FAULT DETECTION IN MOBILE COMMUNICATION NETWORKS USING DATA MINING TECHNIQUES WITH BIG DATA ANALYTICS

Prasanthi Gottumukkala[1] and G.Srinivasa Rao[2]

[1]Department of Information Technology, JNTUK,UCEV, Vizianagaram
[2]GIT ,GITAM University, Visakhapatnam.

ABSTRACT

A collection of datasets is Big data so that it to be To process huge and complex datasets becomes difficult. so that using big data analytics the process of applying huge amount of datasets consists of many data types is the big data on-hand theoretical models and technique tools. The technology of mobile communication introduced low power ,low price and multi functional devices. A ground for data mining research is analysis of data pertaining to mobile communication is used. theses mining frequent patterns and clusters on data streams collaborative filtering and analysis of social network. The data analysis of mobile communication has been often used as a background application to motivate many technical problem in data mining research. This paper refers in mobile communication networking to find the fault nodes between source to destination transmission using data mining techniques and detect the faults using outliers. outlier detection can be used to find outliers in multivariate data in a simple ensemble way. Network analysis with R to build a network.

KEYWORDS

Mobile communication, Data mining, Big Data, R Language , fault detection & outlier

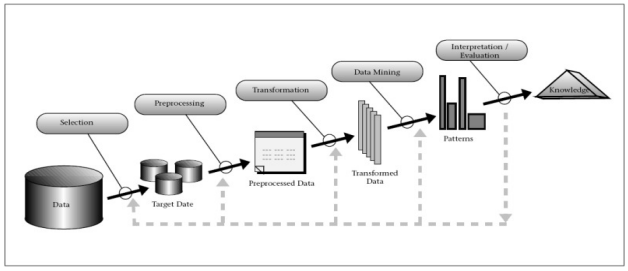
1. INTRODUCTION

1.1 Data Mining

Data mining can be viewed as a result of the natural evaluation of information technology. Data mining also named as knowledge mining from data or knowledge mining, because to extract knowledge information from huge amount of data. To generate large databases and extract huge data in various areas is the information technology development. The approach of decision making on database research and recent information technology is to store and manipulate huge data . Data mining techniques are used to sour huge databases in order to find novel and useful patterns. The process of knowledge discovery from data consists of data cleaning, integration, selection, transformation, data mining, pattern evaluation and knowledge presentation. The logical process of data mining used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.

DOI: 10.5121/ijci.2016.5410

Figure 1 : Knowledge Discovery from data



Three steps involved are Exploration, Pattern identification, Deployment, Exploration: In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined. Pattern Identification: Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction. Deployment: Patterns are deployed for desired outcome. The data mining techniques Association ,classification, clustering, prediction and anomaly detection (anomalies or outliers) are used in data mining research projects. In association, a pattern is to find the items relationship between the same transaction. In Classification on mathematical techniques are used, that is decision tree induction, such as model overfitting and evaluation of classifier. To build classification models from simple techniques such as rule based and nearest-neighbor classifiers and more other advanced techniques such as support vector machines and ensemble methods.

2. MOBILE COMPUTING

The mobile computation process is mobile computing. A technology that allows transmission of data, via a computer, without having to be connected to a fixed physical link. Over the last few years various cellular networks on number of subscribers very rapidly increase the mobile communication. Cellular networks on small size portable computers are used to communicate or send and receive data easy and accurately. A rapid technology involves is the users transmit and receive data from remote area. In this article we give an overview of existing cellular networks and the CDPD Cellular digital packet data technology which allows data communications across these networks. Finally, we look at the applications of Mobile Computing in the real world. Group of distributed computing systems service providing servers participate connect and synchronize through mobile communication.

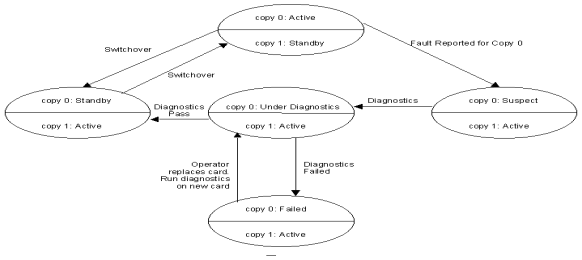
3.FAULT HANDLING TECHNIQUES

This article describes some of the techniques that are used in fault handling software design. A typical fault handling state transition diagram is described in detail. The article also covers several fault detection and isolation techniques.

