Int. J. Advance Soft Compu. Appl, Vol. 8, No. 3, December 2016
ISSN 2074-8523

Network Equipment Failure Prediction with Big Data Analytics

Lam Hai Shuan[1], Tan Yi Fei[1], Soo Wooi King[2], Guo Xiaoning[1], Lee Zhe Mein[1]

[1]Engineering Big Data Lab,
Faculty of Engineering Multimedia University, Malaysia
[2]Faculty of Computing and InformaticsMultimedia University, Malaysia
e-mail: hslam@mmu.edu.my, yftan@mmu.edu.my,
guo.xiaoning@mmu.edu.my, zhemein@gmail.com,
wksoo@mmu.edu.my

Abstract

Minimising network downtime has been a challenge to all telecommunication service providers. One of the major causes for such downtime is equipment failure at various locations and rectification works are required on ad-hoc basis. Therefore, if these failures can be predicted and rectified, downtime can be reduced. The system activities and operation parameters of these equipment are reported over the network and logged at a monitoring station. By studying these data from the equipment, many of the equipment related failures can be predicted to ensure minimal downtime and increase customer satisfaction. However, these data are massive and generated at very high velocity. A dynamic and adaptive algorithm is needed to process the huge amount of data and generate predictions based on trends and patterns. This paper presents a rule based analysis with regression technique and best-fit line methods to predict the equipment failure. The warning occurrence pattern is studied on daily basis and a threshold for alarm signal triggering can be set. The output of this work suggests that the symptom of a failure started as early as 9 days before the failure while for prediction within 4 days before the failure has an accuracy of up to 99.9%.

Keywords: *failure prediction, big data analytics, Hadoop, regression technique, rule based analysis.*

1 Introduction

The prediction of equipment failure has been a critical issue faced by the telecommunication companies because the equipment failure affects the outlook of a company in terms of customer service and reputation [1]. The improvements on customer service and equipment systems are highly desired within the telecommunications industry due to the increasing number of users experiencing Internet service outage in Malaysia [2]. The root cause of Internet service outage is the inability to predict failures in the system due to the massive amount of fault data received by operation staffs and it takes a long duration to process the data to identify the real causes. In this case, there are two key challenges addressed which are the manipulation of large volume of data and handling of high data velocity which can be solved by implementing big data platform, Hadoop [3]. The problem of storing voluminous amount of data can be resolved using Hadoop Distributed File System (HDFS) which enables the data to be stored and analyzed across the distributed databases. With the data rapidly changing, Hadoop's fair and capacity scheduler is assigned to solve the velocity problem in Big Data ([4,5]). Hence, with the help of Hadoop platform handling the incoming flow of data, a series of methods are proposed to analyze the cause of equipment failure and predict any upcoming failure that follows the same trend. With these methods, equipment failures can be predicted and rectification work can be carried out before the outage happens.

The method proposed in this paper includes a rule based analysis to predict the equipment failure and the analysis is conducted on Hadoop platform. In this paper, the warning occurrence pattern leading to the equipment failure is studied in detail and the framework is built and tested using three months of historical data. These historical data include customer reported trouble ticket (CTT), network reported trouble ticket (NTT) and Syslogs of the equipment. Due to confidentiality, the data presented in this paper are masked. The credibility of the model is supported by comprehensive tests. The remaining of the paper are organised in the following sequence: Section 2 describes the related works, Section 3 explains the methodology, Section 4 presents the results and discussions, and Section 5 concludes the overall findings.

2 Related Works

Big data analytics has been a popular field of studies in the recent years due to the value it can generate from the data. With the intensive growth of data sizes and velocity, a big data platform is required to store and manipulate the data. The Apache Software Foundation has developed Hadoop as an open source cloud computing platform that consists of Map Reduce as the software programming framework and HDFS as the distributed file system. With Hadoop, the three key challenges addressed by Big Data such as volume, velocity and variety highlighted by Kamalpreet Singh et al. in a research paper [3] can be solved. A research done by Velmurugan et al. in [6] highlights the importance of maintaining a desired quality of service in communication is by identifying the movement pattern of users based on one month Syslog data of Darmouth College. In this research, a hidden Genetic Algorithm layer-GA-SOFM Neural Network is proposed to predict the movement of users at various locations which can be applied on network prediction in the future. The user's movement pattern is observed by identifying the frequently used path which is known as User Mobility Pattern (UAP). Besides, Bayes Modeling method is also useful in predictive failure analysis. This method is proposed in [7] by C. Carlsson et al. using the possibilistic Bayes models. With the possibilistic Bayes models, a system is developed to assist the monitoring and control personnel to detect possible failures and optimal programs for predictive maintenance to be planned. Meanwhile, Logistic Regression method is highlighted in ([8,9,10]) for the failure prediction. This method is performed to analyze the factors contribute to communication failure and predict failures in the grid metering automation system. In this research, a model based on the logistic regression algorithm is proposed to predict upcoming failure which minimizes the electric power consumption in the metering automation system. The operator is able to observe the data pattern and solve the failure in time. This research concludes that logistic algorithm modeling is a credible model and can be applied directly to predict failure.

