followers and number of comments per post

As there is one outlier brand in the category sportswear (i.e. Nike), we separated this category for a closer look on the relationship number of followers with number of likes per post:

The correlation is stronger (r = 0.9 vs. r = 0.8, both highly significant) excluding Nike as an outlier regarding the number of followers. That leads to the interpretation that Nike despite the high number of followers has a substantive amount of followers who is not active at all (regarding likes or posting comments).

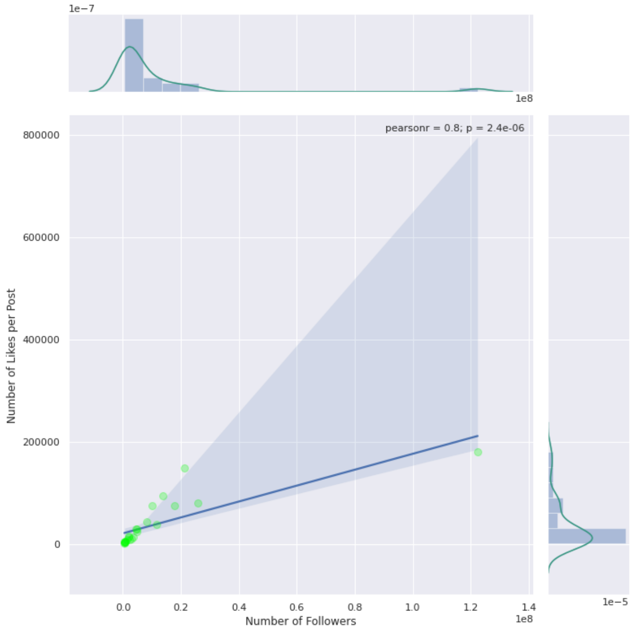


Figure 9: Jointplot of correlation between number of followers and number of likes (including outlier Nike)

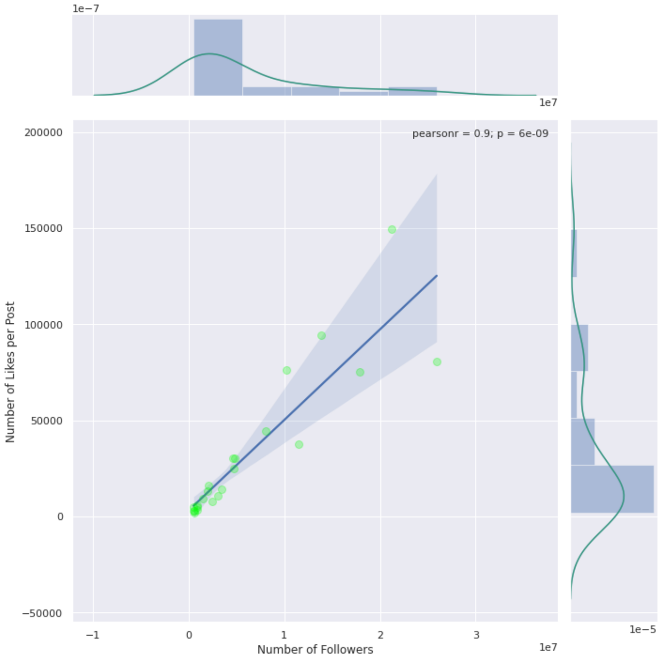
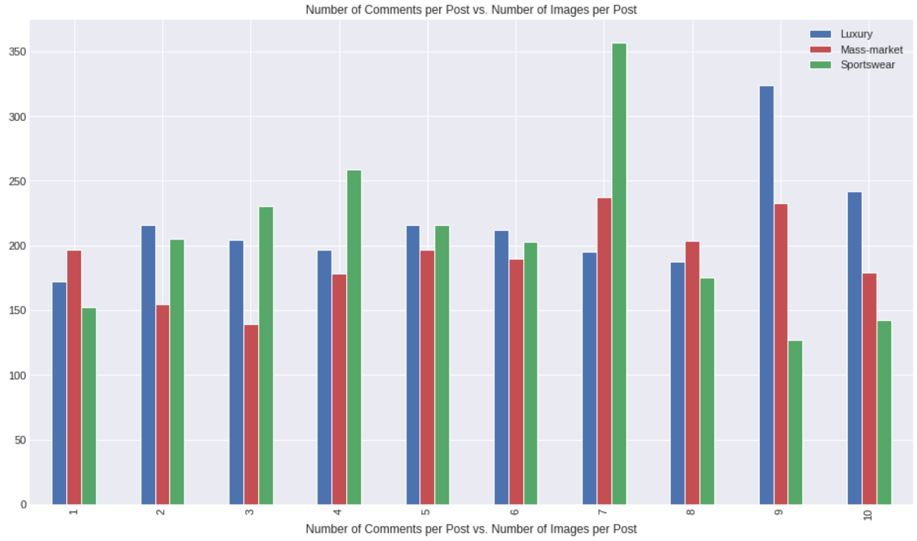


Figure 10: Jointplot of correlation between number of followers and number of likes per post (excluding outlier Nike)

4.2.4. Relation between Number of Comments and Video Vs. Images

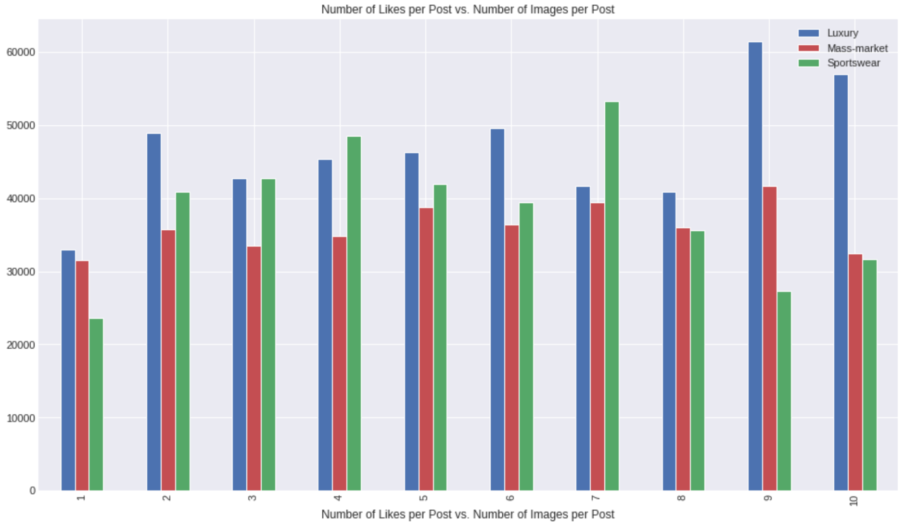
As first noted by Bonilla et al. [32] the number of videos and images in a post has a significant effect on popularity and community interaction.

To see this effect in our dataset, we first consider the number of images in a post plotted against the average number of comments:



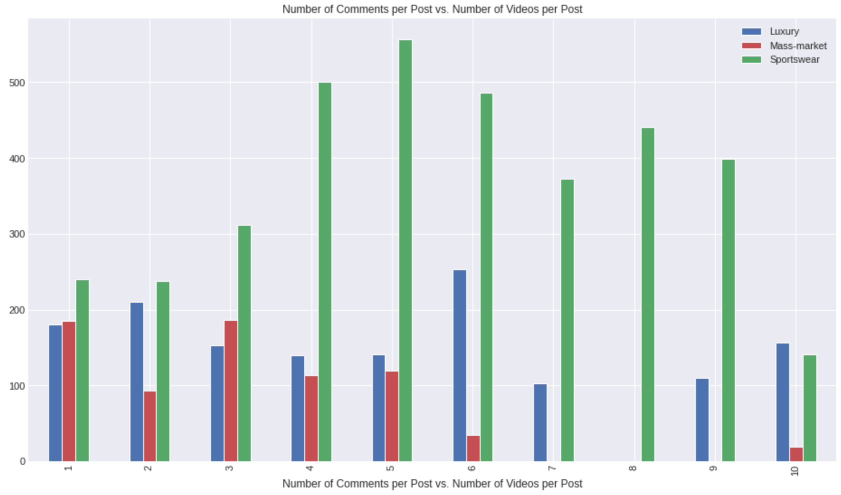
Interestingly, the results depend on the category of the fashion brand. The sportswear community honours the number of images and comments much more frequently on posts with multiple images, reaching a peak at seven images per post. The commenting behaviour in the other categories is a bit more erratic with no clear trend visible.

The story for likes is a bit different:



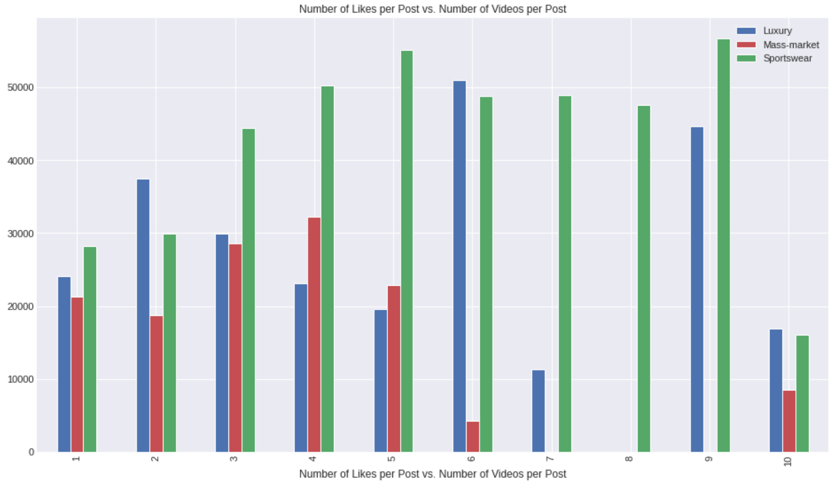
The mass market brands do not get many more likes with increasing number. But both sportswear and luxury brands have a positive effect for likes when posting with a larger number of images.

Videos are much more complicated and expensive to produce and have a similar effect depending on the category of the brand:



Comments increase specifically in the sportswear category with no clear trend in the other categories.

Likes are increased both in sportswear and in to a lesser extent in the luxury category:



4.3. Topic Modelling of Comments

Topic modelling is a technique of artificial intelligence to uncover the structure of a large text corpus [33]. There is a large variety of methods available [33]; for our analysis using non-negative matrix factorization [33] proves to be efficient and provides good results.

We are not so much interested in the posts of the brands because they focus on their own names and values. Rather, we would be interested in the much larger number of comments which are given by the followers (and fans) of the respective brands.

The number of comments is much too large for calculating a complete topic model over all comments. Additionally, that would not be fair as it has a bias toward the brands with a lot of comments. Therefore, we chose a stratification approach and took only 10,000 comments from each brand amounting to a total of 680,000 comments which is enough for calculating valid topic models.

We chose to calculate 10 topics using the NMF method. The topics are shown as wordclouds in the figure below:

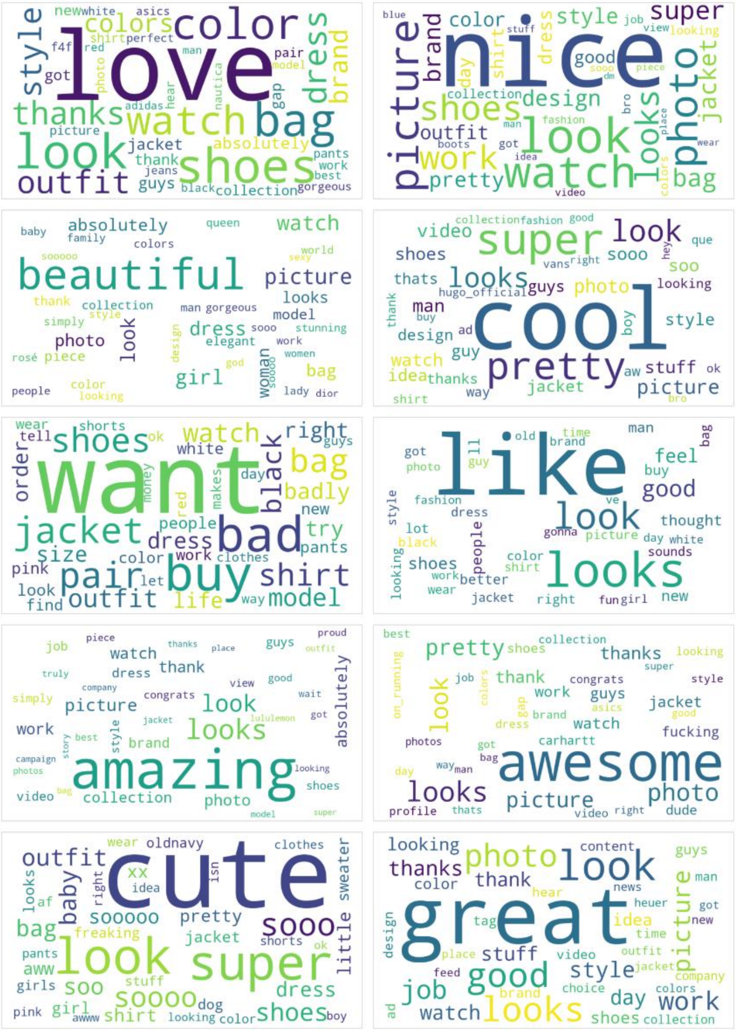


Figure 11: Topic Modelling

Instagram is not a platform which is renowned for extensive content. As expected, the topics contain a lot of generic words, but still offer quite distinctive feedback regarding the products of the respective brands.