3.1 Fault Handling Lifecycle

The following figure describes the fault handling lifecycle of an active unit in a redundancy pair.

Figure 2: fault handling life cycle



Assume that the system is running with copy-0 as active unit and copy-1 as standby.

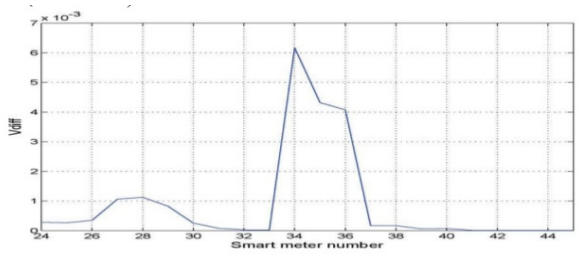
When the copy-0 fails, copy-1 will detect the fault by any of the fault detection mechanisms. At this point, copy-1 takes over from copy-0 and becomes active. The state of copy-0 is marked suspect, pending diagnostics. The system raises an alarm, notifying the operator that the system is working in a non-redundant configuration. Diagnostics are scheduled on copy-0. This includes power-on diagnostics and hardware interface diagnostics. If the diagnostics on copy-0 pass, copy 0 is brought in-service as standby unit. If the diagnostics fail, copy-0 is marked failed and the operator is notified about the failed card. The operator replaces the failed card and commands the system to bring the card in-service. The system schedules diagnostics on the new card to ascertain that the card is healthy. Once the diagnostics pass, copy-0 is marked standby. The copy-0 now starts monitoring the health of copy-1 which is currently the active copy. The system clears the non-redundant configuration alarm as redundancy has been restored.

The operator can restore the original configuration by switching over the two copies. protocol fault is the only fault reported, all the units in the path from source to estimation are probed for health.

3.2.Fault Detection

If the error occurred in the process is fault. Fault detection is indicating if there is a fault. Below graph shows the node fault.

Figure 3 : Fault Detection



Graph showing fault near to node 34

Important role of fault handling is eliminate fault immediately and try to process the fault isolation immediately or as soon as possible. Here are some of the commonly used fault detection mechanisms.

- **Sanity Monitoring:** A unit monitors the health of another unit by expecting periodic health messages. The unit that is being monitored should check its sanity and send the periodic health update to the monitoring unit. The monitoring unit will report faults if more than a specified number of successive health messages are lost.
- **Watchdog Monitoring:** This is the hardware based monitoring technique to detect hanging hardware or software modules.
- **Protocol Faults:** If a unit fails, all the units that are in communication with this unit will encounter protocol faults. The protocol faults are inherently fuzzy in nature as they may be due to a failure of any unit from the source to destination path. Thus further isolation is required to identify the faulty unit.
- **In-service Diagnostics:** Sometimes the hardware modules are so designed that they allow simple diagnostic checks even in the in-service state.
- **Transient Leaky Bucket Counters:** When the hardware is in operation, many transient faults may be detected by the system. Transient faults are typically handled by incrementing a leaky bucket counter. If the leaky bucket counter overflows, a fault trigger is raised.

3.2.1. Fault Table

Generally fault table represented as a matrix contains rows and columns , Let faults C_j represented as colums, test patterns R_i represented as rows, and $P_{ij} = 1$ if the test pattern R_i detects the fault C_j, otherwise if the test pattern R_i does not detect the fault C_j, $P_{ij} = 0$. Denote the actual result of a given test pattern by 1 if it differs from the precomputed expected one, otherwise denote it by 0. The result of a test experiment is represented by a vector where $s_i = 1$ if the actual result of the test patterns does not match with the expected result, otherwise $s_i = 0$. c_j of each column vector equivalent to a fault C_j correspond to a possible result at fault C_j case on test experiment.test experiments on the test patterns quality is depending upon three cases are given below.

a.The test result V matches with a single column vector c_j in FT. This result corresponds to the case where a single fault C_j has been located. In other words, the maximum diagnostic resolution has been obtained.

b.The test result V matches with a subset of column vectors $\{c_i, c_j \ldots c_k\}$ in fault table. This result corresponds to the case where a subset of indistinguishable faults $\{C_i, C_j \ldots C_k\}$ has been located.

c.No match for V with column vectors in fault table is obtained. This result corresponds to the case where the given set of vectors does not allow to carry out fault diagnosis. The set of faults described in the fault table must be incomplete (in other words, the real existing fault is missing in the fault list considered in FT).

