International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166
Volume 5 Issue 4 July 2014

NETWORK FAULT DETECTION - A CASE FOR DATA MINING

Poonam Chaudhary & Vikram Singh
Department of Computer Science
Ch. Devi Lal University, Sirsa

ABSTRACT: Parts of the general network fault management problem, namely, fault detection, isolation, and diagnosis has been taken up in this communication. A model has been proposed to cluster the network fault data using k-mean algorithm followed by its classification through C4.5 algorithm. Side by side the clustered data are used for training using neural network. The proposed model results into betterment in terms of classification of large data set, decreased network faults, enhanced accuracy.

Keywords: K-mean, self organizing maps, data mining, J48, C4.5.

1. INTRODUCTION

Modern mobile communication networks are tremendously composite systems, usually capable of limited self-diagnosis. The self-diagnostic feature generates huge amount of data in terms of status and alarm messages (Fawcett & Provost, 1997). Alarm signals may be false (called false positives) or true. Accordingly, they need to be processed automatically in order to identify true network faults in a timely manner so as to avoid performance degradation of the network. The proactive response is essential to maintaining the reliability of the network. Sheer volume of the data and because a single fault may cause cascading effects, seemingly unrelated renders the network fault isolation a quite difficult task. This is where data mining comes in to picture in generating rules for identifying and classifying network faults (Han, J. et al., 2002).

1.1 Mobile network faults

Mobile network fault can be defined as an abnormal operation or defect at the component, equipment, or sub-system level that is significantly degrades performance of an active entity in the network or disrupts communication. Each and every error is not faults as protocols can mostly handle them. Mostly faults may be indicated by an abnormally high error rate. The fault can be defined as an inability of an item to perform a required function (a set of processes defined for purpose of achieving a specified objective), excluding that inability due to preventive maintenance, lack of external resources, or planned actions.

Researchers and practitioners and don't agree on a generally accepted description of what constitutes behaviour of a normal mobile network fault (Hajji, et al., 2001; Hajji & Far, 2001). However, majority of the researchers tends to characterize the network faults by 1) transient performance degradation, 2) high error rates, 3) loss of service provision to the customers (i.e., loss of signal loss of connection, etc), delay in delivery of services and getting connectivity.

The main causes of network faults differ from network to network. Managing complex hardware and software systems has always been a difficult task. The Internet and the proliferation of web-based services have increased the importance of this task, while aggravating the problem (faults) in at least four ways (Meira, 1997; Thottan & Ji, 1998; Hood & Ji, 1997; Lazar et al., 1992).

In the present work, a model of data mining tool has been designed and developed for network fault detection in mobile communication. To model has been implemented in a simulated environment on Ericson data set of the past history of network fault in mobile communication. Typical data mining techniques of classification, clustering, and training have been used to produce results that have been analysed to determine parameter dependencies and various network anomalies.

2. PROPOSED MODEL

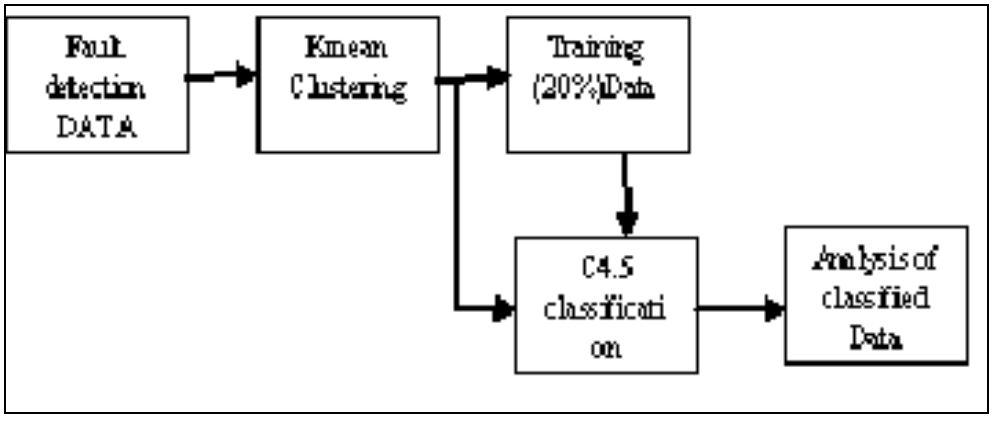

Figure 1. Proposed model showing blocks therein:

Figure 1 above describes the five-block model proposed in this research endeavor. Working of various building blocks of the models has been described in the following sub-sections.