In the next step, we wanted to know the top 5 brands for each of the topics. For this, we calculated the topic distribution for each brand and took only the top 5 brands for each topic:

Table 2: Ranking of Topics and corresponding brands

	top 1	**top 2**	**top 3**	**top 4**	**top 5**
love	SKECHERS USA, Inc.	VEJA	Michael Kors	American Eagle	Gucci Official
nice	Emporio Armani	OMEGA	Nautica	TAG Heuer	Stüssy
beautiful	Dior Official	Victoria's Secret	Tiffany & Co.	Calvin Klein	BVLGARI
cool	hurley	Nike	Volcom	Ray-Ban	Stüssy
want	A BATHING APE®	Supreme	Primark	PALACE	Swatch
like	A BATHING APE®	Jordan	PALACE	Supreme	Under Armour
amazing	Arc'teryx	hurley	The North Face	Billabong	Patagonia
awesome	Billabong	Patagonia	Arc'teryx	hurley	Volcom
cute	oldnavy	Primark	American Eagle	Gap	ZARA Official
great	On	Arc'teryx	The North Face	TAG Heuer	Patagonia

The data can also be visualized more qualitatively but in more detail using a heatmap. The order of the brands has been adjusted to have the luxury brands first, then the mass market and finally the sports brands:

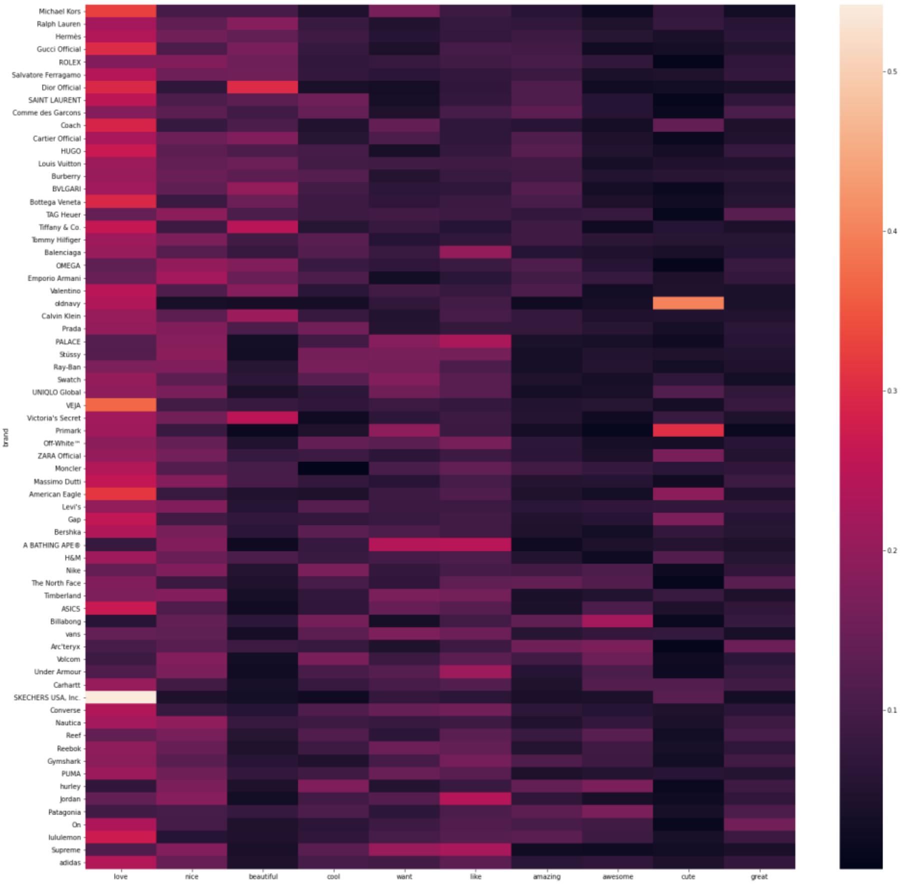
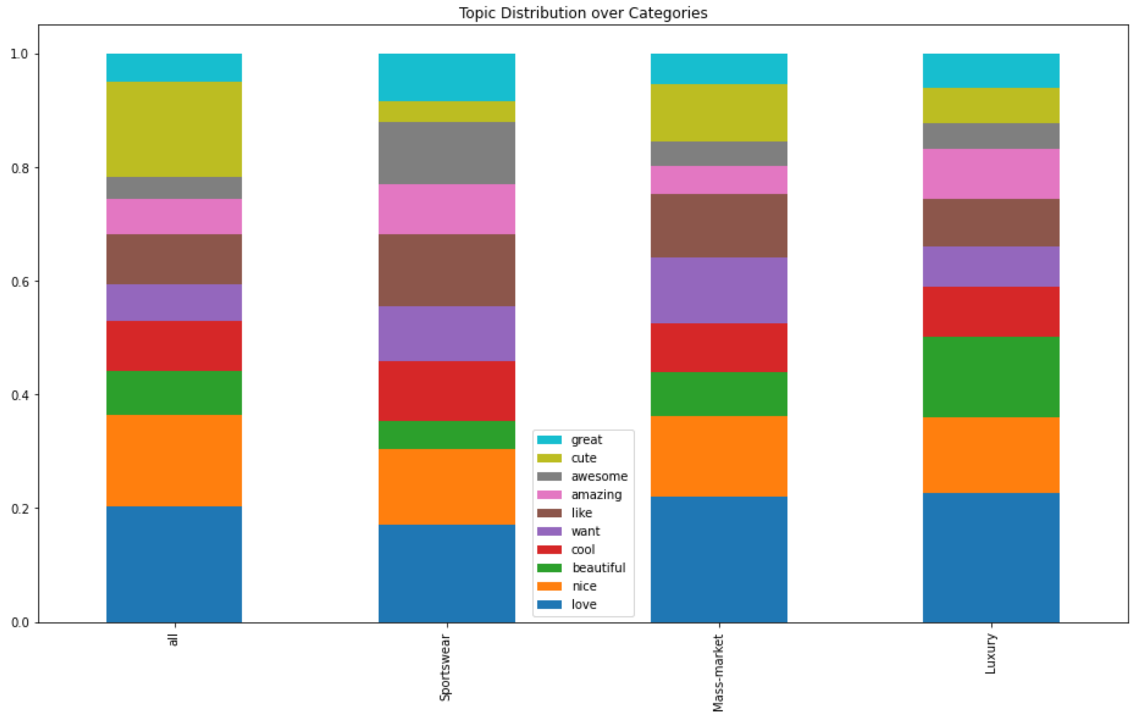


Figure 12: Heatmap of topics and their appearance in brands' UGC

In this representation, it is quite easy to see that cute is most prominent in oldnavy, love is most popular for SKECHERS etc.

In a final step, we calculated the topic distribution of the different categories:



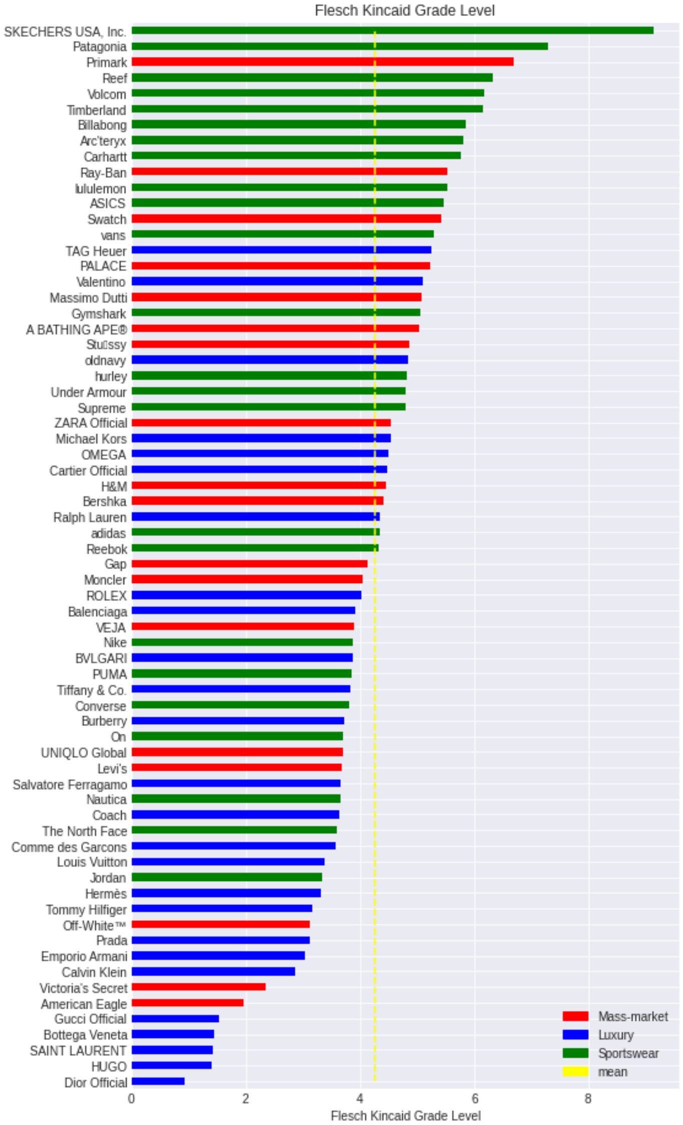
As expected, the "cute" topic is not much associated with "sportswear" but rather with mass-market. The "beautiful" topic is mainly present in the luxury brands, which is also according to our expectations. All in all it looks like the different categories reach their target audience quite well with the feedback going into the same direction as the primary

4.4. Text Complexity in Comments

For analysing the complexity we used the Flesch-Kincaid readability tests [33]. The calculation was perfomed using the Python package textacy [34]. As the necessary linguistic analysis is quite expensive, we used the same stratified dataset as in topic modelling.

It's very interesting to see that the complexity of the comments is much higher in sportswear compared to luxury brands. This can be attributed to many "empty" discussions just containing emojis or trying to gain popularity, which is much more common in the luxury regime.

Taking samples from sportswear, it became clear that the discussions there are much more centred around the presented products and commenters exchange tips regarding their gear. This of course contributes to much more complicated text.



5. Conclusions and Outlook

In this paper, we have shown how Instagram works for fashion brands as a marketing platform and how their individual success can be measured. Coming from descriptive statistics, we have used machine learning models like topic models to quantify the response of the community.

Comments and likes were shown to be good indicators of the success of various brands. We have shown that the engagement of the community depends heavily on the kind of brand with a totally different response from sportswear compared to luxury brands. Interestingly, the comments for sportswear brands have a much higher linguistic complexity compared to luxury brands. However, this has to be taken with a grain of salt as the Instagram comments are always quite short.

Our process was designed in a way to easily facilitate easy updates. All data is aggregated in a central database and used for automatically creating all diagrams. This leads to a scalable, reproducible analysis which is essential if you use the data to plan your (expensive) marketing campaigns.

In the future, we plan to extend the analysis to analyse the Emoji content of the posts and mainly the comments. New work regarding Emoji embedding [35] allows us to find uncommon Emojis and take a specific look at that. Additionally, we can use pre-classified Emojis [36] to have a hierarchy of feelings associated with individual Emojis.

Apart from starting with the brands, there are many other ways to approach Instagram content. We have already started with community detection to find special interests also within the fashion industry (e.g. sneakers). Experts will also convey weak signals and this can be used for very early trend detection.

Yet another possibility is given by local Instagram content. Many posts are annotated with a geolocation. Eventually, this will allow the detection of local and hyperlocal trends.

REFERENCES

[1] Heinonen, K., Medberg, G. (2018). "Netnography as a tool for understanding customers: implications for service research and practice", Journal of Services Marketing

[2] Kozinets, R. V. (1999). „E-tribalized marketing?: the strategic implications of virtual communities of consumption“, European Management Journal, 1999, vol. 17, issue 3, 252-264

[3] Bickart, B., & Schindler, R. M. (2001). „Internet Forums as Influential Sources of Consumer Information“. Journal of Interactive Marketing, 15, 31-40

[4] Catterall, M., Maclaran, P., Stevens, L. (2002). "Consumption and Gender", in GCB - Gender and Consumer Behavior Volume 6, eds. Pauline Maclaran, Paris, France: Association for Consumer Research.

