International Journal of Engineering & Technology, 7 (2.14) (2018) 93-96

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Predicting Network Faults using Random Forest and C5.0

Ji Sheng Tan[1]*, Chin Kuan Ho[1], Amy Hui Lan Lim[1], Mohd Rizal bin Mohd Ramly[2]

[1]Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 631000 Cyberjaya, Selangor, Malaysia.
[2]Telekom Malaysia, Menara TM, Jalan Pantai Baharu, 50672 Kuala Lumpur, Malaysia.
**Corresponding author E-mail: jisheng@tanjisheng.com*

Abstract

The Internet is an enabling technology that assists daily and business activities. However, a network fault could prevent the user from accessing the internet thus creating trouble tickets. Ideally, accurate prediction prior to network fault allows the telco to respond before the customer raises a trouble ticket. Current research focuses on forecasting the quantity of trouble ticket using historical trouble ticket. To improve the prediction of network fault, the customer trouble ticket data is augmented to include internet usage data and signal measurement data. Random Forest (RF) and C5.0 Decision Tree algorithms are used to derive predictive models. Experiment results reveal that RF shows higher AUC score as compared to C5.0 Decision Tree. RF is able to identify the important features while C5.0 Decision Tree is able to list decision rules that describe the relation among selected features.

Keywords: *Broadband Network; C5.0 Decision Tree; Network Fault Prediction; Random Forest; Telecommunication*

1. Introduction

The Internet has become a huge part of everyday life for the current generation. Many people rely on the Internet to buy their groceries, hail a taxi or get their questions answered instantly. Many people today could not imagine life without the Internet. However, a network fault that occurs in the broadband network could prevent the user from accessing the internet.

Not all network faults cause network failure where the customer is prevented from using the service. A network fault may result in the service being delivered deviate from the customer expectation based on the package that they subscribed to. There are multiple possible causes of network faults, (i) customer on-premise equipment (CPE), (ii) core network, (iii) outside plant between CPE and core network, (iv) third party, (v) customer and others.

Three types of network faults are being considered in this research: customer on-premise equipment, core network and outside plant between CPE and core network.

Network faults that could not be resolved by the telco are not being considered in this research. Power outage and power socket malfunction are examples of third party and customer network fault respectively.

1.1. Network terminology

There are several types of fixed line broadband connection used in Malaysia, the main connection includes: Asymmetrical Digital Subscriber Line 2 (ADSL2), Symmetrical Digital Subscriber Line (SDSL), Very-high-bit-rate Digital Subscriber Line 2 (VDSL2), Fiber To The Home (FTTH). Different broadband connection technology generates different signal quality measurements. The signal quality data is collected when the Internet Service Provider (ISP) sends a ping to the customer modem. The ping will go through multiple network components before reaching the network equipment at the customer premise. Information related to the time taken to receive the ping, the signal to noise ratio, the temperature of the network equipment, and many other traceable quality measures are captured. This research focuses on the main high-speed broadband connections, FTTH and VDSL technologies. Broadband Remote Access Server (BRAS) routes traffic from the ISP to the customer. BRAS are used by the ISP to authenticate the customer for their subscription of service and control the quality of service provided to the customer. When a session is terminated, the total upload bytes, total download bytes, total duration, the cause of the session termination and the time where the session starts and ends are being captured. There are about 16 million records of user internet usage session being collected every month.

1.2. Network fault prediction

A customer who has trouble using the service will contact the customer service of the telco to log a trouble ticket. Information about the service affected and symptom that the customer notice is being identified while the customer service agent guides the user through the process of identifying the potential problem. Some minor network fault could be corrected by resetting the modem or checking the cable to make sure that they are connected. The network resolution team will be sent to perform further investigation when the network fault could not be corrected by the customer. A network fault needs to be treated quickly and properly, often within 24 hours from the creation of the trouble ticket. A telco often needs to arrange their workforce to investigate and resolve the trouble tickets created by the customers. The ability to predict network fault occurrences allows a telco to optimize workforce allocation.

