Literature Survey of Network Anomaly Detection

Zhang Shiwei | June 2018

1 Paper Review

1.1 Network Fault Diagnosis Using Data Mining Classifiers [1]

This paper was presented in AIRCC, 2015 by Eleni Rozaki from the Cardiff University.

The first section describes the FCAPS framework and the position of their contribution under that framework. The FCAPS framework stands for fault, configuration, accounting, performance, and security. Their work focus on fault diagnosis.

The second section is the general process of data mining, i.e., data cleaning, section, pattern mining, and knowledge representation. They use Weka to perform the mining.

In the next section several data mining techniques were explained and compaired:

J48 tree (more commonly known as C4.5). It builds decision trees by maximizing information gain greedly at each node.[2]

LAD tree Inducing ADTrees using LogitBoost. An ADTree consists of an alternation of decision nodes, which specify a predicate condition, and prediction nodes, which contain a single number. An instance is classified by an ADTree by following all paths for which all decision nodes are true, and summing any prediction nodes that are traversed.[3]

JRip Alternatively grow and prune rules to build an initial rule set in terms of information gain, Then examine each rule by generate two variants of each rule from randomized data, see which have shorter descrition length.[4]

PART Generating a decision list by buiding a C4.5 decision tree in each iteration and makes the "best" leaf into a rule. Instances are classified at the first match.[5]

Naïve Bayes Using Bayes rule to calculate the conditional probability with the assumption that all attributes are independent of each other.[6]

Bayesnet Also known as belief networks. It use Bayes rule recursively in a DAG to infer the probabilities of the state of a node.[7]

In the fourth section some definitions are given. The most important concept is KPI, which acts as the target value to predict. They define KPI as a variable takes 3 possible values: Normal, Critical and Warning. The value of KPI is determined by DCR (Call Drop Rate), CSSR (Call set up success rate), TR (Traffic Rate), and HOF (Handover Faulures) empirically.

In the fifth and sixth sections the authors showed their results by screenshots of Weka outputs, and made several comparisons between above algorithms.

1.2 Detecting and Localizing End-to-End Performance Degradation for Cellular Data Services [8]

This paper was presented in INFOCOM, 2016 by Michigan State University and AT&T.

Firstly they stated the goal, which is mainly to ascribe E2E performance degradations to one of the four factors: application type, content provider, mobile device, and user location.

Next they gave an overview of their method. The first step is to build 24 * 7 models that predicting the performance of the instances correspond to a specific hour in a week. Then, use these predictions to define degardation. Finally use association rules mining to find patterns that cause degradations.

In the rest of Section 1 they described the 3 main challenges and their solutions. The first challenge is data sparsity. They use recursive grouping to handle this. The second challenge is to localize the cause of degardation, which is what they deploy the association for. The last one is to quantitatively evaluate the result, they solve this by mannually inspect some cases and inject synthetic cases which act as ground truth.

In the second Section the authors discussed some related works in network diagnosis and performance measurement.

2 Relative Work Summary

3 My Proposal

References

[1] Eleni Rozaki. Network fault diagnosis using data mining classifiers. pages 29–40. Academy & Industry Research Collaboration Center (AIRCC).

[2] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

[3] Geoffrey Holmes, Bernhard Pfahringer, Richard Kirkby, Eibe Frank, and Mark Hall. Multiclass alternating decision trees. In *Machine Learning: ECML 2002*, Lecture Notes in Computer Science, pages 161–172. Springer, Berlin, Heidelberg.

[4] William W. Cohen. Fast effective rule induction. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, ICML'95, pages 115–123. Morgan Kaufmann Publishers Inc.

[5] Eibe Frank and Ian H. Witten. Generating accurate rule sets without global optimization. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 144–151. Morgan Kaufmann Publishers Inc.

[6] George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 338–345. Morgan Kaufmann Publishers Inc.

[7] R. Barco, V. Wille, L. Diez, and P. Laizaro. Comparison of probabilistic models used for diagnosis in cellular networks. In *2006 IEEE 63rd Vehicular Technology Conference*, volume 2, pages 981–985.

[8] Mr Faraz Ahmed, Mr Jeffrey J Erman, Dr Zihui Ge, Alex X Liu, Dr Jia Wang, and Dr He Yan. Detecting and localizing end-to-end performance degradation for cellular data services. page 9. IEEE.