Below given example on three test experiments results are V_1, V_2, V_3 explained. V_1 is first case located the single fault, V_2 is second case located the subset of two impossible to differentiate faults, and V_3 is third case located the no fault since the mismatch of V_3 with the fault table on column vectors.

Table 1: example for fault table

	C1	C2	C3	C4	C5	C6	C7
R1	0	1	1	0	0	0	0
R2	1	0	0	1	0	0	0
R3	1	1	0	1	0	1	0
R4	0	1	0	0	1	0	0
R5	0	0	1	0	1	1	0
R6	0	0	1	0	0	1	1

V1	V2	V3
0	0	1
0	1	0
0	1	0
1	0	1
1	0	1
0	0	0

fault C5 located

No match

faults C1 and C4 are not distinguishable

diagnosis not possible

3.2.2. Fault Dictionary

Fault dictionaries (FD) contain fault tables on same data. But the difference is it contains efficient/modernized data. The potential results of test experiments and the faults is mapped. That mapped represented in ordered form and more compressed is fault dictionaries. The given example table shows, the bit vectors columns represent the structured decimal codes or various type of compressed signature.

Table 2:Fault dictionary

No	Bit Vector	Decimal Number	Faults	Test results
1	000001	01	C7	-
2	000110	06	C5	V1=06
3	001011	11	C6	-
4	011000	24	C1,C4	V1=24
5	100011	35	C3	V1=38
6	101100	44	C2	No match
7	110011	32	C8	

3.2.3 Fault Detection Isolation FDI

Fault isolation is determining where the faulty occurred. If the unit or the part of data is faulty then several fault triggers can be generated for that fault unit. The major purpose of fault isolation to correlate the fault triggers and identify the fault in the data. If fault triggers are fuzzy in nature, the isolation procedure involves interrogating the health of several units. For example, if protocol fault is the only fault reported, all the units of the pathway through source towards destination are survey for strength.

Figure : 4 fault isolation

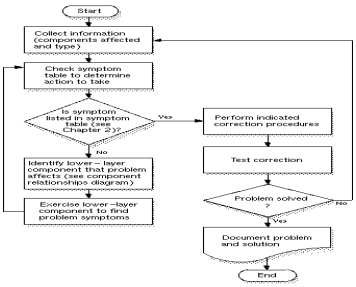
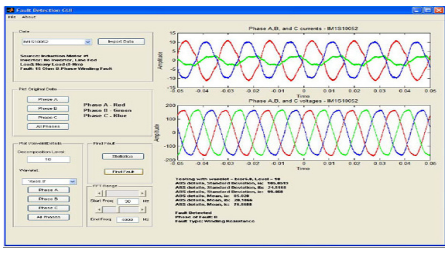


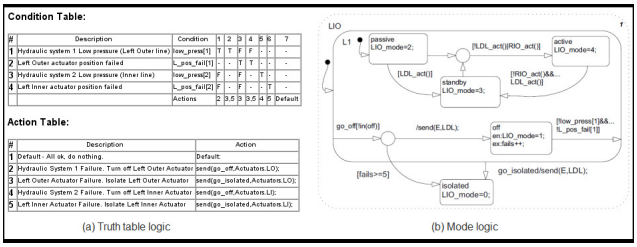
Figure5: fault detection



Fault identification is determine the size of the fault and time of the arrival of fault. Fault detection isolation on model based FDI techniques are used to decide the incident of the fault. The mathematical or knowledge based is the system model. Some of the model-based FDI techniques contain parity-space approach, observer-based approach and parameter based identification methods. There is another trend of model-based FDI schemes, which is called set-membership methods. These methods guarantee the detection of fault under certain conditions. The main difference is that instead of finding the most likely model, these techniques omit the models, which are not compatible with data.The example shown in the figure on the right illustrates a model-based FDI technique for an aircraft elevator reactive controller through the use of a truth table and a state chart. How the controller react to detect faults defines the truth table, and how the controller switches between the different modes of operation (passive, active, standby, off, and isolated) of each actuator defines the state chart.