There are several research conducted that focused on network fault prediction using different data and techniques. Current research work treats fault related data in two ways: (i) using the quantity of trouble tickets created and time series predictive model to predict the quantity of network fault [1,2] and (ii) using system

logs generated by specific network components to predict the likelihood of the components to be faulty [3,4]
In this paper, 7-day sliding window is used to aggregate the sessions of the Internet usage data and the network quality data. The customer trouble ticket archive is used to label the aggregated data as fault or no-fault.
Decision tree algorithms are often used to extract useful information [5], C5.0 can achieve higher accuracy with boosting [6]. RF uses multiple decision trees to achieve higher accuracy [7] and robust with respect to noise [8]. So, C5.0 and RF are applied to learn the features that contributed to the creation of trouble tickets. The features selected by the C5.0 decision tree and RF algorithm are important for the telco to take a preventive action in the future.

2. Related work

2.1. Customer trouble ticket

All The quantity of customer trouble ticket related to network fault can be represented with time series. Sampling the quantity of trouble ticket using equal interval allows the data to be presented in a discrete format. Using the trend exist in the time series, future sequence can be forecasted. There was a focus in Indonesia's telco [1] on predicting the quantity of network fault related to customer on-premise equipment in a monthly interval using Autoregressive Integrated Moving Average (ARIMA). The multivariate recurrent neural network model can also be used to forecast the network fault in both short-term (hourly and daily interval) and long-term (weekly and monthly interval) [2].
Given enough training data, the number of trouble tickets that will be created can be forecasted in advance allowing the telco to manage their workforce allocation. However, telcos are interested in identifying which user is going to experience a network fault so that they could do preventive maintenance to minimize the down time.

2.2. System logs

Besides looking at the forecasting of expected trouble tickets as a time-series problem, system logs generated by network component is another good indicator that can be used by telco to identify the network components that will soon be faulty.
Using system logs generated by Digital Subscriber Line Access Multiplexer (DSLAM), network fault can be identified by training one-class Support Vector Machine (SVM), Decision Tree and Bayesian Network [3]. System log generated by network components are often viewed as a series of high-frequency events, so the problem is approached using Hadoop clusters and simple model like linear regression4. The result using system logs shows that the number of trouble tickets related to the device analyzed can be accurately identified few days before the network component failure.
System log gives in-depth details about the network component but there are several difficulties in implementation. The difficulties of using system logs include (i) The need to acquire the data which often not available on a real-time basis. (ii) The storage required for processing all the data. (iii) Different manufacturers and different models will generate different kinds of system log. Major preprocessing is required with expert guidance in understanding the verbose logs prior to training a prediction model.

3. Proposed methodology

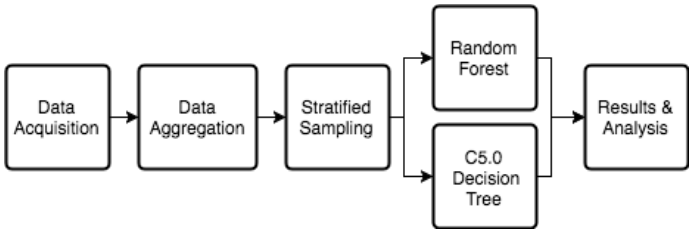

Fig. 1: Methodology

Figure 1 shows the overall methodology of the network fault prediction approach taken in the research. After acquiring historical data from the telco, the data is aggregated. Stratified sampling is applied to split the data for training and for testing. Four different sampling ratios are used, 80% for training, 70% for training, 60% for training and 50% for training. The training set is then used to train the classifiers. Two different classifiers are used, random forest and C5.0 decision tree which belongs to the decision tree family. The testing data set are used to evaluate the classifiers. The decision tree is then analysed to understand variables that significantly contribute to a network fault.

3.1. Data acquisition

Six months of historical data were being taken and analysed from July 2016 to December 2016. There are three main datasets that are being used, (i) customer trouble ticket data, (ii) customer internet usage data and (iii) signal quality data. Table 1 below shows the key features of each data.