3 Methodology

Three months of historical data were sampled and first analysed to identify the relationship between the columns in the tables. The CTT contains details of customer complaints about the service interruptions, the NTT consists of technical information about the equipment breakdown and the Syslogs contain event logs, warnings, and alarms generated by the equipment. All records are matched according to Network ID of the equipment which serves as the primary key for all the data. However, the historical data do contain data of equipment failure due to external factors such as weather problems, power failure, theft and vandalism. These information are filtered as it will result in prediction inaccuracy. By matching CTT and NTT data, the equipment failure cases can be identified. The fault of equipment recorded in CTT data is cross checked with the causes of faults in the NTT data which has the detail description on the equipment failure. Then, the warnings generated by the equipment are analyzed. Hence, important features that are related to communication failure due to equipment failure are extracted out from the historical data using Sequential Query Language (SQL).

The framework of the rule based analysis model is illustrated in Fig. 1.

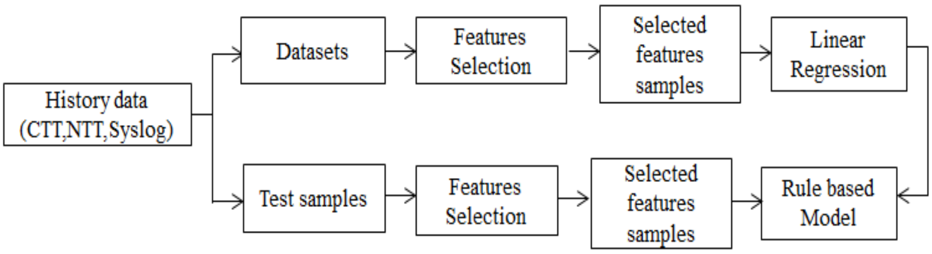

Fig. 1: Framework of rule based analysis model

Datasets from the first and second months are used to construct a rule based analysis model with a linear regression algorithm. Similar features are extracted as the test samples to map against the data from the third month to validate the model's accuracy. The selected features listed in Table 1 are prepared using SQL language on Hadoop by tabulating the cumulative total number of each warning from Syslogs for logged data on particular days based on CTT occurrences.

Table 1: Features selected from historical data

CTT	**NTT**	**Syslog**
Date	Date	Date
Fault of equipment	Causes of faults	Types of warnings
Types of equipment	Types of equipment	Types of equipment
Network ID	Network ID	Network ID

The types of warnings are represented as Warning L and Warning R. Then, based on the observations of the warning occurrences, two hypotheses were proposed for testing such as:

- The pattern obtained from the cumulative increment of total for each warning several days prior to equipment failure shall be similar for every equipment with the same model and with the same failure
- When certain range of cumulative total number of each warning is observed, equipment failure will happen on the next day

The proposed hypotheses are tested and discussed in the results section.

3.1 Linear Regression Model

Due to the large scale of data logged by the equipment, the time needed to predict the failure is lengthy. Considering the fact that the data logs occur in real-time and continuous, the data needs to be processed with the simplest and effective method. Therefore, linear regression model is suitable for predicting an outcome in this project compared to Bayes and Logistic Regressions methods in [7,8,9]. In linear regression theory, the relationship between independent and dependent variables have to be determined first before constructing the model. The variables which are not related to each other will not provide a functional model. The variable that is used to base the prediction on is called the predictor variable and is referred to as X. As for the variable that is used to predict is called the criterion variable and is referred to as Y. A straight line called as regression line is formed when predictions of Y is plotted as a function of X. In general, the formula for a regression line is shown as follows:

$$Y = Ax + C \qquad (1)$$

where Y is the predicted outcome or value, A is the gradient, x is the predictor variable and c is the intercept. In this project, simple linear regression is performed to predict the cumulative total number of each warning the next day. The variable Y is the cumulative total number of each warning per day and X is the day cumulative total number of each warning recorded. The regression line will be computed and the square of correlation coefficient denoted as R^2 is determined. The square of correlation coefficient indicates how well the regression line represents the data. It gives the variance of one variable that is predictable from the other variable and can be used as a measure to determine the certainty in making the predictions from a certain model or graph. The range for R^2 is 0 to 1. As R^2 approaches 1, the regression line fits better on the data and has stronger linear relationship.