[5] Kozinets, R. V. (2006). „Click to Connect: Netnography and Tribal Advertising“. Journal of Advertising Research, vol. 46, no. 3, 279–288

[6] Simmons, G. (2008). „Marketing to postmodern consumers: Introducing the Internet chameleon“, European Journal of Marketing 42(3/4)

[7] Tikkanen, H., Hietanen, J., Henttonen, T., Rokka, J. (2009). „Exploring virtual worlds: Success factors in virtual world marketing“. Management Decision 47(8):1357-1381

[8] Rokka, J. (2010). „Netnographic inquiry and new translocal sites of the social“. International Journal of Consumer Studies, Volume34, Issue4, July 2010, 381-387

[9] Musabirov, I., Bulygin, D. (2020). „Prototyping Text Mining and Network Analysis Tools to Support Netnographic Student Projects“. International Journal of Emerging Technologies in Learning, Vol. 15, No. 10 (2020)

[10] Statista (2020), Most popular social networks worldwide as of July 2020, ranked by number of active users, [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[11] Later (2020), The Top 50 Instagram Hashtags of All Time, [Online]. Available: https://later.com/blog/ultimate-guide-to-using-instagram-hashtags/#top

[12] Leaver, T., & Highfield, T. (2018). Visualising the ends of identity: Pre-birth and post-death on Instagram. Information, Communication & Society, 21(1), 30–45.

[13] Roncha, A., &Radclyffe-Thomas, N. (2016). How TOMS’ “One day without shoes” campaign brings stakeholders together and co-creates value for the brand using instagram as a platform. Journal of Fashion Marketing and Management, 20(3), 300–321.

[14] Paine, K. D. (2011). Measure what matters: Online tools for understanding customers, social media, engagement, and key relationships. Hoboken, New York: John Wiley and Sons.

[15] Dhanesh, G.S. (2017), “Social media and the rise of visual rhetoric: implications for public relations theory and practice”, in Bridgen, E. and Vercic, D. (Eds), Experiencing Public Relations, Routledge, New York, NY, pp. 137-150.

[16] Mersey, R., Malthouse, E.C. and Calder, B.J. (2010), “Engagement with online media”, Journal of Media Business Studies, Vol. 7 No. 2, pp. 39-56.

[17] Muntinga, D.G., Moorman, M. and Smit, E.G. (2011), “Introducing COBRAs: exploring motivations for brand-related social media use”, International Journal of Advertising, Vol. 30 No. 1, pp. 13-46.

[18] FashionUnited (2020), „Most valuable fashion brands“ [Online]. Available: https://fashionunited.com/i/most-valuable-fashion-brands

[19] BBC (2019), Facebook staff 'flagged Cambridge Analytica fears earlier than thought', [Online]. Available: https://www.bbc.com/news/technology-47666909

[20] GitHub (2020), Instaloader, [Online]. Available: https://instaloader.github.io/

[21] Squid (2020), Optimising Web Delivery, [Online]. Available: http://www.squid-cache.org/

[22] Raspberry Pi 4 (2020), [Online]. Available: https://www.raspberrypi.org/

[23] OpenVPN (2020), Building a Strong Community, [Online]. Available: https://openvpn.org

[24] JSON (2020), JavaScript Object Notation, [Online]. Available: https://www.json.org

[25] SQLite (2020), What Is SQLite?, [Online]. Available: www.sqlite.org

[26] Instagram (2011, Profile Michael Kors, [Online]. Available: https://www.instagram.com/p/Bfs65/

[27] Project Jupyter (2020), The Jupyter Notebook, [Online]. Available: www.jupyter.org

[28] Python (2020), Compound Data Types, [Online]. Available: www.python.org

[29] NumFOCUS (2020), pandas, [Online]. Available: https://pandas.pydata.org/

[30] NumFOCUS (2020), SciPy, [Online]. Available: https://www.scipy.org/

[31] Scikit-Learn (2020), Machine Learning in Python, [Online]. Available: https://scikit-learn.org/stable/

[32] María Del Rocío Bonilla, José Luis del Olmo Arriaga & David Andreu (2019) : The interaction of Instagram followers in the fast fashion sector: The case of Hennes and Mauritz (H&M), Journal of Global Fashion Marketing]

[33] Albrecht, J., Ramachandran, S. and Winkler, C. Blueprints for Text Analysis Using Python, Sebastopol: O’Reilly, 2020

[34] textacy 0.10.1 (2020). textacy: NLP, before and after spaCy, [Online]. Available: https://pypi.org/project/textacy/

[35] Al-Halah, Z., Aitken, A., Shi, W., Caballero, J. (2019). „Smile, Be Happy :) Emoji Embedding for Visual Sentiment Analysis“ [Online]. Available: https://arxiv.org/abs/1907.06160

[36] Unicode (2020). „Full Emoji List, v13.1“ [Online]. Available: https://unicode.org/emoji/charts/full-emoji-list.html

AUTHORS

Prof. Dr. Stefanie Scholz

Professor for social economy at Wilhelm Loehe University of Applied Sciences, Fuerth, Germany. Focussed on data driven marketing, netnography, consumer behaviour and customer ermpowerment

Dr. Christian Winkler

Christian Winkler holds a PhD in Theoretical Physics. He has worked in software and AI for 20 years, specializing in intelligent algorithms for unstructured data and text. He is a frequent speaker at conference and author of many articles and tutorials.

REGULARIZATION METHOD FOR RULE REDUCTION IN BELIEF RULE-BASED SYSTEM

Yu Guan

College of Mathematics and Computer Science,
Fuzhou University, Fuzhou, China

ABSTRACT

Belief rule-based inference system introduces a belief distribution structure into the conventional rule-based system, which can effectively synthesize incomplete and fuzzy information. In order to optimize reasoning efficiency and reduce redundant rules, this paper proposes a rule reduction method based on regularization. This method controls the distribution of rules by setting corresponding regularization penalties in different learning steps and reduces redundant rules. This paper first proposes the use of the Gaussian membership function to optimize the structure and activation process of the belief rule base, and the corresponding regularization penalty construction method. Then, a step-by-step training method is used to set a different objective function for each step to control the distribution of belief rules, and a reduction threshold is set according to the distribution information of the belief rule base to perform rule reduction. Two experiments will be conducted based on the synthetic classification data set and the benchmark classification data set to verify the performance of the reduced belief rule base.

KEYWORDS

Knowledge-based system, Belief rule base, Regularization method, Rule reduction.

1. INTRODUCTION

The inference system of the belief rule base proposed by Yang et al. with a belief distribution structure and evidential reasoning method is based on the research results of D-S evidence theory, fuzzy theory, and generative IF-THEN rules. Belief rule-based inference system can effectively synthesize the missing, fuzzy, and uncertain parts of the input information. In the inference process of the belief rule base, the attribute weight, rule weight, belief distribution, and other parameters in the system directly affect the accuracy of the final inference prediction result. To improve the inference accuracy of the belief rule base, Yang et al. proposed a parameter optimization model of the belief rule base. Later researches also proposed a series of belief rule base parameter optimization models using different machine learning algorithms. The early belief rule base can only construct rules based on the specific domain knowledge of the human expert, and cannot construct a reasoning system containing a large number of rules. The extended belief rule-based inference system uses the training data set to construct the rule base based on the data-driven concept.

This paper proposes the optimized belief rule base further simplifies the belief distribution structure of the belief on the basis of the extended belief rule base system, further improves the efficiency of constructing belief rules through training data sets, and reduces the complexity of rule storage and operation. Then a parameter training method based on regularization is proposed to reduce redundant rules. The rest of this paper is organized as follows: In Section II, we

Natarajan Meghanathan et al. (Eds): CSE, BMLI, EDU, AI & FL, CCNET - 2020
pp. 51-61, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101705

reviewed the basic structure and parameter training model of the belief rule base. Then Section III proposes the optimized belief rule structure, evidence reasoning method, and parameter training method based on regularization. The two experiments in Section IV verify the performance of the reduced belief rule base system, and Section V concludes the paper.

2. Overview of Belief Rule Base Inference System

The construction of the belief rule base is based on the IF-THEN generative rules and the belief distribution framework, and the professional domain knowledge is obtained through expert setting or other methods. The inference engine uses the D-S evidence synthesis theory to synthesize the conclusions of different rules to obtain the final reasoning result. This section will specifically introduce the structure of the belief rule base and the process of the evidential reasoning of the belief rule base.

2.1. Belief Rule Base

The belief rule base proposed by Yang et al. is based on the traditional IF-THEN generative rules, and introduces the structure of belief distribution on the result attribute, and introduces the weight of the antecedent attributes and the rule weight, which can effectively express uncertain information. The structure of the belief rule base has L rules, T attributes, and N results is as follows:

$$\begin{gathered} R_k\text{: if } \{x_1 is A_1^k \wedge \cdots \wedge x_T is A_T^k\} \text{ then } \{(D_1,\beta_1^k), \cdots, (D_N,\beta_N^k)\} \\ \text{with rule weight } \theta_k, k=1, \cdots, L \text{ and attribute weight } \delta_1, \cdots, \delta_T \end{gathered} \tag{1}$$

$(x_1, \cdots, x_T)$ belong to a certain reference candidate of their antecedent attributes, for any A_i^k satisfies $A_i^k \in \{A_{i,1}, \cdots, A_{i,J_i}\}$ .In the extended belief rule-based inference system, the belief distribution structure is further introduced into the antecedent attributes, which improves the model's ability to express fuzziness and incomplete information. The th rule in the extended belief rule base can be expressed as:

$$\begin{gathered} R_k \text{ if: } \{[(A_{11}^k,\alpha_{11}^k), \cdots, (A_{1J_1}^k,\alpha_{1J_1}^k)] \wedge \cdots \wedge [(A_{T1}^k,\alpha_{T1}^k), \cdots, (A_{TJ_T}^k,\alpha_{TJ_T}^k)]\} \\ \text{then } \{(D_1,\beta_1^k), \cdots, (D_N,\beta_N^k)\} \\ \text{with rule weight } \theta_k, k=1, \cdots, L \text{ and attribute weight } \delta_1, \cdots, \delta_T \end{gathered} \tag{2}$$

The rules in the extended belief rule base can be directly generated from the training data set. For the input data, convert the th attribute parameter to construct the th antecedent attribute of the corresponding rule in the form of belief distribution using the corresponding reference values of candidate attribute $\{(A_{i1},\gamma_{i1}), \cdots, (A_{iJ_i,\gamma_{iJ_i}})\}$:

$$\begin{gathered} \alpha_{ij}^k = \frac{\gamma_{i(j+1)-x_i^k}}{\gamma_{i(j+1)-\gamma_{ij}}}, \gamma_{ij} \le x_i^k \le \gamma_{i(j+1)} \\ \alpha_{i(j+1)}^k = 1 - \alpha_{ij}^k, \gamma_{ij} \le x_i^k \le \gamma_{i(j+1)} \\ \alpha_{it}^k = 0, t=1, \cdots, (j-1), (j+2), \cdots, J_i \end{gathered} \tag{3}$$

2.2. Reasoning Method using Evidential Reasoning

The belief rule-based reasoning system uses the evidential reasoning method to synthesize rule results. The inference process consists of the following steps in sequence:

1) belief rule activation weight calculation

For the component of input on any antecedent attribute, convert it to the belief distribution on the corresponding attribute, and the method is as follows:

$$\begin{gathered}\alpha_{ij}^{k}=\frac{\gamma_{i(j+1)-x_i}}{\gamma_{i(j+1)-\gamma_{ij}}},\gamma_{ij}\le x_i\le\gamma_{i(j+1)}\\ \alpha_{i(j+1)}=1-\alpha_{ij},\gamma_{ij}\le x_i\le\gamma_{i(j+1)}\\ \alpha_{it}=0,t=1,\cdots,(j-1),(j+2),\cdots,J_i\end{gathered}\tag{4}$$

Using the belief distribution after input conversion, the individual matching degree of the k th rule on i th antecedent attribute is calculated as:

$$S_i^k=1-d_i^k=1-\sqrt{\frac{\sum_{j=1}^{J_i}(\alpha_{i,j}-\alpha_{i,j}^k)^2}{2}}\tag{5}$$

After the individual matching degree of each attribute is calculated, the individual matching degrees of all attributes are aggregated. The aggregation function in the form of conjunctive rules is:

$$\alpha_k=\prod_{i=1}^{T_k}(S_i^k)^{\bar{\delta}_i},\bar{\delta}_i=\frac{\delta_i}{\max\limits_{j=1,\cdots,T_k}\delta_j}\tag{6}$$

The activation weight of this rule is calculated by the following formula:

$$w_k=\frac{\theta_k\alpha_k}{\sum_{l=1}^{L}\theta_l\alpha_l}\tag{7}$$

Rule weight normalization operation makes every activation weights satisfy $0\le w_k\le 1,\sum w_k=1$.