Table 1: Key Features of Data Used

Customer Trouble Ticket	Customer Internet Usage	Signal Quality Measurement	
		VDSL	FTTH
Customer ID	Customer ID	Customer ID	Customer ID
Created Date	Upload Bytes	Signal-to-noise Ratio	ONU Power
Resolved Date	Download Bytes	Attenuation	OLT Power
Type of Network Fault	Session Duration	Maximum Speed	ONU Temperature
	Termination Causes	Achieved Speed	ONU Ranging
	Subscribed Speed		ONU Cyclic Redundancy Check

3.2. Data aggregation

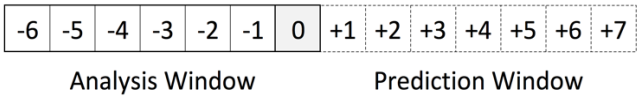

Fig. 2: 7-day sliding window method for weekly analysis

The user internet usage data can range from a second to multiple days, while the network signal quality data ping will only happen once every week, hence there is a need to aggregate the data to perform analysis and prediction.
A 7-day sliding window was chosen because most of the customer will have different internet usage on weekday and weekend. The customer will spend their time in the office or away from home during the working hours and surf the internet more during the nights and the weekends.
Figure 2 shows the 7-day sliding window that can be used in real life scenario. The data related to the customer internet usage and signal quality collected from the day of analysis and 6 days prior to the day of analysis will be aggregated. The number of internet

usage sessions terminated are grouped with the termination cause and counted. The total upload and download bytes of the sessions are converted into upload and download bytes per second before averaging with the mean operation.
The measurement of the signal quality often does not change very drastically. Hence, there is often very little ping from the telco to the customer. The data available in a sliding window is often one or two. So, the mean aggregation operator is used to aggregate the data.
The aggregated data is then labelled with the customer trouble ticket data. If there is a customer trouble ticket created within the prediction window, the next seven days, the aggregated data are labelled as fault otherwise no-fault.

3.3. Learning the prediction model

It is important to optimize the parameters of the prediction model to achieve the best detection rate of network fault. C5.0 decision tree can often be boosted to provide better prediction accuracy by optimizing the number of trials [6]. Experiments are carried out using different numbers of trials to identify the optimal value of trials where the accuracy is the highest.
There are two parameters in RF that could be optimized [8]. The number of variables in the random subset at each node (mtry) and the number of trees to be constructed in the forest (ntree). To identify the optimal value of mtry and ntree, experiments are carried out with different mtry and ntree at each iteration. Higher ntree will often lead to a better accuracy but the cost of computation will increase as the ntree increases. Hence, the optimal parameters are selected based on the ability to predict network fault and the computational overhead.

3.4 Evaluation method

The area under the receiver operating characteristic (ROC) curve, known as the AUC, is suggested against accuracy [9]. An AUC score of 1.0 represents a perfect classifier, and a score of 0.5 is a random classifier. ROC is constructed by plotting the true positive rate (TPR) and the false positive rate (FPR). Taking into account both TPR and FPR, a model that predicts everything as positive when the majority of the data are positive will result in a high accuracy but low AUC score, making AUC better metric against accuracy in this scenario. Hence, the AUC is used as the evaluation metric.

4. Result and analysis

4.1 Optimizing parameters of prediction model

In order to identify the optimal parameters to use with the RF prediction model. Experiments are carried out with different *mtry* from 1 to 10, while *ntree* are increased by 5 for every experiment from 5 to 500.

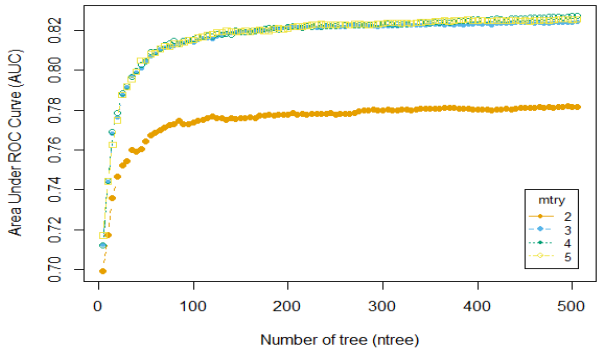

Fig. 3: AUC of RF Prediction on VDSL data using different *ntree* and *mtry*.

Figure 3 shows the AUC of the RF prediction model using VDSL data at every iteration using different values of *mtry* and *ntree*. The *mtry* of 3 and *ntree* of 500 has the highest AUC score of 0.8249. However, the AUC appears to be stable at score of 0.8193 when the number of tree is at 200. The time taken to train the model increases from 206.77 seconds at 200 *ntree* to 470.47 seconds at 500 *ntree*. Considering the time taken difference on the two different parameters, the *mtry* of 3 and *ntree* of 200 are selected as the parameters for the VDSL data.