3.2 Implementation on Hadoop

The predictive analysis for these works are implemented on Hadoop platform. The system consists of 1 primary Name Node and 4 Data Nodes. The data hosting is with Hive which serves as the data warehouse infrastructure on HDFS and the queries were done with HiveQL. Zookeeper was used as the main coordinator to manage the cluster services. The data collected is first stored using HDFS and then distributed across the cluster. All the nodes will process the data using SQL codes in Hive. Grouping or reduction of data is unnecessary as SQL has a native function to classify and process the data in all-in-one process to determine the total number of occurrences for each data. The results are collected and graphs are generated when all the values of occurrences are accumulated.

The raw data for the equipment Syslog consists of more than 4 million records per day and 2 months of data were used for developing the prediction model. The mappers key/value pairs were created with Warning type and the counting values. The reducer were set to aggregate the counting values giving the daily accumulative total for each Warning type and for each equipment.

4 Results and Discussions

In this section, results using the methods above are presented and discussed. In addition, the validity of the hypotheses is tested as well.

4.1 Hypotheses Testing

The dataset consists of approximately 480 million records for 3 months of logged data containing equipment event records. The hypotheses are tested on Equipment S because this equipment has the highest number of customer complaints in the first month and equipment failure is recorded in the second month and third month as well. This provides a wide range of information that can be analysed. To ensure the consistency of equipment failure trend, this method is validated on another equipment with same model. The first hypothesis was tested and the results are illustrated in Fig. 2 and Fig. 3.

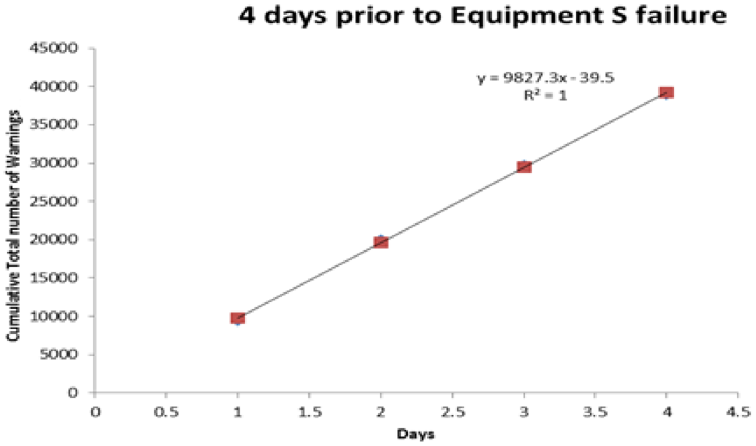

Fig. 2: Regression line computed for Warning L from 4 days prior to Equipment S failure

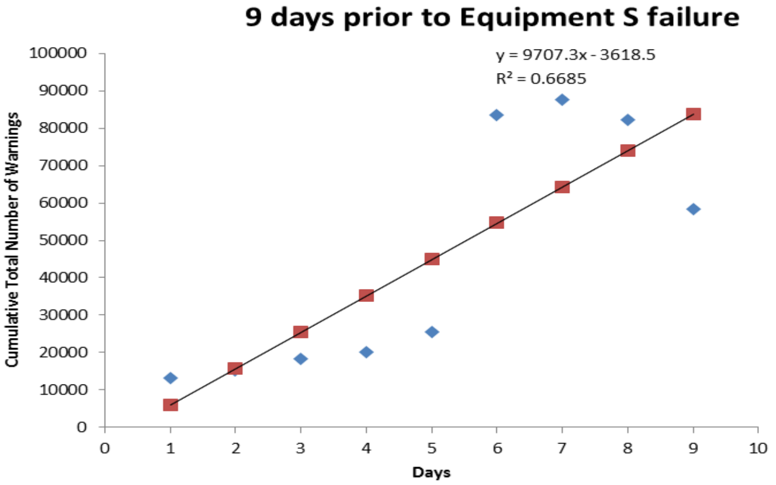

Fig. 3: Regression line computed for Warning L from 9 days prior to Equipment S failure

Fig. 2 and Fig. 3 present the pattern for cumulative total number of Warning L prior to 4 and 9 days respectively before equipment S failure. The same method was tested on Warning R of Equipment S and the results for both are compared in Table 2.