2) evidential reasoning of belief rule base

After the rule weight calculation is completed, all the rules are synthesized and the inference result is obtained. First, the belief distribution of the rule is transformed into the corresponding probability mass information:

$$
\begin{aligned}
m_{j,k} &= w_k \beta_j^k, j = 1, \cdots, N \\
m_{D,k} &= 1 - \sum_{j=1}^{N} m_{j,k} = 1 - w_k \sum_{j=1}^{N} \beta_j^k \\
\bar{m}_{D,k} &= 1 - w_k \\
\tilde{m}_{D,k} &= w_k \Big(1 - \sum_{j=1}^{N} \beta_j^k\Big)
\end{aligned} \tag{8}
$$

where $m_{j,k}$ represents the credibility of the kth rule on the jth consequent attribute, $\bar{m}_{D,k}$ represents the credibility that the kth rule is not assigned to any consequent attribute, and $\tilde{m}_{D,k}$ represents the credibility of the missing reference attribute of the kth rule. The total uncertainty credibility is given by$m_{D,k} = \bar{m}_{D,k} + \tilde{m}_{D,k}$.

Synthesize the credibility information of all rules and obtain the final belief result of each consequent attribute:

$$
\begin{aligned}
m_j &= k\Big[\prod_{i=1}^{L}(m_{j,i} + m_{D,i}) - \prod_{i=1}^{L} m_{D,i}\Big], j = 1, \cdots, N \\
\bar{m}_D &= n\Big[\prod_{i=1}^{L} \bar{m}_{D,i}\Big] \text{ and } \tilde{m}_D = k\Big[\prod_{i=1}^{L} m_{D,i} - \prod_{i=1}^{L} \bar{m}_{H,i}\Big] \\
k &= \Big[\sum_{j=1}^{N}\prod_{i=1}^{L}(m_{j,i} + m_{D,i}) - (N-1)\prod_{i=1}^{L} m_{D,i}\Big]^{-1} \\
\beta_j &= \frac{m_j}{1 - \bar{m}_D}, j = 1, \cdots, N \text{ and } \beta_D = \frac{\tilde{m}_D}{1 - \bar{m}_D}
\end{aligned} \tag{9}
$$

3. Belief Rule Base Structure Optimization and Regularization Method

This section optimizes the traditional belief rule base structure by simplifying the belief structure of antecedent attributes and introduces the Gaussian membership function to optimize the calculation of activation weights. This section also proposes a group-level evidential reasoning method to avoid reasoning failure when there are too many rules. For the optimized inference system of the belief rule base, the regularization method is used to restrict and select different parameters in the belief rule base step by step during the training process, and the rules are further screened and reduced according to the rule parameters after training.

3.1. Structural Optimization of Belief Rule

The rule structure used by the conventional belief rule-based inference system is based on the belief distribution form. When constructing the belief rule and inferring the input data, it is necessary to convert the data into the corresponding belief distribution form, which requires additional computing and storage resources. It is also necessary to set the attribute candidate reference values in advance, and the empirical knowledge of human experts is required. The unreasonable setting of the number of candidates and values will reduce the accuracy of the system's reasoning.

Using the Euclidean distance method to calculate the attribute similarity may calculate abnormal activation weights, which may cause the inference system to fail due to the rule zero activation problem. To prevent the rule zero activation problem, it is necessary to construct more rules to cover all the possibilities, which causes the explosion of the number of rules caused by the increase in the number of attributes.

Because of the above shortcomings, this section first optimizes the rule structure by simplifying the belief distribution structure to avoid using the activation weight calculation method based on the similarity of the belief distribution, avoiding the problem of reasoning failure that it may cause, and simplifying the complexity of the rule construction. For training data $X_k=(x_1^k,\cdots,x_T^k)$, the corresponding belief rule that directly simplifies the belief distribution structure of the antecedent attributes is as follows:

$$\begin{gathered} R_k\text{: if } \{x_1^k \wedge \cdots \wedge x_T^k\} \text{ then } \{(D_1,\beta_1^k),\ \cdots,(D_N,\beta_N^k)\} \\ \text{with rule weight } \theta_k, k=1,\ \cdots,L \text{ and attribute weight for each rule } \delta_1^k,\cdots,\delta_T^k \end{gathered} \tag{10}$$

The simplified belief rule directly uses the attribute information corresponding to the original data to construct without the conversion of the belief distribution, avoiding additional computing and storage resources. The optimized belief rule structure cannot calculate the individual matching degree of the rule attributes by calculating the Euclidean distance of the belief distribution. This section proposes to use the Gaussian membership function to optimize the activation weight calculation process. The Gaussian membership function is a function in fuzzy theory that calculates the degree to which a specified element belongs to a specific set. Its form is as follows:

$$gaussianmf(x;\delta,c)=e^{-\frac{(x-c)^2}{\delta^2}} \tag{11}$$

For the input data$X=(x_1,\cdots,x_T)$, the individual matching degree on the ith attribute of kth rule is calculated using the Gaussian membership function:

$$S_i^k=e^{-\frac{[\delta_i^k\times(x_i-x_i^k)]^2}{\theta_k^2}} \tag{12}$$

The individual matching degree is determined by the distance from the input attribute information to the corresponding rule attribute, the rule attribute weight parameter, and the rule weight parameter. Unlike the conventional belief rule-based inference system that uses uniform attribute weights, each rule has its own attribute weight parameter, which can achieve finer rule activation granularity. Under the confidence rule setting of the conjunctive relationship, the final activation weight of this rule is calculated by the following formula:

$$w_k=\prod_{i=1}^{T}S_i^k=e^{-\frac{\sum_{i=1}^{T}[\delta_i^k\times(x_i-x_i^k)]^2}{\theta_k^2}} \tag{13}$$

The calculated activation weight is within the range of $(0,1]$ without the need for weight normalization, which simplifies the reasoning process. There will be no rule activation value zero, which avoids the potential rule zero activation problem and improves the robustness of the inference system.

3.2. Parameter Training of Regularization Method

This section uses the step-by-step regularization method. First, the regularization penalty of the rule antecedent attributes and attribute weights is introduced in the training process to make the belief rule distribution of the inference system more widely representative. Then the antecedent attributes and attribute weights are fixed, and the regularization penalty on the rule weight is introduced. The distribution of the activation weight of the belief rule base is restricted through training, and the rule weight is used to measure the importance of the corresponding rule after training. After training, the rule weight is used to determine whether to retain the corresponding rule, and the rule reduction of the belief rule base is realized.

1) complexity expression of belief rule base

According to the belief rule structure in equation (10), both the rule antecedent attributes and the corresponding attribute weights need to be restricted within a certain range. The restriction of antecedent attributes makes the rule not far away from the data distribution, which makes the rule redundant and useless. The restriction of the attribute weights makes the activation of the rule always represent the distribution of a part of the data, avoiding too high attribute weight to make the rule fit noisy data. The objective function can be expressed when using the L_2 regularization method to construct the belief rule base penalty:

$$obj_1 = \min\{loss + \lambda_x||x||_2^2 + \lambda_\delta||\delta||_2^2\} \tag{14}$$

The regularization coefficients λ_x and λ_δ are used to control the degree of the penalty of the model, to prevent the model penalty from being too large or too small from affecting the final inference accuracy. The calculation of the rule activation weight in equation (13) is also affected by the rule weight. When using the objective function, the rule weight needs to be fixed in advance to avoid model penalty failure. When using the L_1 regularization method to construct the objective function that includes the model penalty, according to the belief rule structure in (10) and the calculation method of activation weight in equation (13), The penalty of rule weight can effectively restrict the activation range of the rule, so that the activation range of the redundant rule that activates the same distributed data is limited and reduced. By comparing the weight parameters of the rules after training, important rules can be selected for rule reduction, and the corresponding objective function can be expressed as:

$$obj_2 = \min\{loss + \lambda_\theta||\delta||_1\} \tag{15}$$

In the same activation weight calculation process, antecedent attributes and attribute weights are involved. The use of the objective function requires fixed antecedent attributes and attribute weights.

2) rule reduction of step-by-step regularization training method

Combining the newly proposed belief rule structure, regularization penalty construction method, and group-level rule evidential reasoning method, This chapter proposes a step-by-step method of parameter training with rule reduction capabilities. Each step fixes different parameters and sets a specific regularization penalty to filter out representative belief rules. The specific steps of the step-by-step regularization parameter training method are as follows:

```
Algorithm 1: Regularization Parameter Training Method
Input: the attributes of belief rule base and regularization coefficients $x,\delta,\theta,\lambda_x,\lambda_\delta,\lambda_\theta$, the training data set $D$ and$obj_1,obj_2$ function and model $loss = Loss(D;x,\delta,\theta)$
Output: Reduced belief rule base
1: Fixed rule weight$\delta$, set $obj_1$ as the objective function with$\lambda_x,\lambda_\delta$, Use training set $D$ for training
2: Fixed antecedent attribute $x$ and attribute weights$\delta$, Set $obj_2$ as the objective function with $\lambda_\delta$ , Use training set $D$ for another training
3: Set reduction threshold $ts$ based on rule weight after training and reduce rules with rule weight less than $ts$
4: Set $loss$ as the objective function, Use training set $D$ to train the reduced belief rule base
Return the reduced belief rule base attributes
```

4. EXPERIMENTS

This section first uses a synthetic binary classification data set containing two numerical attributes to calculate the activation weight distribution of the belief rule base after each training step is completed and verify the effectiveness of the regularization method to limit the activation range of the rule. Then the UCI public classification data set is used to compare the inference performance of the belief rule base after reduction under different reduction threshold parameter settings.

4.1. Synthetic Binary Data Set Experiment

This section selects the artificially generated binary data set with two numerical attribute variables used in [1], which contains a total of 250 data, with 125 positive and negative samples each. To facilitate the setting of the attribute weights, the two numerical attribute variables are standardized before the experiment. The standardized data distribution is shown in Figure 1:

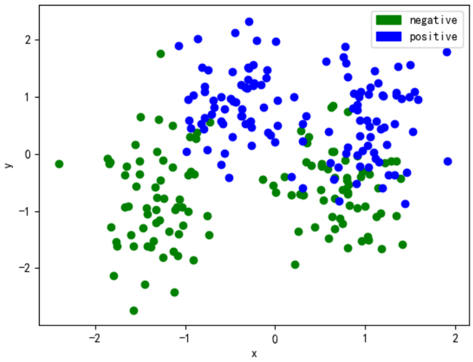


Figure 1. Synthetic binary data set distribution

Four samples are randomly selected from the standardized positive and negative training samples and the corresponding belief rules are constructed. The attribute weights and rule weights of all confidence rules are initialized to 1.0. Figure 2 shows the activation weight distribution of the belief rule corresponding to the negative and positive samples.

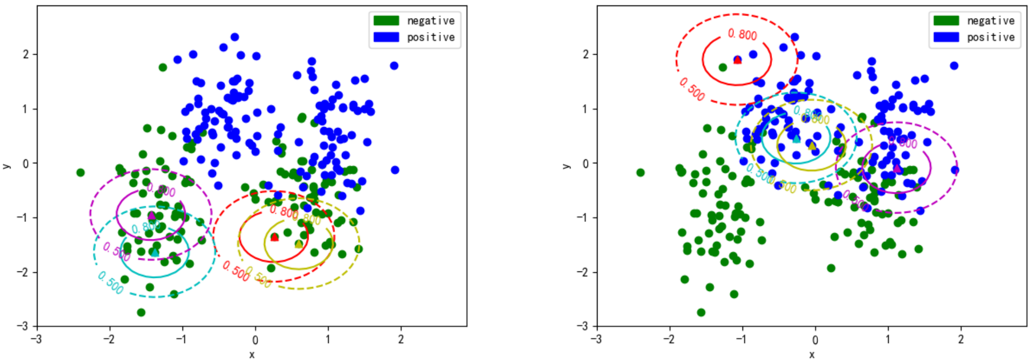


Figure 2. The activation weight distribution of the belief rule generated by the negative and positive sample

In the first step of regularization training, the rule weight is fixed and the penalty coefficients of the antecedent attributes and attribute weights are both set to 0.001. Use cross-entropy as the classification loss function and gradient method as the parameter training method.

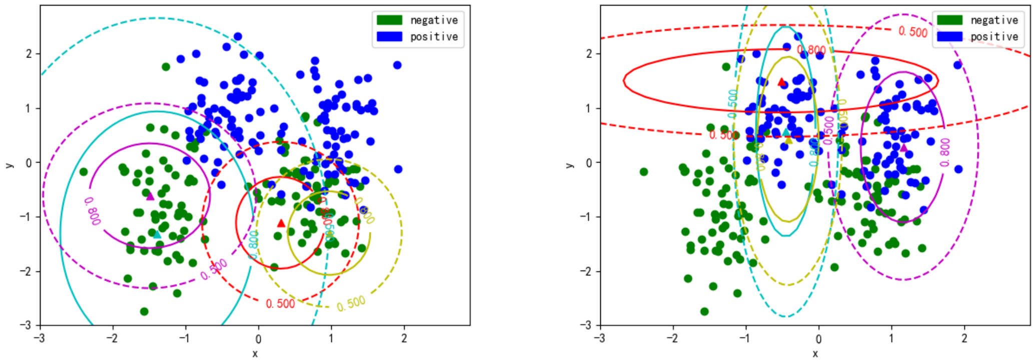


Figure 3. The activation weight distribution of the belief rule generated by the negative and positive sample after the first step training

Figure 3 shows the activation weight distribution of the belief rule corresponding to the positive and negative samples after the first training. It can be found that the regularization method makes the activation weight distribution of each rule approach each other, and the antecedent attribute distribution and attribute weight of the belief rule are also more similar. It provides favourable conditions for further rule reduction.