For example, if in a hydraulic system 1 on fault is detected, then truth table send an incident to the state chart that the left inner actuator should be turned off. The model-based FDI technique most important benefit is reactive controller also connected to a continuous-time model of the actuator hydraulics and it allow the learning of switching transients

Figure 6: model based FDI for Aircraft example



3.3. Fault Diagnosis

Fault detection and fault isolation is the fault diagnosis. To trim down huge computational effort concerned in construct a fault dictionary, the detected faults are *dropped* from the set of simulated faults in fault simulation. Hence, all the faults detected for the first time by the same vector will produce the same column vector (signature) in the fault table, and will be included in the same equivalence class of faults. In this case the testing experiment can stop after the first failing test, because the information provided by the following tests is not used. Such a testing experiment achieves a lower diagnostic resolution. A tradeoff between computing time and diagnostic resolution can be achieved by dropping faults after k>1 detections. Example: In the fault table produced by fault simulation with fault dropping, only 19 faults need to be simulated compared to the case of 42 faults when simulation without fault dropping is passed out (the simulated faults in the fault table are shown in shadowed boxes). As the result of the fault dropping, however, the following faults remain not noticeable: $\{C_2, C_3\},\{C_1, C_4\},\{C_2, C_6\}$.

Table 3: fault diagnosis

	C1	C2	C3	C4	C5	C6	C7
R1	0	1	1	0	0	0	0
R2	1	0	0	1	0	0	0
R3	0	0	0	0	0	1	0
R4	0	0	0	0	1	0	0
R5	0	0	0	0	0	0	0
R6	0	0	0	0	0	0	1

4. DATA MINING FOR FAULT DETECTION

Data mining is an expanding area of research in artificial intelligence and information management whose objective is to extract relevant information from large databases .Data mining and analysis tasks include classification, regression, and clustering of data, aiming at determining parameter or data dependencies and finding various anomalies detection from the data.

4.1 Grid Computing: Grid computing has been proposed as a novel computational model, distinguished from conventional distributed computing by its focus on large-scale resource sharing, innovative applications, and, in few cases, high-performance orientation. Nowadays grids can be used as effective infrastructures for distributed high performance computing and data processing. A grid is a geographically distributed computation infrastructure composed of a set of heterogeneous machines that users can access via a single interface. Grids therefore, provide common resource-access technology and operational services across widely distributed virtual organizations composed of institutions or individuals that share resources.

4.2 Self-Organizing Map: SOM is an important unsupervised competitive learning algorithm, being able to extract statistical regularities from the input data vectors and encode them in the weights without supervision (Feher, K., 1995). Such a learning machine will then be used to build a compact internal representation of the mobile network, in the sense that the data vectors representing its behavior are projected onto a reduced number of prototype vectors (each representing a given cluster of data), which can be further analyzed in search of hidden data structures. The main advantages of their solution are the limited storage and computing costs. However, SOM requires processing time which increases with the size of input data.

4.3 Discrete Wavelet Transform: Discrete Wavelet Transform (DWT) is used to reduce the input data size, features of the data can be extracted without losing the significant data can be used for anomaly detection. Wavelets have been extensively employed for anomaly and fault detection DWT has also been integrated with SOM to detect system faults .

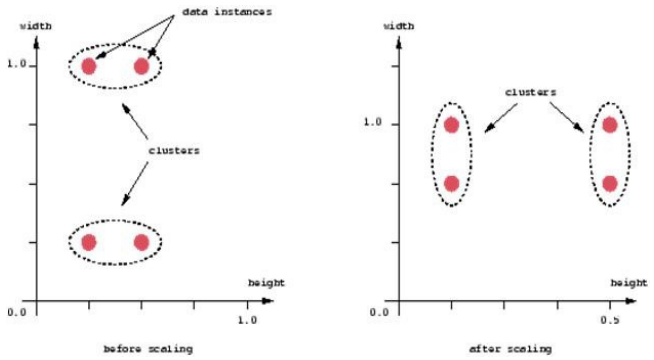
In particular, feature vectors of the faults have been constructed using DWT, sliding windows and a statistical analysis. DWT is a mathematical transform that separates the data signal into fine-scale information known as detail coefficients, and rough-scale information known as approximate coefficients.

Its major advantage is the multi-resolution representation and time-frequency localization property for signals. Usually, the sketch of the original time series can be recovered using only the low-pass-cut off decomposition coefficients; the details can be modelled from the middle-level decomposition coefficients; the rest is usually regarded as noises or irregularities.