2.1 Fault Detection Data

A Datasheet from Ericson has been taken that consists of the various problems caused in the network and the reason due to which the problem is caused. The data consists is large, still to enhance the data some entrees are replicated. The data set consists of the various alarm and their child alarms.

2.2 K-mean Clustering

The C4.5 algorithm is suitable to classify the small dataset. That's why k-mean clustering is used to arrange the data in the form of clusters. K-mean clustering is performed as

1. Determine the centroid coordinates
2. Determine the distance of each object from the centroid by using the formulae

   1. $$d_{ij} = \sqrt{\sum_{k=1}^{N} \left(x_{ik} - x_{jk} \right)^2}$$
3. Group the object based on minimum distance (find the closest centroid)

2.3 Training using neural network

Training has been applied to the each cluster.20% of the total trained data of each cluster has been using the self organizing maps.

2.4 C4.5 Classification

The classification is performed on each cluster by using the c4.5 decision tree algorithm. The attribute selection is on the basis of the training data. The steps involved in C4.5 classification are:

1. create a node *N*
2. **if** *samples* are all of the same class C
   then return *N* as a leaf node labeled with the class C;
3. **if** *attribute-list* is empty
   then return *N* as a leaf node labeled with the most common class in *samples*;//majority voting
4. select *test-attribute*, the attribute among *attribute-list* on the basis of trained data.
5. label node N with *test-attribute*;

```
6. for each known value ai of test-attribute; grow a branch
   from node N for the condition test-attribute = ai;
7. let si be the set of samples in samples for which test-
   attribute = ai; partition
8. if si is empty
   then attach a leaf labeled with the most common class in
   samples;
   else attach the node returned by Generate_decision_tree
   (si, attribute-listtest-attribute);
```

2.5 Data Analysis

The classification data has been analysed on the basis of Tp rate, Fp rate, and accuracy (precision, recall, and ROC) performance metrics for estimating the accuracy of a given classification model.

2.5.1 Accuracy (precision and recall)

The general percentage accuracy as a performance measure has been proven to be misleading. For example, a classifier that labels all regions as the majority class will achieve an accuracy of 94%, because 96% of the majority may belong to that region. Meanwhile, the classifier may have incorrectly classified some of the minority class instance as the majority because of the bias nature of the dataset but may appear to be accurate. It is therefore imperative to compare the accuracy using an alternative method - Precision and Recall.

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\% \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \times 100\% \quad (2)$$

Where, TP, TN, FP, and FN are as represented in the confusion matrix. Precision in this context refers to the actual percentage of responses to mails that were predicted by the classification model, which translates into the returns on cost of mailing. The Recall, on the other hand, measures the percentage of customers that were identified and needed to be targeted.

2.5.2 Accuracy (ROC analysis)

An alternative method that was used to evaluate classifier performance is the Receiver Operating Characteristic (ROC) analysis (Flach and Gamberger, 2001). It compares visually the performance of the classifier across the entire range of probabilities. It shows the trade-off

between the false-positive rate on the horizontal axis of a graph and the true-positive rate on the vertical axis. From the values obtained from the confusion matrix above, the true-positive rate and false positive rate could be defined as equations (3) and (4) respectively

$$\frac{TP}{TP+FN} \quad (3)$$

$$\frac{FP}{FP+TN} \quad (4)$$

As a standard method for evaluating classifiers, the primary advantage of ROC curve is that they are used to evaluate the performance of a classifier independent of the naturally occurring class distribution or error cost.

