Network fault prediction using MRMR and Random Forest classifier

Tawsif K.*, Hossen J., Emerson Raja J.

Faculty of Engineering and Technology, Multimedia University, Bukit Beruang, 75450, Melaka, Malaysia

*Corresponding e-mail: tawsif.online@gmail.com

Keywords: Network Fault; Machine learning

ABSTRACT – Day by day the world is being digital and the dependency on internet is increasing. As internet became likely one of the basic demands of citizens, so internet connection should be uninterrupted. In this paper, an approach using MRMR (Minimum Redundancy Maximum Relevance) for feature selection and Random Forest Classifier is proposed to predict network failures. This approach predicts internet session failure based on current state of session data once in a day. The merits of our approach is demonstrated by accuracy, precision and recall.

1. INTRODUCTION

Establishment of smart city and smart homes requires continuous uninterrupted internet connection and more bandwidth to be functional. To solve the issue network monitoring should be real-time and predictive, so that ISP (Internet Service Provider) can take action before occurring any failure. The traditional process of telecom service that requires deployment of hardware devices in telecom network is costly and slow. Network softwarization is cost effective and efficient approach for telecom industry to provide advance services [1]. The two key technologies of network softwarization are SDN (Software-Defined Network) and NFV (Network Function Virtualization) [2]. SDN is popular not only for reducing OPEX/CAPEX but also introducing feature of Datacenter Interconnection (DCI) [3].

2. METHODOLOGY

According to our approach, this proposed system will predict network failure based on current log of user session. In our dataset, we have the session data retrieved at every end of days from every user's connection. Using Random Forest classifier, we trained our model to predict the network failure. At the beginning in our dataset we had seven classes to classify, but out of them only 'Lost-Carrier' is human independent and responsible for the most session termination.

Table 1 Terminate cause

Terminate-cause	Occurrence (%)
Lost-Carrier	55%
Admin-Reset	10%
NAS-Error	0.1%
NAS-Request	18%
User-Request	16%
Session-Timeout	0.2%
Port-Error	0.7%

As presented in Table 1, "Terminate-cause" indicates the reason of internet session termination and occurrence contains percentage of failure for particular reason, according to our observation 'Lost-Carrier' has the highest number of occurrence (55%) and our model predicts whether it's going to be fault because of 'Lost-Carrier' or not.

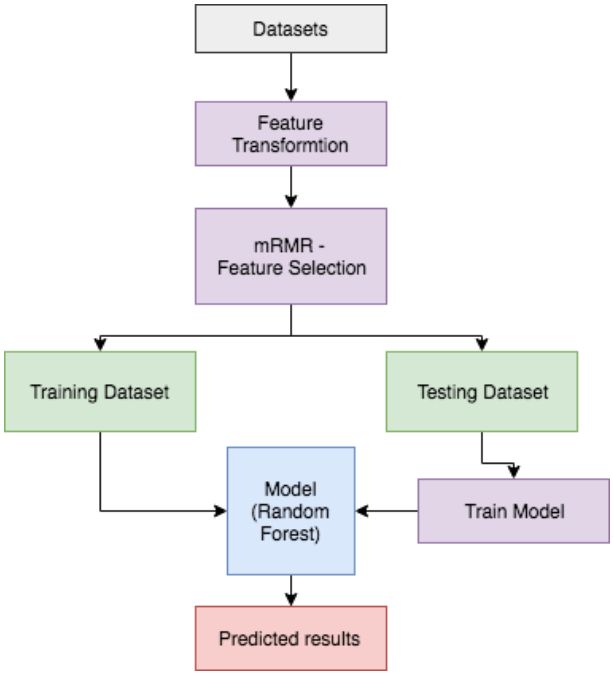

Figure 1 Methodology

2.1 Feature Transformation & Selection

I. Feature Transformation:

Few attributes from dataset we have chosen to apply on feature selection have string values holding categorical values. We encoded the string columns of labels to column of label indices. The indices are ordered by label frequencies, so the highest frequent label gets index 0 and the lowest frequent label will get numLabels. That means, the indices are in [0, numLabels].

II. Feature Selection:

In classification applications, selected attributes from datasets are used as the input to the classification algorithm. In our application we used MRMR algorithm which selects the closest relevant attributes with class label [4]. It is a filtering method trying to minimize redundancy among attributes [5]. This algorithm uses mutual information, I(x,y) to check similarity level

between an attribute and class label vector or two attributes. Mutual information is defined as:

$$I(X,Y) = \sum_{i,j} \left(p(xi, yi) \log \frac{p(xi,yi)}{p(xi)p(yi)} \right) \quad (1)$$

$p(x_i)$ & $p(y_i)$ are marginal probability function and $p(x_i,y_i)$ is joint probability distribution. MRMR algorithm wants to meet two conditions to select attributes. They are maximum relevance, maxMR and minimum redundancy, minMV.

$$maxMR, \quad MR = \frac{1}{|S|} \sum_{i \in S} I(h,i) \quad (2)$$

$$maxMV, \quad MV = \frac{1}{|S|2} \sum_{i,j \in S} I(i,j) \quad (3)$$

where h = {h_1, h_2,…, h_k}.