In the second step of regularization training, fix the antecedent attributes and attribute weights, and set the rule weight penalty coefficient to 0.001. Use the same parameter optimization method for parameter training. Figure 4 shows the rule activation weight distribution corresponding to the positive and negative samples after the training. It can be found that the activation range of the rules with similar activation weight distribution is reduced to negligible, and the remaining rules with a larger activation range cover the activation area of the rule with the reduced activation range.

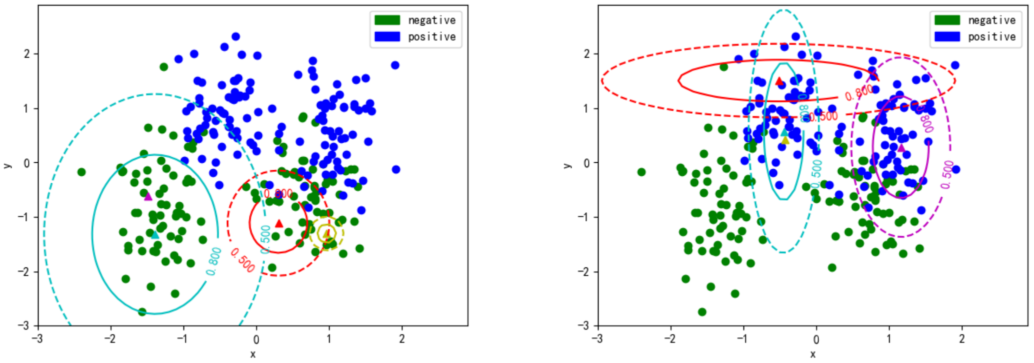


Figure 4. The activation weight distribution of the belief rule generated by the negative and positive sample after the second step training

The rule weights corresponding to the positive and negative samples after regularization training in the second step are listed in Table 1. The maximum rule weight is 0.2287, and the threshold is set to 50% of the maximum rule weight, which is 0.1143 for rule reduction. Rules 1 and 2 for positive samples and rule 4 for negative samples are retained.

Table 1. Rule weight after the second step training

	Rules(from positive samples)				Rules(from negative samples)			
No.	1	2	3	4	1	2	3	4
Weight	0.13530	0.1617	0.0040	0.0879	0.0997	0.6780	0.0327	0.2287

The reduced belief rule base contains three rules, and the third step of training is performed to adjust the reduced belief rule base. The activation weight distribution of the three rules and classification contour maps after training are shown in Figure 5.

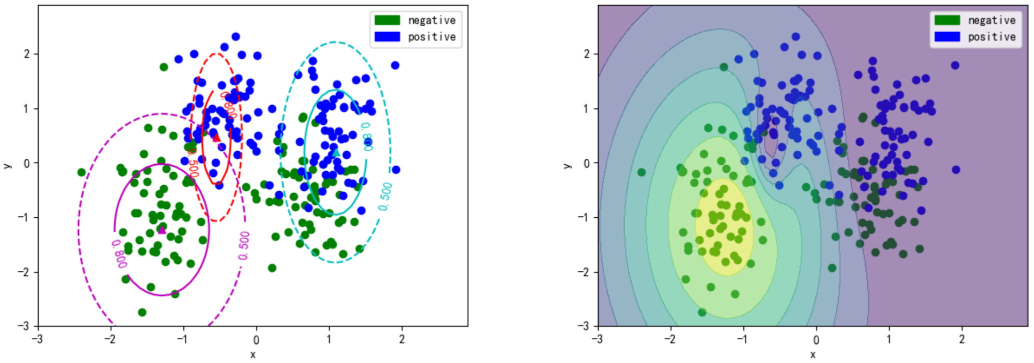


Figure 5. The activation weight distribution of the three rules and classification contour maps after training

4.2. Benchmark Data Sets Experiment

This section uses four UCI classification data sets to verify the reduction performance of the regularization training method on the belief rule base. Table 2 lists the detailed information of the data. Each data set repeats ten independent ten-fold cross-validation experiments to obtain the final results.

Unified data standardization before the experiment, using all training sets to construct the belief rule base. The attribute weight and rule weight are both set to 1.0, and the penalty coefficient of each step of the regularization method is set to 0.001. The experiment compares the reduction size and inference accuracy of the belief rule base under different reduction threshold settings of 10%, 30%, 50%, 70% and compare the number of rules and inference accuracy with the compact belief rule-based classification system using evidence clustering(CBRBCS)[5].

Table 3. Details of the classification datasets

Dataset	#Instances	#Features	#Classes
Iris	150	5	3
Wine	178	14	3
Ecoli	336	8	8
Glass	214	10	6

Table 3 lists the number of rules and the corresponding inference accuracy after the reduction of belief rule base using regularization training method (BRB-R) and the CBRBCS inference system after regularization reduction training on the Iris, Wine, Ecoli, and Glass datasets. Observing the data in the table, we can find that as the reduction threshold increases, the method in this chapter has no significant decrease in inference accuracy and can reduce more rules.

Table 4. Comparison of the results for classification datasets

Method	Aspects	Iris	Wine	Ecoli	Glass
CBRBCS	Accuracy	93.33	94.80	82.62	68.15
	Reduction rate	21.43%	86.86%	17.78%	45.00%
BRB-R(10%)	Accuracy	**95.03**	94.84	84.94	71.61
	Reduction rate	72.15%	77.63%	72.30%	68.79%
BRB-R(30%)	Accuracy	94.83	95.03	85.12	71.37
	Reduction rate	84.88%	87.81%	69.57%	70.52%
BRB-R(50%)	Accuracy	93.39	**95.10**	85.64	70.34
	Reduction rate	90.88%	93.56%	78.26%	74.06%
BRB-R(70%)	Accuracy	93.63	93.14	**86.13**	71.81
	Reduction rate	99.57%	96.56%	**87.37%**	84.16%

5. CONCLUSIONS

This paper proposes a step-by-step regularization parameter training method by constructing the complexity penalty of the reasoning system based on belief rules. Each step of the training is fixed with different parameters, and different penalties are selected to achieve data fitting and rule reduction. The experimental results show that compared with other reduction methods based on belief rule base, this method has a higher simplification rate and higher inference accuracy.

ACKNOWLEDGEMENTS

This work was supported by the Natural Science Foundation of Fujian Province, China (No. 2019J01647), and the Industrial Internet Innovation and Development Project of the Ministry of Industry and Information Technology of China (No. TC19083WB)

REFERENCES

[1] B.D. Ripley "Pattern Recognition and Neural Networks," Cambridge University Press, 1996.

[2] J. B. Yang, J. Liu, J. Wang, "Belief rule-base inference methodology using the evidential reasoning approach-RIMER," IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 36, no. 2, pp. 266-285, 2006.

[3] J. Liu, L. Martinez, A. C. Calzada, "A novel belief rule base representation, generation and its inference methodology," Knowledge-Based Systems, vol. 53, pp. 129-141, 2013.

[4] H. Zou, T. Hastie, "Regularization and variable selection via the elastic net," Journal of the royal statistical society: series B (statistical methodology), vol. 67, no. 2, pp. 301-320, 2005.

[5] L. Jiao, X. Geng, Q. Pan. "Compact belief rule base learning for classification with evidential clustering," Entropy, vol. 21, no. 5, pp. 443, 2019.

[6] A. P. Dempster, "A generalization of bayesian inference," Journal of the Royal Statistical Society: Series B (Methodological), vol. 30, no. 2, pp. 205-232, 1968.

[7] G. Shafer, A. F. M. Smith, "A mathematical theory of evidence," Biometrics, vol. 32, no. 3, pp. 703, 1976.

[8] L. Jiao, T. Denoeux, and Q. Pan, "A hybrid belief rule-based classification system based on uncertain training data and expert knowledge," IEEE Trans. Syst., Man, Cybern. Syst., vol. 46, no. 12, pp. 1711–1723, 2016.

[9] L. Jiao, Q. Pan, T. Denoeux, Y. Liang, and X. Feng, "Belief rule based classification system: Extension of FRBCS in belief functions framework," Inform. Sciences, vol. 309, no. 1, pp. 26–49, 2015.

AUTHOR

Yu Guan received the B.S. degree from Fuzhou University, in 2018, where he is currently pursuing the master's degree. His research interests include intelligent decision technology, rule-based inference, big data analysis, and machine learning.

MACHINE LEARNING ALGORITHM FOR NLOS MILLIMETER WAVE IN 5G V2X COMMUNICATION

Deepika Mohan[1], G. G. Md. Nawaz Ali[2] and Peter Han Joo Chong[1]

[1]Department of Electrical and Electronics Engineering,
Auckland University of Technology, Auckland 1010, New Zealand
[2]Department of Applied Computer Science,
University of Charleston, WV 25304, USA

ABSTRACT

The 5G vehicle-to-everything (V2X) communication for autonomous and semi-autonomous driving utilizes the wireless technology for communication and the Millimeter Wave bands are widely implemented in this kind of vehicular network application. The main purpose of this paper is to broadcast the messages from the mmWave Base Station to vehicles at LOS (Line-of-sight) and NLOS (Non-LOS). Relay using Machine Learning (RML) algorithm is formulated to train the mmBS for identifying the blockages within its coverage area and broadcast the messages to the vehicles at NLOS using a LOS nodes as a relay. The transmission of information is faster with higher throughput and it covers a wider bandwidth which is reused, therefore when performing machine learning within the coverage area of mmBS most of the vehicles in NLOS can be benefited. A unique method of relay mechanism combined with machine learning is proposed to communicate with mobile nodes at NLOS.

KEYWORDS

5G, Millimeter Wave, Machine Learning, Relay, V2X communication

1. INTRODUCTION

Significant advancement has been made in the field of vehicular communication in recent years. Information about live traffic updates, warnings, safety messages and even more information are exchanged through the communication node among vehicles and Road side units (RSU) [12]. All these communications between vehicles and network are done as the vehicles communicates through cell phone towers for information about traffic and routes. Every mobile node communicates with each other to avoid accidents and to calculate the speed of their neighbours, hence from this it is clear that the vehicular communication and mobility are approaching for a new era. The 5G provides a higher speed with lower latency which creates a revolution in the digital world of wireless technology and the advancements in 5G has benefited in the improvements in vehicular networks for autonomous driving [11]. Ultra-high reliability is provided for many applications that utilize 5G as it involves new technology with a usage of high frequency and antenna improvement. The vehicle-to-everything (V2X) plays an important role in improving high vehicle utilization, supporting accident free transportation thus providing zero emission vehicles which are more efficient [16]. The major use of 5G technology in future is V2X communication where the vehicles communicate with its surroundings to get information from outside world. For basic security and non-safety applications the V2X technology is operated below 6 GHz [26]. Many V2X applications make use of the Millimeter Wave

Natarajan Meghanathan et al. (Eds): CSE, BMLI, EDU, AI & FL, CCNET - 2020
pp. 63-76, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101706

(mmWave) band as it functions below 6GHz with higher data rate, increased network capacity and bandwidth. The major drawback mmWave faces is its coverage area as it suffers a higher penetration loss hence it cannot transmit signal when obstructed by buildings, therefore mmWave cannot send information to vehicles in NLOS. The mmWave coverage area can be increased either by deploying mmWave Base Station (mmBS) or designing an antenna in such a way which will transmit information to vehicles using radio frequency [13]. But a suitable solution for overcoming this drawback is by using Machine Learning (ML); through which mmBS understands and learns about its environment and transmits data to vehicles in NLOS range [6]. Even though mmWave have more advantages in the communication field, it suffers a big loss due to shadowing, blockages, fast fading and path loss [5]. More number of approaches is proposed for these drawbacks but for predicting blockages and transmitting messages is still a challenge in mmWave, which will be focused in this paper. Therefore in a particular coverage area if the broadcast information such as basic safety messages (BSM) are transmitted to all vehicles then only the vehicles at LOS receives the information but other vehicles in NLOS cannot receive the message from mmBS. Many researches are going on to overcome this problem in mmWave. One solution is by making use of ML so that all the vehicles can communicate with the mmBS without any interruptions as ML trains the base station to identify blockages and transmits data to NLOS nodes in terms of automatic and semi-autonomous driving which is the main motive of this paper. From the best of our knowledge none of the existing works proposed this idea and shown any result. Hence the organizational structure of this paper is as follows, this paper introduced the functions and applications of mmWave and Machine Learning in section 1. Section 2 discusses the related works which helps the reader to understand the different methodologies used in mmWave and V2X Communication using ML. The System Architecture and its work functions are summarized in Section 3 and Section 4 discusses the RML algorithm and the experimental results. Finally Section 5 concludes the overall work along with the limitations of RML and its future developments.