3 RESULTS

Ericson data set has been used for comparing the performance of proposed tool vis-à-vis J48 and MLP tools available in the public domain. Tables 1, 2 and 3 list the results of performance of three tools when run on the same test data.

Table 1: Parameter Analysis using J48

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	1	0.458	1	0.629	0.406	Major
	0	0	0	0	0	0.182	minor
	0	0	0	0	0	0.425	Minor
	0	0	0	0	0	0.212	Critical
Weighted Avg.	0.458	0.458	0.21	0.458	0.288	0.374	

Table 2: Parameter Analysis using MLP

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.636	0.692	0.438	0.636	0.519	0.439	Major
	0.5	0.023	0.667	0.5	0.571	0.733	minor
	0.235	0.29	0.308	0.235	0.267	0.49	Minor
	0	0	0	0	0	0.233	Critical
Weighted Avg.	0.417	0.422	0.365	0.417	0.38	0.46	

Table 3: Parameter Analysis using new Algorithm

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.182	0.231	0.4	0.182	0.25	0.51	Major
	0.5	0.091	0.333	0.5	0.4	0.807	minor
	0.235	0.194	0.4	0.235	0.296	0.566	Minor
	0.4	0.465	0.091	0.4	0.148	0.544	Critical
Weighted Avg.	0.25	0.23	0.362	0.25	0.268	0.558	

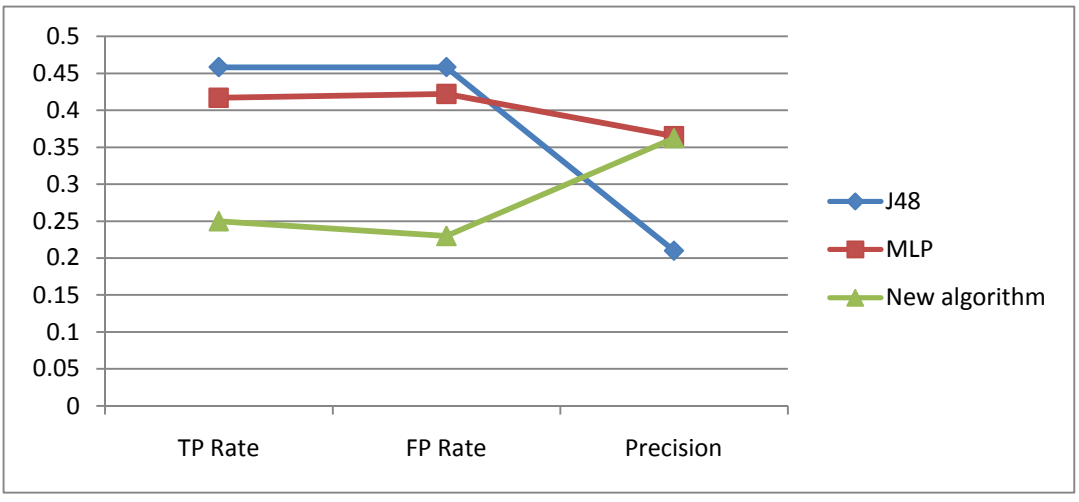

Figure 2. TP rate, FP rate and Precision of J48, MLP and proposed algorithm.