4.4 Cluster Analysis: Clustering is a process which partitions a given data set into homogeneous groups based on given features such that similar objects are kept in a group whereas dissimilar objects are in different groups. With the advent of many data clustering algorithms in the recent few years and its extensive use in wide variety of applications, including image processing, computational biology, mobile communication, medicine and economics, has lead to the popularity of this algorithms. Main problem with the data clustering algorithms is that it cannot be standardized. Algorithm developed may give best result with one type of data set but may fail or give poor result with data set of other types. Although there has been many attempts for standardizing the algorithms which can perform well in all case of scenarios but till now no major accomplishment has been achieved. Many clustering algorithms have been proposed so far. However, each algorithm has

its own merits and demerits and cannot work for all real situations. Before exploring various clustering algorithms in detail let's have a brief overview about what is clustering.

Figure 9 : Clustering scaling

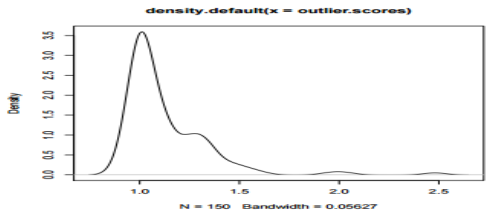


4.5 Outlier Detection by Clustering

The way to detect outliers is clustering. By grouping data into clusters, those data not assigned to any clusters are taken as outliers. For example, with density-based clustering such as objects are grouped into one cluster if they are connected to one another by densely populated area. Therefore, objects not assigned to any clusters are isolated from other objects and are taken as outliers. We can also detect outliers with the k-means algorithm. With k-means, the data are partitioned into k groups by assigning them to the closest cluster centers. After that, we can calculate the distance (or dissimilarity) between each object/nodes and its cluster center, and pick those with largest distances as outliers.

4.6 Outlier Detection with LOF LOF (Local Outlier Factor) is an algorithm for identifying density-based local outliers. With LOF, the local density of a point is compared with that of its neighbours. If the former is significantly lower than the latter (with an LOF value greater than one), the point is in a sparser region than its neighbours, which suggests it be an outlier. A shortcoming of LOF is that it works on numeric data only. Function lofactor() calculates local outlier factors using the LOF algorithm, and it is available in packages DMwR and dprep. An example of outlier detection with LOF is given below, where k is the number of neighbours used for calculating local outlier factors. Figure 10 shows a density plot of outlier scores.

Figure 10 : Density of outlier factors



5. BIG DATA

The most recent trend in the IT world and business right now is Big Data. The term that refers to combinations of data sets whose size, variability, and velocity make them difficult to be captured, managed, processed or analyzed by standard technologies and tools, these relational databases and desktop statistics, within the time necessary to make them useful. To analyse the datasets using R language. Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The upcoming new technologies Big Data ,if the failure occurred it should be within acceptable threshold. Thus the major task is to limit the probability of failure to an "acceptable" level. But it is very expensive to reduce the probability of failure.

6. CONCLUSION

The purpose of this paper is to use data mining tools for identifying defective parts in data communication. First find faults points in transmission nodes and then using data mining techniques detect the faults. Fault detection, isolation, recovery is a subfield of control engineering which concerns itself with monitoring a system, identifying when a fault has occurred, and pinpointing the type of fault and its location. To analysis of datasets use big data tools example R language. R is a programming language and software environment for statistical analysis, graphics representation and reporting. Very fast growing industry is mobile computing. Very limited patterns could be found from real data by human analysts thereby paving way for avenues of data mining research for pattern hunting in mobile communication data sets. Various data mining techniques are discussed for fault detection in mobile communication and further new technique will be introduced for fault detection. The paper also focuses on technical challenges with Big Data processing. using big data analytics faults also reduced.

REFERENCES

[1] Wireless Digital Communications: Modulation and Spread Spectrum Applications‖.Upper Saddle River, NJ: Prentice Hall.
[2] Data mining and ware housing tan han
[3] Introduction to Clustering Techniques by Leo Wanner
[4] Data Clustering: A Review by A.K. Jain, M.N. Murty and P.J. Flynn.
[5] Albert Bifet "Mining Big Data In Real Time" Informatica 37 (2013).
[6] Introduction to R for Data mining 2012 spring webinar series,Joseph B.Rickert Revolution analytics june 5,2012
[7] R and Data Mining : Examples & case studies Yanchang Zhao, http://www.RDataMining.com ,April 26 2013