Table 2: Comparison of number of days to predict equipment S failure

Equipment	R^2 values			
	4 days prior to failure		9 days prior to failure	
	Warning L	Warning R	Warning L	Warning R
S	1	0.995	0.6685	0.6489

The results show that Warning L is better in terms of predicting the equipment failure compare to Warning R and the warning occurrence pattern will converge to a particular pattern as the equipment failure is about to occur. In Figure 2, R^2 =1 for 4 days prior to equipment S failure whereas R^2 = 0.6685 for 9 days prior to Equipment S failure. In this case, a smaller number of days is taken into consideration to predict equipment S failure. This has been tested on different equipment with different number of days and the results are listed in Table 3.

Table 3: Comparison of number of days to predict different equipment failure

Equipment	R^2 values					
	4 days prior to failure		6 days prior to failure		9 days prior to failure	
	L	R	L	R	L	R
A	0.999	0.994	0.812	0.908	0.773	0.671
B	0.998	0.996	0.825	0.836	0.776	0.645
C	0.992	0.998	0.838	0.855	0.779	0.669
D	0.997	0.997	0.810	0.849	0.756	0.678
E	0.995	0.999	0.815	0.807	0.794	0.684
F	0.999	0.998	0.821	0.911	0.702	0.665
G	0.996	0.999	0.830	0.822	0.707	0.687

Table 3 shows the values of R^2 for cumulative warnings occur for different number of days prior to equipment failure. It shows the consistency of smaller number of days fitted better by the regression line. Next, for the cumulative total number of each warning occurs, the pattern of regression line in Month 1 and Month 2 (datasets) is compared with the data from Month 3 (test samples). Since the regression lines are linear, gradient is enough to illustrate the pattern of the line. Thus, the gradients for those regression lines from the datasets are calculated based on the data of Month 1 and Month 2 as listed in Table 4.

Table 4: Gradients obtained for Equipment S

Datasets	Gradient (M_n)	
	Warning L	Warning R
1	9814	12532
2	34997	38770
3	13287	7947
4	11335	8444
5	9906	5155

The gradients obtained in Table 4 are based on the date of equipment S failed as recorded in NTT. In Month 1, equipment S encountered failure once and in the second month, equipment S encountered failure on 4 different days. Therefore, 5 sets of data were collected for analysis.

The Month 3 data are used for testing and validating the hypothesis. It is identified that equipment S encountered failure on 3 different days in Month 3. Hence, the gradients obtained in Table 4 are further used for predicting the failure on Month 3 to justify the proposed hypothesis. The results are listed in Table 5.

Table 5: Validate the gradients with test samples for Equipment S

Test Samples	Gradient (M_n)		Similarity	Gradient Difference (%)
	Warning L	Warning R		
1	14454	7131	Dataset 3	9.00
2	13267	8863	Dataset 3	5.22
3	10799	5929	Dataset 5	9.03

The gradients obtained in Month 3 are cross validated with the gradients obtained in Month 1 and Month 2. The gradient difference is about ±10%. A range of cumulative total number of each warning is then specified based on the gradient obtained in Table 4. The specified range of each warning is incorporated with the gradient obtained to formulate a rule. The formation of the model based on these parameters is listed in Table 6.

Table 6: Rule based Analysis Model

Rule	Warning L	Warning R
1	M_L=9814: 22000<R_c< 26000	M_R=12532: 28000< R_c <32000
2	M_L =34997: 83000< R_c < 87000	M_R =38770: 91000< R_c <95000
3	M_L =13287: 31000< R_c <35000	M_R =7947: 17000< R_c <21000
4	M_L =11335: 26000< R_c <30000	M_R =8444: 20000< R_c <24000
5	M_L =9906: 22000< R_c <26000	M_R =5155: 11000< R_c <15000

Table 6 presents the rules formulated in the model to predict the upcoming failure where M_L and M_R are referred to the gradient of regression line based on Warning L and Warning R respectively. The range of cumulative number of warnings is denoted as R_c. If the equipment cumulative total numbers of Warning L and Warning R and gradients fall within the range of one of the rules in Table 6, the possibility of the equipment to fail the next day is high. These rules are validated with the test samples of Equipment S listed in Table 7.

Table 7: Validation of Model with test samples of Equipment S

Test	Warning L	Warning R	Similar to Rule
1	M_L=14454:36000< R_c <40000	M_R=7131:17000< R_c <21000	3
2	M_L =13267:32000< R_c <36000	M_R =8863:20000< R_c <24000	3
3	M_L =10799:25000< R_c <29000	M_R =5929:12000< R_c <16000	5

The test samples of equipment S produced gradients and range of each warning exhibit high similarity as the parameters in the rule based analysis model. In fact, equipment S encountered failure the next day as recorded in CTT data. The model is further validated with different equipment listed in Table 8.