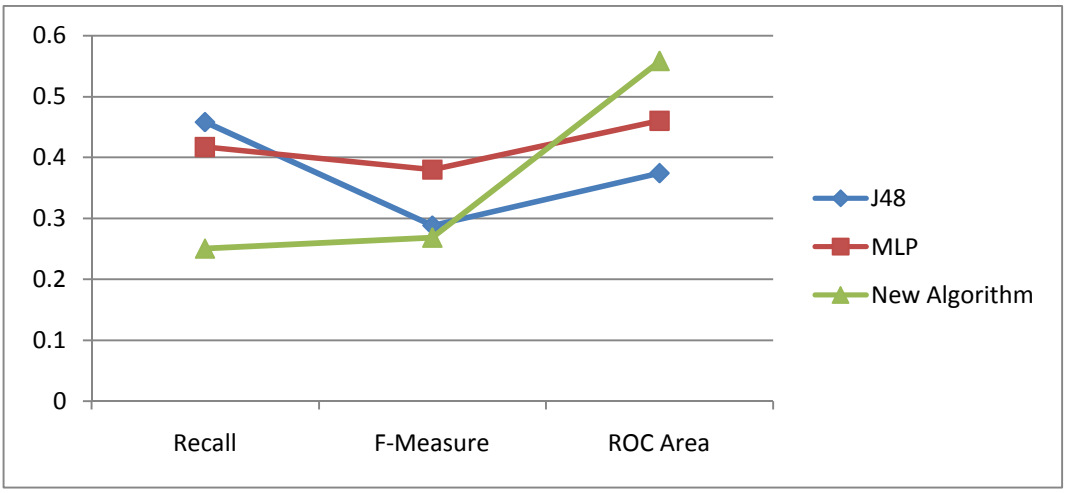

Figure 3. Recall, F-Measure, ROC area of J48, MLP and proposed algorithm.

4. CONCLUSION

A new model has been proposed and implemented in the WEKA environment. The results of proposed model have been compared with two existing techniques, namely, J48 and MLP. All the three techniques were fed the same data set from Ericson. Concepts of k-mean clustering, neural network training and J4.5 classification have been used in the proposed model.

From the results contained in the above listed tables it can be made out that proposed tool yields lesser true-positive rates of 0.25 against the 0.458 and 0.417 of J48 and MLP respectively. Similarly, lesser false-positive rates of 0.23 have been yielded against the 0.458 and 0.422 of J48 and MLP respectively. As far precision rate goes the performance of proposed tool (0.362) is comparable to MLP (0.365) whereas it lesser than J48 (0.21). Recall rate of proposed tool (0.25) is far less than both J48 (0.458) and MLP (0.417).

REFERENCES

Lazar, A.; Wang, W. and Deng, R., “Models and algorithms for network fault detection and identification: A Review”, Proceedings of IEEE ICC, Singapore, pp.999-1003, November 1992.

Provost, F., Fawcett, T., and Kohavi, R., "The Case Against Accuracy Estimation For Comparing Classifiers", in Proceedings of the 15th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann, 1998.

Provost, F. and Fawcett, T., "Robust Classification for Imprecise Environments", Machine Learning 42(3), p 203-231, 2001.

Flach, P. and Gamberger, D., "Subgroup Evaluation And Decision Support For A Direct Mailing Marketing Problem", Aspects of Data Mining, Decision Support and Meta- Learning, [Online] available from: http://www.informatik.unifreiburg.de/~ml/ecmlpkdd/WS-Proceedings/w04/paper8.pdf, 2001.

Meira, D. M., "A Model for Alarm Correlation in Telecommunications Networks", PhD Thesis, Federal University of Minas Gerais, Belo Horizonte, Brazil, Nov. 1997.

Thottan, M. and Ji, C., "Proactive Anomaly Detection Using Distributed Intelligent Agents" IEEE Network, Sept./Oct. 1998.

Hood, C. S. and Ji, C., "Proactive Network Fault Detection", Proceedings of the IEEE INFOCOM, pp. 1139-1146, Kobe, Japan, April 1997.

Hajji, B. and Far, B. H., "Continuous Network Monitoring for Fast Detection of Performance Problems", Proceedings of 2001 International Symposium on Performance Evaluation of Computer and Telecommunication Systems, July 2001.

Fawcett, T. and Provost, F. ,"Adaptive fraud detection. Data Mining and Knowledge Discovery" 1997; 1(3):291-316.

Han, J., Altman, R. B., Kumar, V., Mannila, H. and Pregibon, D., "Emerging scientific applications in data mining". Communications of the ACM 2002; 45(8): 54-58.

Hajji, B., Far, B. H. and Cheng, J.,"Detection of Network Faults and Performance Problems", Proceedings of the Internet Conference, Osaka, Japan, Nov. 2001.