In the implemented study, Mutual Information Difference (MID) performs the feature selection, which is defined as max(MR - MV) [6].

2.2 Classification

Random Forest

During the classification process, multiple decision tree classification is used because of increasing number of classification value. The strength of the individual trees in the forest is responsible for generalization error of a forest of tree classifier [7]. The combination of all the tree predictors is the final classification of the algorithm. The class with highest votes are made from each decision tree in the decision forest is the final decision and test data is included in the class.

3. RESULT AND DISCUSION

The dataset is used for the implemented system is collected from Telecom Malaysia. We normalized all the datasets and identified effective attributes for network failure. Then, as shown in Figure 1, we transformed the features and executed MRMR for feature selection. Using Random Forest Classifier, we trained the model with selected features.

Table 2 Evaluation

Method	Formula	Percentage
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN} * 100$	95.44%
Precision	$\frac{TP}{TP+FP} * 100$	95%
Recall	$\frac{TP}{TP + TN} * 100$	95%

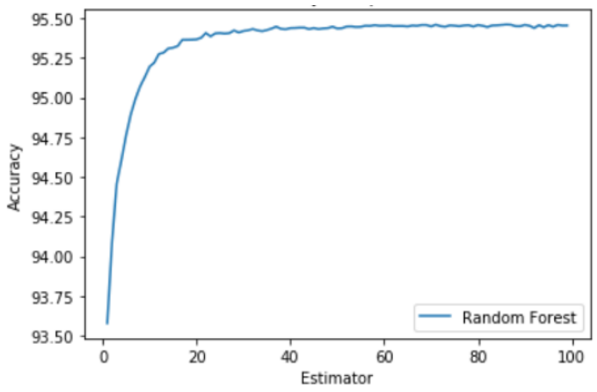

Figure 2 Accuracy Graph

We evaluated the performance by confusion matrix, see Table 2. Where, TP, TN, FP and FN represents True Positive, True Negative, False Positive and False Negative respectively. The accuracy is increased with the number of estimator as shown in Figure 2. We implemented our application with 60 estimators. Because, after that the accuracy achieved stability.

Therefore, following the aim of this research, the implemented study can predict network faults to resolve the issue internet terminations.

4. CONCLUSION

In this research, a new approach is introduced to get prediction daily based on session log status. Based on analysis and MRMR algorithm selected features and Random forest classifier performed a good result with 95.44% accuracy. This approach can improve customer satisfaction if ISP take action on time according to the predictions. Ensemble learning with multiple techniques can be applied for better result in future.

REFERENCES

[1] Wang, T. H., Chen, Y. C., Hsu, C. M., Hsu, K. S., and Young H.C., (2017). Design and implementation of a service-oriented network provisioning system for network as a service, *19th Asia-Pacific Network Operations and Management Symposium (APNOMS)*, 280–283.

[2] Kind, M., Szabó, R., Meirosu, C., and Westphal, F.-J., (2015). Softwarization of carrier networks, *it - Inf. Technol. 57(5)*.

[3] Open Networking User Group (ONUG). (2014) Software-Defined WAN Use Case, *ONUG SD-WAN Work. Gr.*, 1-10.

[4] Peker, M., Arslan, A., Sen, B., Celebi, F. V., and But, A., (2015). A novel hybrid method for determining the depth of anesthesia level: Combining ReliefF feature selection and random forest algorithm (ReliefF+RF), *2015 Int. Symp. Innov. Intell. Syst. Appl. Proc. (INISTA)*.

[5] Ding, C. and Peng, H., (2003). Minimum redundancy feature selection from microarray gene expression data, *CSB2003. Proc. 2003 IEEE Bioinforma. Conf. CSB2003, 3(2)*, 523–528.

[6] Lu, Z., Zhao, Q., and Yang, L., (2009). A Segmentation MEmOD For Crossing Ambiguity String Based On Mutual Information And T-TES T Difference, *2009 IEEE Youth Conference on Information, Computing and Telecommunication*, 371–374.

[7] Breiman, L., (2001). Random Forests, Stat. Dep. Univ. Calif. Berkeley, 1–33.