2. RELATED WORKS

For improving the efficiency of road safety, traffic and infotainment options in vehicles, the V2X communication has been identified as the most suitable technology between road infrastructure and vehicles [25]. The V2X provides better QoS and coverage when compared to the other short-range dedicated communication. The mobile nodes handle more amount of data due to crowding and video streaming on traffic and entertainment applications with frequent internet usage [7]. Hence for this each automatic application requires good reliability with reduced latency which is embedded within individual packet preference level. In order to lower the pressure of complexity in vehicular applications, the ML is developed for V2V communication in which every transmitter acts as an agent whose decision is based on the observation from the surroundings. Even though ML and 5G are two different fields but on combining both together better results are obtained for solving various application problems. The complexity of design and procedures in V2X can be handled by data learning, modifying and replacing the rule list with ML routines which learns readily from previous stored data. The problems in coverage probability on spectrum location are differentiated into small parts by using Poisson Point Process and Cox Point process [10]. For the vehicular Fog computing methodology by using Q-learning provides higher reward as the system reaches a stable state faster when compared to other existing methods [27]. Considering the expected high traffic demand of vehicular networks, the spectrum resources of mmWave are used more efficiently for spectrum sharing mechanism. Therefore a model based on sensing is developed on mmWave spectrum for vehicular networks [24]. For constructing this model innovatively, an algorithm based on beam alignment was proposed using the temporal correlation. During the implementation of this method, the sensing outputs generated in different sensing time are affected by various environmental conditions.

On utilizing the capable solutions like the beam recovery, tracking and alignment, the higher data rate of 10 to 100 mbps and lower latency of 10ms to 100 ms in mmWave can be used by vehicular communication [19]. In order to overcome the limitations of hardware in mmWave, the beam forming mechanism is proposed in which the antenna weight is restricted to phase shifts of lower resolution and the best transreceiver pairs are selected [21]. The training process helps to detect the capable beam pairs that increase the standard of the link using numerical algorithm [22]. The ML designs can utilize the uplink signal gathered at the base station to learn about the mapping structure related to the outdoor scenario [9]. Even though there are many solutions supporting mmWave communication but the actual problem lies in the system design [8]. It suffers a higher penetration loss when blocked by obstacles and hence results in inaccurate beam selection. There are some methods that are time consuming and unscalable for 5G implementation but these approaches cannot detect regular pattern of traffic and blockages. An approach of contexted multi- armed bandit was developed that works on the network information based on the vehicles arrival direction which uses the mmBS to automatically learn and understand from its environment. Many NLOS scenarios have dissimilar amplitude giving rise to functions consisting of various local optima and hence by improving channels of NLOS nodes certain changes and extensions are proposed in Nelder-Mead beam based training technique [14]. Therefore the starting point, of the training mechanism is derived as in Equation (1) as,

$$P_{1,nc}^{2} = (p_{1,nc}^{2} , q_{1,nc}^{2}), n_c \in [1, N_c] \tag{1}$$

$$p_{1,nc}^{2} = 2 \text{ x } (2 \text{ x } P_{opt,nc}^{1} - 1) - 1 \tag{2}$$

$$q_{1,nc}^{2} = 2 \text{ x } (2 \text{ x } q_{opt,nc}^{1} - 1) - 1 \tag{3}$$

where N_c, is the output optima which is defined as,$P_{opt,nc}^{2} = (p_{opt,nc}^{2} , q_{opt,nc}^{2})$, declared in descending order which is derived from Equation (2) and (3). The ML algorithm based on Manhattan Poisson line process for street model is defined as a better solution for differentiating blockages in mmWave urban networks [17]. In order to analyze the blocking assistance the following equations are calculated for simulating triangles as in Equation (4).

$$\frac{B_C^m - \Omega_m - \frac{w_v^2}{2}}{B_C^m} = \frac{H_{BS} - H_L}{H_{BS} - H_S} \tag{4}$$

Therefore from Equation (5), the final value of blocking assistance is derived as,

$$B_C^m = \frac{\Omega_m + \frac{W_v^L}{2}}{[\frac{H_{BS} - H_L}{H_{BS} - H_S}]} \tag{5}$$

For a group of vehicles travelling on the same lane, some mobile nodes experience fading if the length of the group is long and hence a relay option is best suited to overcome this path loss [3]. The most efficient method for penetration loss problems in mmWave is to utilize ML. A relay mechanism based on soft-information was delivered for vehicular networks which use the estimate and forward strategy in which the minimum mean square error of the signal received is obtained with less complexity which achieves a good trade off among vehicular networks [1]. The DSRC and LTE-V2V nodes use the first relay system which provides a performance gain of 91% with respect to the communication distances [2]. For signal blocking in a direct link a deep learning model was formulated to solve the challenges in optimal relay selection [4]. This method of prediction is utilized in the proposed research in terms of relay selection but using RL.

3. SYSTEM MODEL

Since more number of solution were formulated for sending information to NLOS vehicles, these solutions may sometimes make the system complex due to increased system specification in mmWave links. In this paper, we propose Relay using Machine Learning (RML) algorithm to train the mmBS for identifying the blockages and to learn about the best direction and location for broadcasting the messages to vehicles at NLOS. A relay mechanism is used for transmitting information to vehicles that cannot communicate with the mmBS. This proposed ML algorithm chooses a relay that is capable of transmitting data to all the vehicles at NLOS with strong signal strength and minimum congestion. This kind of Relay mechanism utilizing ML is new in mmWave 5G V2X communication as the BS learns about its environment through observations and transmit information to vehicles at NLOS through a single hop relay depending on the position of blockages. Instead of deploying more number of BS to broadcast messages, it is easier for formulating ML on mmBS and make it learn about its surroundings in order to transmit data accordingly which saves much time, cost and energy. An urban scenario is considered for the research where there are rapid changes in traffic and the network is often interrupted by permanent and temporary blockages. Fig. 1 shows the design scenario of the proposed research. The mmBS which is deployed has a coverage range of 300 meters and the vehicles on different lanes are subjected to blockage effect. The vehicles closer to mmBS are free from blockages as they are at direct LOS with the BS. But the vehicles at a distance are obstructed by blockages is the NLOS node that does not receive any data from the mmBS.

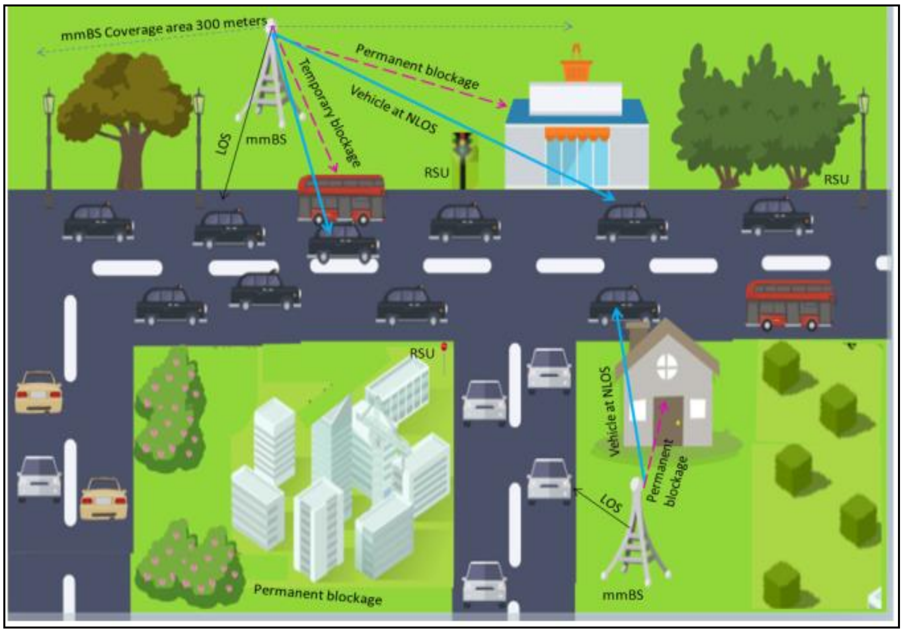


Fig. 1. Architectural design scenario.

3.1. Identification of Blockages

As the mmWave is sensitive to blockages, it leads to severe shadowing and produces various characteristic of path loss between LOS and NLOS links. In order to reduce this effect, the blockage is identified [18]. Using the RML the mmWave identifies the blockages. Initially the mmBS broadcasts signal through antenna and the blockages are identified if the radiation pattern is reflected back from its path. After a deep research analysis on the temporary blockage an assumption is made in this paper that the maximum height of vehicle is 16 feet, and only large vehicles like bus, truck and semi-trucks are considered as temporary blockages. Hence a threshold value of 16 feet is taken to differentiate blockages as permanent or temporary. Once when the radiation pattern is reflected below the threshold the mmBS identifies it as temporary blockage and if the reflection occurs greater than the threshold it determines it as permanent blockage. In the case of temporary blockages the obstacle itself acts as a relay to transmit information to NLOS nodes, but for permanent blockages the BS calculates the distance between itself and the barrier along with the direction and hence stores the detail for future use. This stored data on the position of blockages is updated at regular intervals of time which makes the mmBS understand the best location and position to communicate with vehicles at NLOS. Fig. 2 shows the distance estimation for identifying the blockage location.

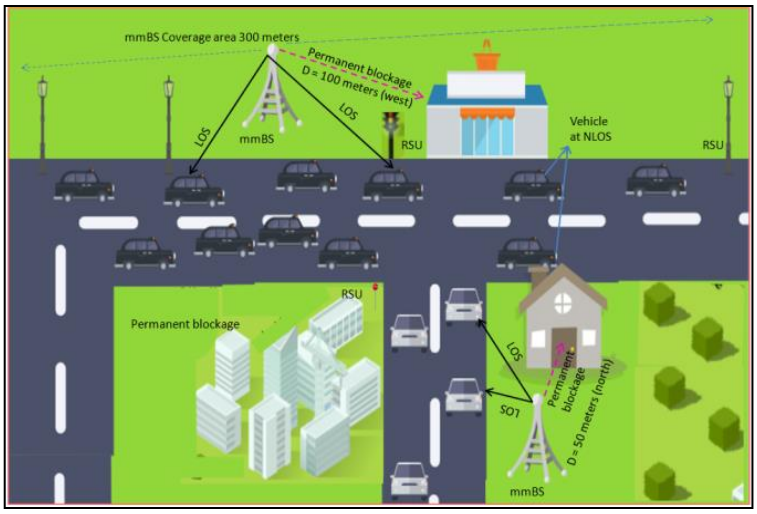


Fig. 2. Identification of the location of blockage based on distance estimation.

3.2. Efficient Relay Selection

The mmBS identifies the location of LOS and NLOS vehicles in its coverage area through V2V communication. Once when a LOS vehicle communicate with the BS, all the data about the LOS is gathered by the mmBS, hence the BS confirms the position of NLOS node and chooses an efficient relay for transmission of information in the following manner [20]. In handoff, a BS targets a particular mobile node and manages data delivery through relay technique. From Equation (6), the handoff algorithm is formulated at time t, with N, number of flow at m (constant) and hence the estimation state is given as,

$$N_n = N + N_r - N_t - m \tag{6}$$

$$N_r = t_s * \frac{numberofarrival}{T_e} \tag{7}$$

$$N_t = t_s * \frac{numberofdeparture}{T_e} \tag{8}$$

where $t_s$the next sampling time and T_e is the estimated time as in Equation (7) and (8). Using RML the mmBS chooses the suitable relay based on the following conditions,

a. Since most vehicles uses GPS receiver system, each node in the network gets its position in real time.

b. The link metric is calculated by exchanging information about the vehicles position through control messages which is stated in Equation (9) as,

$$\text{Link metric} = \ln[\,1 - \textstyle\prod_{V_{Ln} \in Path}(1 - P_{VLn})] \qquad (9)$$

c. Based on the information exchanged the Relay calculation is performed with the metrics choosing Relay path as calculated in Equation (10).

$$\text{Path} = \{V_{L1}, V_{L2}, \ldots \ldots . V_{Ln}\} \qquad (10)$$

Here $V_{L1}, V_{L2}, \ldots \ldots . V_{Ln}$ are different suitable LOS relay nodes. Hence the RML algorithm aims to choose a relay based on distance and position which is defined in Equation (11), (12) and (13).

$$P_{path} = 1 - \textstyle\prod_{V_{Ln} \in Path}(1 - P_{VLn}) \qquad (11)$$

$$P_{path} = \ln[1 - \textstyle\prod_{V_{Ln} \in Path}(1 - P_{VLn})] \qquad (12)$$

$$\textstyle\sum_{V_{Ln} \in Path} \ln P_{VLn} = \sum_{V_{Ln} \in Path} LinkMetric_{VLn} \qquad (13)$$

4. RML ALGORITHM

Machine learning algorithm is used in the beam to direct the messages and performs relay mechanism so that all the vehicles make use of the mmWave technology. On performing Beam selection or switching operations there is fewer guarantee that all the data transmitted reaches the mobile node without any delay but by using relay mechanism the data reaches the nodes at a faster rate without any packet loss. In this scenario all the network operations and functions are built on the same data foundations which will make the system simpler [15]. The advantage of RML is that the solution to an issue can be learnt directly from the data generated as this algorithm runs on the mmBS which broadcasts information to the nodes in its coverage area. The background of RML algorithm is the Reinforcement Learning which is formulated based on the correlation among the location of NLOS vehicles and the action taken by the mmBS based on Relay mechanism is the key for future decision.