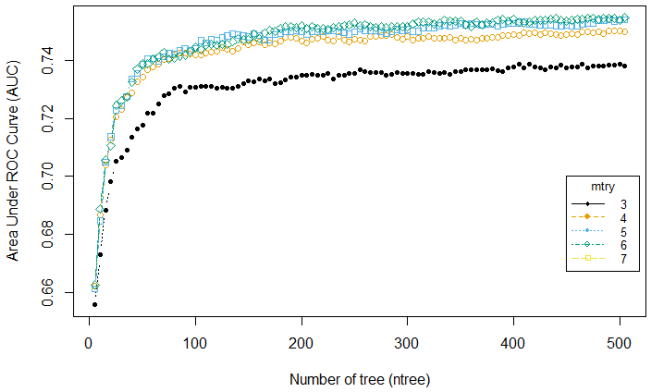

Fig. 4: AUC of RF Prediction on FTTH data using different *ntree* and *mtry*.

Figure 4 shows the AUC of the RF prediction model using FTTH data. The *mtry* of 6 and *ntree* of 500 has the highest AUC score of 0.7527. However, the AUC appears to be stable at score of 0.7484 when the number of tree is at 200. The time taken to train the model increases from 96.68 seconds at 200 *ntree* to 240.81 seconds at 500 *ntree*. The *ntree* parameter is set to 200 after considering the computational time difference, while *mtry* of 6 is chosen.

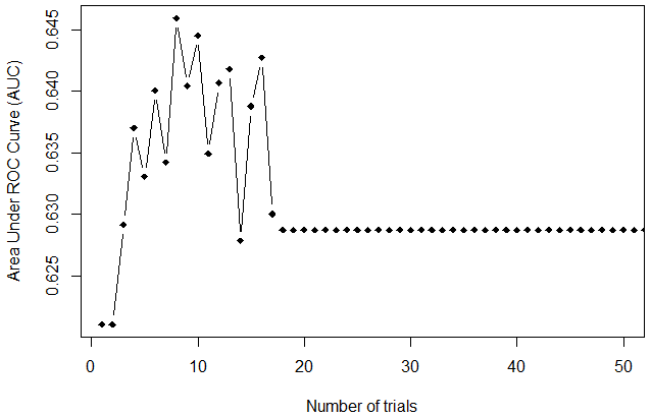

Fig. 5: AUC of C5.0 Prediction on VDSL data using different number of trials.

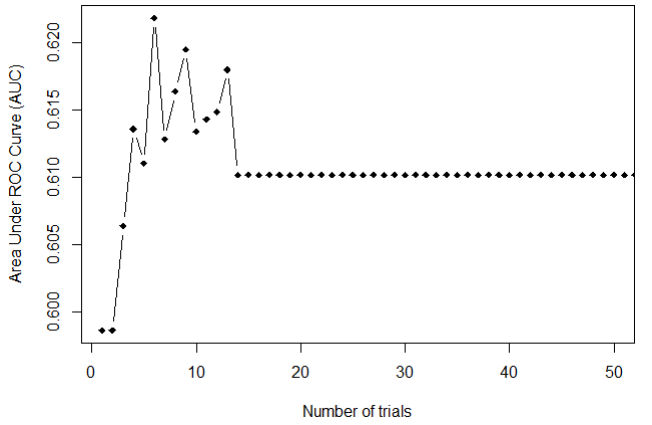

Fig. 6: AUC of C5.0 Prediction on FTTH data using different numbers of trails.

To identify the optimal number of boosting iteration for C5.0 decision tree, different numbers of trials were tested from 1 to 100. The AUC score of each iteration is recorded. Figure 5 and 6 shows

the AUC of the C5.0 prediction model using VDSL and FTTH data respectively. Since the AUC score remained constant after 18th and 14th iterations, the graph is trimmed to show only the first 50 iterations. The optimal number of trials are 8 and 6 respectively.

4.2 Classifier performance

Table 2: Prediction Results (AUC)

Train/Test Ratio	VDSL		FTTH	
	RF	C5.0	RF	C5.0
80 / 20	0.8220	0.6459	0.7500	0.6218
70 / 30	0.8052	0.6337	0.7375	0.6074
60 / 40	0.7872	0.6245	0.7221	0.6029
50 / 50	0.7642	0.6147	0.7108	0.6055

The prediction model is evaluated with the testing sets that are set aside when training the prediction model. The prediction result is presented in Table 2. The left column indicates the ratio where the training and testing are separated with stratified sampling. The AUC score of RF and C5.0 prediction model on VDSL data and FTTH data are presented. Overall, RF classification performed better compared to C5.0 in predicting the likelihood of the customer to raise a trouble ticket related to network fault in the prediction window.