Table 8: Validation of Model on different equipment

Equipment	Warning L	Warning R	Similarity
A	M_L =10083 23000< R_c < 27000	M_R =6989 13000< R_c <17000	Rule 5
B	M_L =8192 17000< R_c <21000	M_R =5301 11000< R_c <15000	Rule 5
C	M_L =9234 21000< R_c <25000	M_R =10981 26000< R_c <30000	Rule 1
D	M_L =14852 32000< R_c <36000	M_R =8513 17000< R_c <21000	Rule 3

In Table 8, the gradients and range of each warning are identified for different equipment 4 days prior to failure. It was found that the equipment listed in Table 8 encountered failure the next day. However, to further validate the model, the model is tested on different equipment that did not experience any breakdown even though there are warnings recorded in Syslog. The results are listed in Table 9.

Table 9: Validation of Model on different equipment

Equipment	Warning L	Warning R	Similarity
T	M_L =542 1100< R_c <1500	M_R =2516 5000< R_c <9000	None
U	M_L =311 100< R_c <500	M_R =61 90< R_c <400	None
V	M_L =476 800< R_c <1200	M_R =83 800< R_c <1200	None

Table 9 shows that none of the equipment without failure tested similar to any of the rules. Therefore, it shows the rules establish did not generate any false alarm. It is proven that although there are warnings recorded but the gradients and range of each warning did not fall in the range listed in the model. Based on the results presented in this paper, the rule based analysis model is regarded as a reliable model considering the fact that it can be used as a reference directly to predict equipment failure.

5 Conclusion

In conclusion, the proposed hypotheses are valid and the rule based analysis demonstrated consistency on predicting equipment failure. This is proven by observing the pattern obtained from the cumulative total of each warning for 4 days prior to equipment S failure which is similar with other equipment using simple linear regression. The R^2 values obtained are closer to 1 for 4 days prior to failure and shows the strongest linear relationship compared to 6 days and 9 days prior to failure. This means that the regression line fits closely to the actual data. With the gradients obtained from the regression lines 4 days prior to failure, certain range of cumulative total number of each warning is specified to predict the failure in the next day. The range of cumulative total of each warning is incorporated with the gradients obtained to construct a rule based analysis model which can effectively predict the equipment the next day. The future work can be summarized as followed: more rules can be developed to cater for model types of equipment. Meanwhile, the rule based method can be expanded to include additional warning types to identify significant relationship to various types of equipment failure.

References

[1] Chang, P.K, Chong, H.L. 2011. Customer satisfaction and loyalty on service provided by Malaysian telecommunication companies. *Electrical Engineering and Informatics (ICEEI).*

[2] Peng Ching Huai. 2009. The Role of Customer Satisfaction, Customer Value and Service Experience in Telecommunication Industry. *ISECS International*

Colloquium on Computing, Communication, Control, and Management, Volume 3, pp. 295-299.

[3] K. Singh, R. Kaur. Hadoop. 2014. Addressing Challenges of Big Data. *2014 IEEE International Advance Computing Conference (IACC)*. pp. 686-689.

[4] Thibaud Chardonnens, Philippe Cudre-Mauroux, Martin Grund, Benoit Perroud. 2013. Big Data Analytics on High Velocity Streams: A Case Study. *IEEE International Conference on Big Data*. pp. 784-787.

[5] Deepa Gupta, Sameera Siddiqui. Big Data Implementation and Visualization. *IEEE International Conference on Advances in Engineering & Technology Research (ICAETR-2014)*, August 01-02,2014, Dr.Virendra Swarup Group of Institutions, Unnao, India.

[6] Velmurugan, L. and P. Thangaraj, 2012, A Hidden Genetic Layer Based Neural Network for Mobility Prediction, *American Journal of Applied Sciences* 9 (4): 526-530

[7] Christer Carlsson, Markku Heikkilä and Jozsef Mezei, 2014 Possibilistic Bayes Modelling for Predictive Analytics, *Computational Intelligence and Informatics (CINTI), IEEE 15th International Symposium on*, pp 15-20

[8] Harrell FE. 2001. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. *Springer.*

[9] Tao LIU,Shaofeng Wang, Shaocheng WU, Jing MA, Yueming LU. 2014. Prediction of Wireless Communication Failure in Grid Metering Automation System Based on Logistic Regression Model. *China International Conference on Electricity Distribution, Shenzhen.*

[10] Tung Le, Ming Luo, Junhong Zhou, Chan H.L. 2014. Predictive maintenance decision using statistical linear regression and kernel methods. *Emerging Technology and Factory Automation (EFTA), IEEE*