4.1. Detailed Description

In detail, our proposed RML algorithm work function is as follows (see Algorithm 1): First during the initialization process the trained model accepts the input broadcast messages B_{mn}, height, distance and V_h. The mmBs coverage area is assigned to 300 meters and hence the system observes the environment based on regular traffic patterns. The model checks for permanent and temporary blockages based on the estimated threshold ε.

Algorithm 1 Pseudocode of RML algorithm.

```
Input: $B_{mn}$, h, D, Ref, angle (θ) and $V_h$
Initialize vehicles: $V_h = Vh_1, Vh_2, \ldots \ldots \ldots \ldots Vh_n$
Initialize LOS nodes: $V_L = V_{L1}, V_{L2}, \ldots \ldots \ldots \ldots \ldots . V_{Ln}$
Initialize NLOS nodes: $V_{NL} = V_{NL1}, V_{NL2}, \ldots \ldots \ldots \ldots \ldots . V_{NLn}$
Output: Ar (antenna radiation), $R_t$
Initialize Model = model (input, output)
State = Observation
```

```
8: **if** State random.rand () <$\varepsilon$
9: Select action temporary blockage
10: **else** Take action >$\varepsilon$
11: Action = D + Ref + θ //Direction of permanent blockage
12: Return action
13: **end if**
14: Exchange$D_v, V_L, V_{NL}$ // vehicles exchange information about its distance and information about its neighbour nodes.
15: Calculate $V_D = D_r - D_v$ // distance between the vehicle and the blockage is calculated
16: State S = $V_D + +$ // update the state
17: Select Relay node $D_v \leq D_r$
18: **if** $(D_v(V_{L1}) \leq D_v(V_{NL1}))$ // calculates the distance between vehicle1 (LOS) with vehicle1 (NLOS)
19: Transmit Bm1 = $V_{NL1}$ // Send received broadcast message from vehicle1 (LOS) to vehicle1 (NLOS)
20: **else** $(D_v(V_{L2}) \leq D_v(V_{NL1}))$ // calculates the distance between vehicle2 (LOS) with vehicle1 (NLOS)
21: **end if**
22: Update $D_r = D_v$ // update the best distance between vehicles at LOS to NLOS based on the direction of permanent blockage.
23: S = np.array (a[0]) // Select States
24: S+1 (new state) = np.array (a[n]) // Select new States for replay
25: Q = self.model predict (states)
26: Q-new = self.model predict (new_states)
27: Replay_distance = len (replay)
28: Target = Q[i]
29: Target [action_r] = reward_r
30: **while True,**
31: Total reward = reward
32: S = S+1
33. State_s, action_a, reward_r, done_r = =replay [i] // To construct training set
34: **for** each $D_r$ // identify the best suitable distance of data transmission
35: Repeat Step 2
36: **end for**
37: $B_m$= $V_L$ = $V_{NL}$ // All the vehicles at both LOS and NLOS receives broadcast information from the mmBS
38: **end**
39: **Return**
```

When ε is higher than 16 then the action is taken with respect to distance D, Ref and θ to estimate the location of the blockage. In the mean time, the vehicles V_L(LOS) and V_{NL}(NLOS) exchange information about each other and the overall data is termed as V_h. The mmBS calculates the distance between D_r and D_v to predict V_D and hence the state S is ready for update. The Relay node is selected based on the distance between D_v and D_r, hence initially the distance between the LOS node V_{L1} and NLOS node $V_{NL1}$is calculated and if the distance are near then B_{m1} is transmitted to V_{NL1} else the next LOS node V_{L2} is compared until a suitable relay is selected. In order to transmit data to V_{NL1} the target Q[i] is chosen which is the relay node and the action_r is performed and for every action a reward_r is obtained. Finally B_{mn} is transmitted to V_{NL} through $V_L$with a new state S+1. The time complexity of RML algorithm measures the speed at which this algorithm performs its operations for the given input size n. This proposed RML algorithm has two types of time complexity: training and run time complexity. When the amount of data is larger, the time complexity for training the mmBS is represented as in Equation (14).

$$\text{Training time complexity} = O\ (n*\log\ (n)*d*k) \qquad (14)$$

where n is the number of training examples which is chosen from the environment using RL and d is the directional dimensionality of blockages, and k is the number of available LOS relay nodes. The complexity in run time is expressed with respect to the traffic pattern and the neighbour nodes which is represented in Equation 15 as,

$$\text{Run time complexity} = O\ (\text{depth of traffic}*k) \qquad (15)$$

where the depth of traffic denotes the overall nodes in the coverage area. In order to decrease the complexity of vehicles the action space is considered when a maximum update is reached the model resets itself without erasing the data from its memory.

4.2. Simulation Setup

The scenario is simulated for the terrain area with dimension 300m X 300m where the transmission range of mobile nodes and eNB are 100m. Table 1 briefs the important parameters used for simulation.

Table 1. Simulation parameters.

Simulation Parameter	Values
Simulator tool	ns-3
Machine Learning	Reinforcement Learning
Terrain Dimension	300m x 300m
No. of eNB's	1
Position of eNB (x, y)	(55,55) (115,115) (175,175) (235,235) (295,295) for blockages (2, 4, 6, 8, 10)
Mobility model of eNB	Constant position model
No. of vehicles	10, 20, 30, 40, 50
Mobility model for vehicles	Random waypoint model
Speed of vehicles	15 Min = 0.1 Max = 15
No. of Relay	1
No. of Blockages	2, 4, 6, 8, 10
Propagation Model	3GPP Propagation model
Building width	X = 50m, Y = 50m
Building height	Z = 10m (32 feet)
Channel Model	3GPP Channel model
Transmission radius	100m
Bandwidth of mmWave signal	200 MHz
Transmit power	30 dBm
Packet size of messages	1024 bytes
Data Rate	100 Gb/s
MTU	1500 bytes
Interpacket arrival	200ms
Queue size	100
Simulation Time	50Seconds

The number of blockages is varied within the coverage area based on the location of eNB which is deployed using constant position model. The mobile nodes are placed in the simulation setup using the Random waypoint model in which the speed of vehicle is set as 15 (constant) with a pause time of 5 seconds after which the vehicle takes new direction. An assumption is made in this simulation scenario that when a vehicle hits the blockage or the borders of the coverage area it bounces back and takes a new position as the collision is ignored in this setup because the main concentration of our work is relay selection and transmission of information to vehicles at NLOS. Here the maximum number of obstacles are given as 10 which is designed using the 3 GPP building model with a height of the obstacle, z = 10 meters. The simulation time is set as 50 seconds but the actual simulation runs for 45 minutes which is the time taken for training and run time purposes. Fig. 3 shows the design of the simulated network. For the location of permanent blockages a 3GPP propagation model is used to calculate the co-ordinates of x and y.

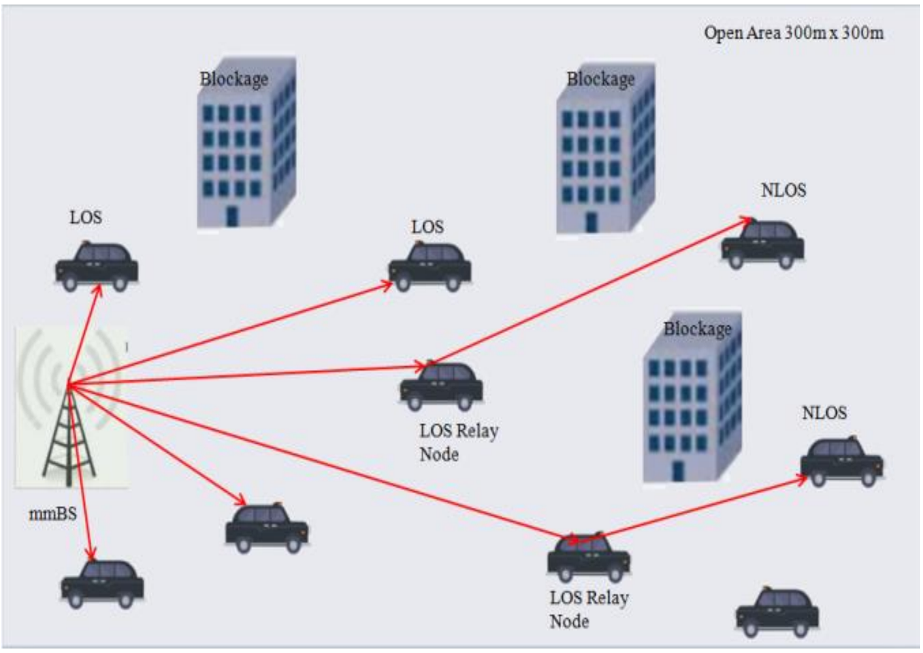


Fig. 3. Layout of the Simulated Network.

4.3. Experimental Results and Discussions

From the formulated RML algorithm the simulation is done using the specifications of the parameters. The results are obtained on comparing the simulation scenario using RML and without using RML. For the model without RML, the simulation is done with normal mmWave design which is prone to blockages and penetration loss. Hence the simulation is performed under 3 stages and the results are obtained in terms of throughput, latency and PDR. First the mobile nodes are kept constant at 20 and the blockages are varied from 2, 4, 6, 8 and 10 for the simulation time of 50 seconds. Fig. 4 shows that the average latency is less than 0.05 ms for the simulation that uses RML with 10 obstacles but the latency increases to 0.7 ms for the model without using RML.

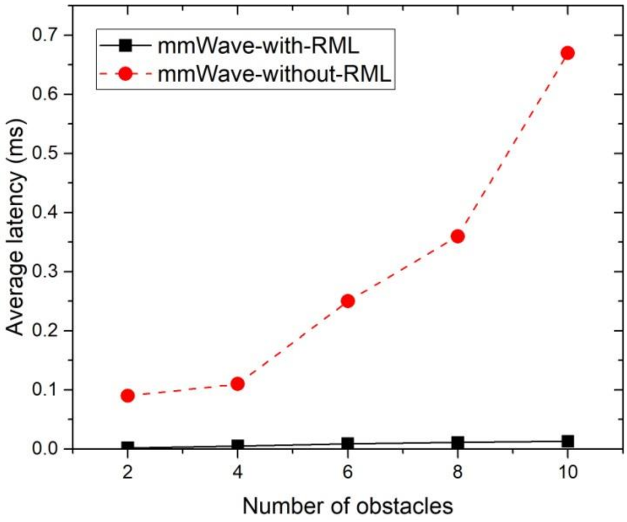


Fig. 4. Average latency comparision under different number of blockages for mmWave with and without RML.

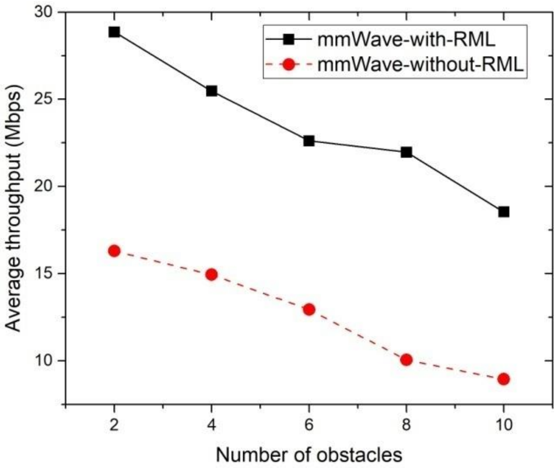


Fig. 5. Average throughput evaluation for mmWave with and without RML.

The throughput is calculated by varying the blockages and from Fig. 5 the throughput is 27.44 mbps when using RML but it is only 18 mbps when RML is not used. Hence the performance of the model without RML is low. Fig. 6 shows the result of PDR for constant mobile node where the performance is good as it reached 100% when using RML in mmWave than without using RML.