4.3 Important variables

Even though the C5.0 decision tree algorithm performs poorly compared to RF in terms of AUC. C5.0 produces easily readable and understandable models in the form of trees. The RF algorithm, on the other hand, is able to identify the features that are important based on all the decision trees that grew using a random subset of data and random choices of features at each branch. The feature importance is being sorted based on the importance of the feature.

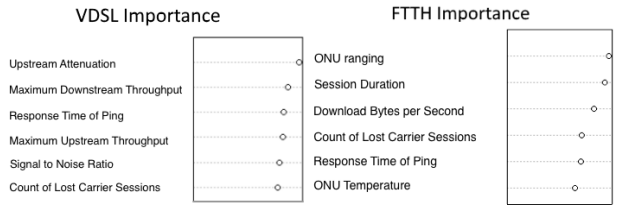

Fig. 7: Feature importance based on Random Forest

Figure 7 shows the top 6 important features identified by the RF algorithm. RF estimates that the VDSL line upstream attenuation is a good indicator, followed by the maximum downstream throughput and the response time of the ping. On the contrary from FTTH, RF suggested that the Optical Network Unit (ONU) ranging is the most important feature followed by the average duration of user internet usage session and the average download bytes of the sessions. RF suggested that the type of customer whether they are home or business to be not relevant to the prediction.

5. Conclusions

In this paper, the commonly used data for network fault prediction, customer trouble tickets are augmented to include more variables which are related to the individual customer, the internet usage data which describes the customer behaviour and the network signal quality data which describes the quality of the signal of the network components associated with the user.

By applying the sliding window of 7 days as the analysis window, the data can be trained by C5.0 decision tree and random forest. The results show that the random forest algorithm is able to predict the network fault with high accuracy. Moreover, the random forest algorithm is able to estimate the variables that are important for prediction. C5.0, on the other hand, could present us with the expression that leads to the predicted outcome, giving an understanding of what could be the potential causes that the network management team could look into. By performing the prediction, broadband service providers can be benefited with better workforce allocation. With the help of the features identified by both prediction algorithm, the network management team could identify the possible cause of the trouble ticket.

The internet usage sessions and the network quality data in the analysis window are aggregated using mean function and that could not capture the outliers or any sudden changes that appear in some of the sessions. Therefore, the future work is focused on using better feature or representation for the internet usage data and network quality data.

Acknowledgement

This work was supported by Telekom Malaysia under the TM Research & Development (TM R&D) Grant.

References

[1] Sonny Yuhensky, Rendy Munadi, et al. Forecasting formulation model for amount of fault of the cpe segment on broadband network pt. telkom using arima method. In Control, Electronics, Renewable Energy and Communications (ICCEREC), 2016 International Conference on, pages 185–191. IEEE, 2016.

[2] Z˘eljko Deljac, Mirko Randic´, and Gordan Krc˘elic´. A multivariate approach to predicting quantity of failures in broadband networks based on a recurrent neural network. Journal of network and systems management, 24(1):189–221, 2016.

[3] Angelos K Marnerides, Simon Malinowski, Ricardo Morla, and Hyong S Kim. Fault diagnosis in dsl networks using support vector machines. Computer Communications, 62:72–84, 2015.

[4] Lam Hai Shuan, Tan Yi Fei, Soo Wooi King, Guo Xiaoning, and Lee Zhe Mein. Network equipment failure prediction with big data analytics. International Journal of Advances in Soft Computing & Its Applications, 8(3), 2016.

[5] Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Philip, S.Y. & Zhou, Z.H (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

[6] Du, M., Wang, S. M., & Gong, G. (2011). Research on decision tree algorithm based on information entropy. In Advanced Materials Research (Vol. 267, pp. 732-737). Trans Tech Publications.

[7] Yang, B. S., Di, X., & Han, T. (2008). Random forests classifier for machine fault diagnosis. Journal of mechanical science and technology, 22(9), 1716-1725.

[8] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[9] Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on knowledge and Data Engineering, 17(3), 299-310.