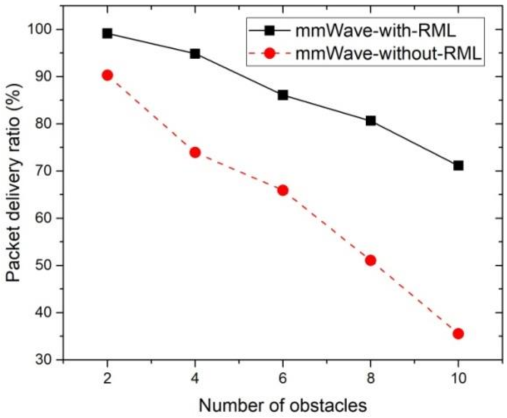


Fig. 6. PDR performance of mmWave through RML and without RML for constant mobile nodes.

In the next stage of the simulation the blockages are made constant at 10 and the vehicles are varied and from the result obtained in Fig. 7, it is clear that for maximum blockages the latency output using RML is at 0.1ms but for model without using RML it goes beyond 1.6ms.

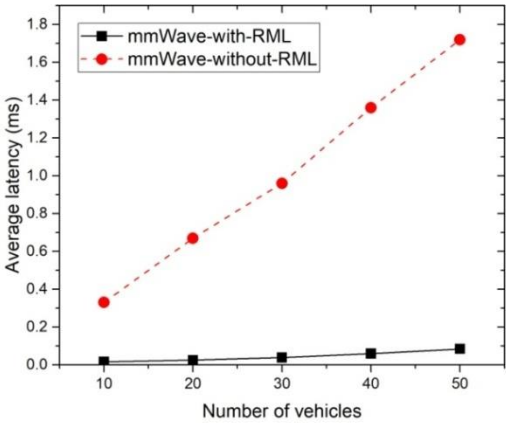


Fig. 7. Average Latency evaluation for Blockage = 10 by varying vehicles in mmWave with and without RML.

The Fig. 8 shows throughput result in which the model using RML outperforms the other.

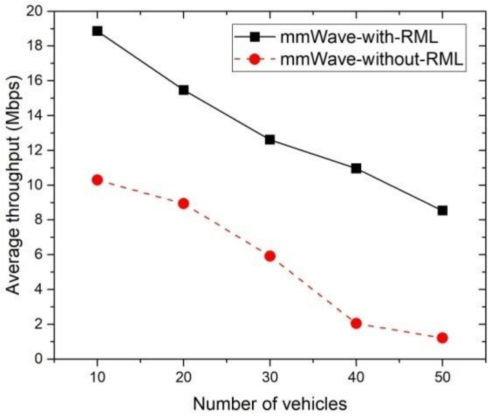


Fig. 8. Average throughput comparision graph for number of vehicles at blockage = 10 in mmWave with and without RML.

For the number of blockages as 10 and vehicles as 10 the PDR is 75% for the scenario using RML where as it is only 50% for the output without using RML as in Fig. 9. From the obtained result it is clear that on using RML as the blockages increases the throughput and the PDR also rises with a fall in latency, therefore the proposed RML confirms that this kind of solution can have the ability to solve real-world applications.

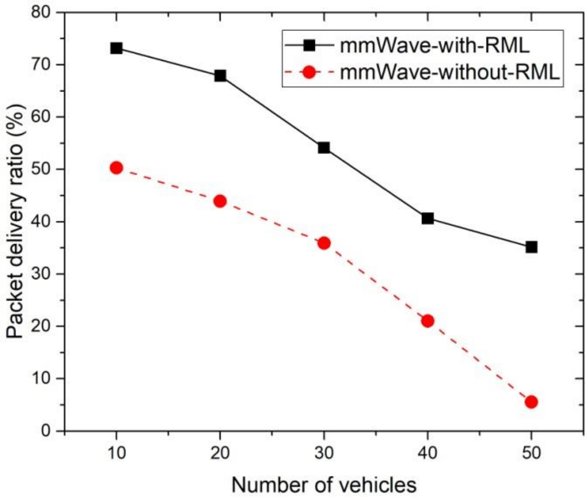


Fig. 9. PDR perfromance for mmWave with and without RML.

5. CONCLUSION

In this paper, we address the drawbacks in mmWave due to blockages from buildings and vehicles in real-time urban scenario. To this aim, we propose RML, a single hop relay mechanism based on Reinforcement learning which trains the mmBS to identify the location of blockages and to select a relay node based on the distance calculation from the BS and to broadcast the data to NLOS vehicles using LOS relay node. Using RML 1) the mmBS learns about its surroundings through continuous observations 2) the mmBS becomes efficient to predict the relay node earlier when it identifies vehicles at NLOS. The RML achieves 75% of the PDR when the blockage is at its maximum and the results demonstrate the reliability of using Machine Learning in 5G V2X applications. Even though the RML performance is good, it lacks accuracy when it sees rapid changes in traffic pattern. When more number of NLOS nodes is present a single relay may not be suitable to transmit information to all of them especially when the coverage range and the transmission distance is higher. Hence our RML model can be further extended by using Deep RL or Q-learning to accurately predict the temporary blockages and to efficiently select the relay nodes for transmitting data to NLOS vehicles based on the daily changes in traffic pattern especially during the peak hours.

REFERENCES

[1] Yingchun Wang, Tingting Yao, Fenghua Zhu , Gang Xiong, Dongpu Cao, Fei-Yue Wang, (2017) “Relaying Algorithm Based On Soft Estimated Information For Cooperative V2X Networks,” International Conference on Intelligent Transportation Systems (ITSC), pp. 509-514.

[2] Byungjun Kim, Seongwon Kim,Hoyoung Yoon,Sunwook Hwang,M. Xavier Punithan,Byeong Rim Jo,Sunghyun Choi, (2019) “Nearest-First: Efficient Relaying Scheme in Heterogeneous V2V Communication Environments,” in IEEE Access, vol. 7, pp. 23615-23627.

[3] Jingwei Fu, Gang Wu, Ran Li, (2020) “Performance Analysis of Sidelink Relay in SCMA-based Multicasting for Platooning in V2X,” IEEE International Conference on Communications Workshops (ICC Workshops).

[4] A. Abdelreheem, O. A. Omer, H. Esmaiel, U. S. Mohamed, (2019) "Deep Learning-Based Relay Selection In D2D Millimeter Wave Communications," International Conference on Computer and Information Sciences (ICCIS), pp. 1-5.

[5] Peyman Siyari, Hanif Rahbari, Marwan Krunz, (2019) "Machine Learning for Intelligent and Agile Wireless Communications," IEEE Journal on Selected Areas in Communications (JSAC) - Special Issue on Machine Learning in Wireless Communications, vol. 37, no. 11, pp. 2544-2558.

[6] Marwan Krunz, Mingjie Feng, (2019) "Protocols, Adaptation, and Spectrum Allocation for 5G Millimeter-wave Systems," wireless communication and networking laboratory.

[7] X. Zhang, M. Peng, S. Yan, Y. Sun, (2020) "Deep-Reinforcement-Learning-Based Mode Selection and Resource Allocation for Cellular V2X Communications," IEEE Internet of Things Journal, vol. 7, no. 7, pp. 6380-6391.

[8] G. H. Sim, S. Klos, A. Asadi, A. Klein, M. Hollick, (2018) "An Online Context-Aware Machine Learning Algorithm for 5G mmWave Vehicular Communications," IEEE/ACM Transactions on Networking, vol. 26, no. 6, pp. 2487-2500.

[9] Liang Xian, Alexander Maltsev, Gia Khanh Tran, Hiroaki Ogawa, Kim Mahler, Robert W. Heath Jr, (2017) "Where, When, and How mmWave is Used in 5G and Beyond," IEICE Transactions on Electronics.

[10] S. Yan, X. Zhang, H. Xiang, W. Wu, (2019) "Joint Access Mode Selection and Spectrum Allocation for Fog Computing Based Vehicular Networks," IEEE Access Fog Radio Access Networks (F-RANs) for 5G: Recent Advances and Future Trends, vol. 7, pp. 17725-17735.

[11] H. Ullah, N. Gopalakrishnan Nair, A. Moore, C. Nugent, P. Muschamp, M. Cuevas, (2019) "5G Communication: An Overview of Vehicle-to-Everything, Drones, and Healthcare Use-Cases" IEEE Access Roadmap to 5G: Rising to the Challenge, vol. 7, pp. 37251-37268.

[12] Kshitiz Shrestha, (2019) "5G: The Future of Improved Road Safety and Autonomous Vehicle," Metropolia University of Applied Sciences.

[13] S. K. Agrawal, K. Sharma, (2016) "5G millimeter wave communications," International Conference on Computing for Sustainable Global Development (INDIACom), pp. 3630-3634.

[14] C. R. Storck, F. Duarte-Figureueiredo, (2018) "5G V2X Ecosystem Providing Entertainment on Board Using MmWave Communications," IEEE 10th Latin-American Conference on Communications (LATINCOM), pp. 1-6.

[15] Tianyu Wang, Shaowei Wang, Zhi-Hua Zhou, (2019) " Machine Learning for 5G and Beyond: From Model- Based to Data-Driven Mobile Wireless Networks," Emerging Technologies & Applications.

[16] Y. Wang, K. Venugopal, A. F. Molisch, R. W. Heath, (2017) "Blockage and Coverage Analysis with mmWave Cross streetBSs Near Urban Intersections, " IEEE International Conference on Communications (ICC), pp. 1-6.

[17] Grants CNS-1702957and CNS-1320664, and by the Wireless Engineering Research and Education Centre Volume 21, Issue 3.

[18] H. Elkotby, M. Vu, (2017) "A Probabilistic Interference Distribution Model Encompassing Cellular LOS and NLOS mmwave Propagation," IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 738-742.

[19] W. Khawaja, O. Ozdemir, Y. Yapici, F. Erden, I. Guvenc, (2020) "Coverage Enhancement for NLOS mmWave Links Using Passive Reflectors," IEEE Open Journal of the Communications Society, vol. 1, pp. 263-281.

[20] X. Wang, E. Turgut, M. C. Gursoy, (2019) "Coverage in Downlink Heterogeneous mmWave Cellular Networks With User-Centric Small Cell Deployment," IEEE Transactions on Vehicular Technology, vol. 68, no. 4, pp. 3513-3533.

[21] K. Zeman, M. Stusek, P. Masek, J. Hosek, (2018) "Improved NLOS Propagation Models for Wireless Communication in mmWave bands,"International Conference on Localization and GNSS (ICL-GNSS), pp. 1-6.

[22] Yekaterina Sadovaya, Dmitrii Solomitckii, Wei Mao, Oner Orhan, Hosein Nikopour, Shilpa Talwar, (2020) "Ray-Based Modeling of Directional Millimeter-Wave V2V Transmissions in Highway Scenarios," IEEE Access, vol. 8, pp. 54482-54493.

[23] H. Zhang, C. Guo, (2019) "Beam Alignment-Based mmWave Spectrum Sensing in Cognitive Vehicular Networks," IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 1-5.

[24] Rakesh R T, Debarati Sen, Goutam Das, (2018)“On Bounds of Spectral Efficiency of Optimally Beamformed NLOS Millimeter-Wave Links,” IEEE Transactions on Vehicular Technology, vol. 67, no. 4.
[25] Fabio Arena, Giovanni Pau, Alessandro Severino, (2019) “V2X Communications Applied to Safetyof Pedestrians and Vehicles,” Journal of Sensor and Actuator Networks.
[26] S. Zhou, Y. Sun, Z. Jiang, Z. Niu, (2020) “Exploiting Moving Intelligence: Delay-Optimized Computation Offloading in Vehicular Fog Networks,” IEEE Communications Magazine, vol. 57, no. 5, pp. 49-55.

AUTHORS

Deepika Mohan is currently a Master Student at the Department of Electrical and Electronics Engineering, Auckland University of Technology, NZ. Deepika received her previous Master Degree (Communication systems) in India 2011. Her research interest is Machine Learning, mmWave, vehicular communication and 5G.

G. G. Md. Nawaz Ali (Member IEEE) is working as an Assistant Professor with the Department of Applied Computer Science of University of Charleston, WV, USA. Prior to joining to UCWV, he was working as a post doctoral research fellow with the Department of Automotive Engineering of Clemson University, SC, USA (March 2018 – July 2019). His research interests are in the areas of Vehicular Adhoc networks, Scheduling and Broadcasting and data analytics.

Peter Han Joo Chong (Senior member IEEE) is an Associate Head of School (Research) and a Head of Department of Electrical and Electronic Engineering at Auckland University of Technology, Auckland, New Zealand. He received his Ph.D. degrees in Electrical and Computer Engineering from the University of British Columbia, Canada. His research interests are in the areas of wireless/mobile communications systems including radio resource management, multiple access, MANETs/VANETs, green radio networks and 5G-V2X networks. He has published over 200 journal and conference papers, 1 edited book and 9 book chapters in the relevant areas.

AUTHOR INDEX