International Journal of Engineering &Technology, 7 (3.12) (2018) 178-179

SPC

International Journal of Engineering & Technology

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

Research paper

Network Anomaly Detector using Machine Learning

K.M. Uma Maheswari[1*], Ashwin Pranesh[2], S. Govindarajan[3]

[1]*Assistant Professor (Sr. G), Department of CSE, SRM Institute of Science & Technology, Kattankulathur, Tamil Nadu, India*
[2]*UG student, Department of CSE, SRM Institute of Science & Technology, Kattankulathur, Tamil Nadu, India*
[3]*Professor, Department of EDP, SRM Institute of Medical Sciences, Kattankulathur, Tamil Nadu, India*

Abstract

The 4G network consists of a network of routers on each tower that decides where a certain packet must be switched to. These routers like any other hardware device is subject to failure due to number of factors such as threshold violations and problems with its tuning. The routers and other relevant hardware devices undergo various maintenance cycles that can sometimes be wasteful as the hardware may be replaced even when in complete working condition. This is a measure taken to ensure the network is always up and running. This measures has proven to be expensive and alternative solutions have been looked for. To alleviate the costs involved in the maintenance of these routers, a system will be developed to perform applications such as report failures, find the root cause and implement a remedial action automatically.The prediction of failures in the routers is achieved by unsupervised machine learning while will be trained to pick up anomalies from a continuous stream of log messages sent to system which is then analyzed. The anomaly data is then used to schedule maintenance runs more effectively.

Keywords: *syslogs, detector, predictor, random forest, supervised*
.

1. Introduction

The Jio network consists of a large number of hosts spread across the country. The hosts send an endless stream of syslog messages, With various types of data such as hardware data like fan speed, host temperature or software data such as packets forwarded, packets dropped, etc. These log messages are used to continually monitor the status of the hosts. The log data are also processed accordingly and used to perform various automatic maintenance tasks. One such task is the anomaly detector. The anomaly detector utilizes a supervised machine learning process to detect anomalous values in log messages from hosts. The network anomaly detector works on pooled syslog data from various hosts. This detector is proposed to work once every week on syslogs collected for a whole week. The detector trains a machine learning component that will learn to detect and point out anomalous values syslogs. These values are detected to go beyond a threshold that the detector has learned to be normal. The entire platform consists of various modules that work together to create a system that automatically performs host maintenance. The dotted lines are the scope of this project. The platform consists of a module that pools syslog messages from all the hosts across the country and makes them available for analysis. This data is used to perform anomaly detection. This utilizes a machine learning algorithm that learns the log values that appear to be "normal" and point out values that appear to go below or beyond the threshold that has been learned.
The anomalies detected and then passed on to another module that points out the host that was responsible for the failure. The required remedial action is then performed. Router behavior is monitored through log messages. Each host keeps sending a steady stream of log messages that details the condition of various hardware and software components.

Each circle in the Jio network consists a number of hosts and the entire network is encompassed with numerous circles. Each circle consists of its own syslog server which constantly collects the data. Each of these servers transmit all the syslog data to a central bus cluster or Apache Kafka cluster. All the syslog data is pooled in the bus cluster with a default expiry of 2 weeks. The syslogs are retrieved from the Kafka cluster for analysis.

2. Random Forest Clustering

A random forest (RF) predictor is an ensemble of individual tree predictors. As part of their construction, RF predictors naturally lead to a dissimilarity measure between the observations.
One can also define an RF dissimilarity measure between unlabeled data: the idea is to construct an RF predictor that distinguishes the "observed" data from suitably generated synthetic data. The observed data are the original unlabeled data and the synthetic data are drawn from a reference distribution. An RF dissimilarity can be attractive because it handles mixed variable types well, is invariant to monotonic transformations of the input variables, and is robust to outlying observations. The RF dissimilarity easily deals with a large number of variables due to its intrinsic variable selection.
Machine learning methods are often categorized into supervised (outcome labels are used) and unsupervised (outcome label are not used) learning methods. Interestingly, many supervised methods can be turned into unsupervised methods using the following idea: one creates an artificial class label that distinguishes the

"observed" data from suitably generated "synthetic" data. The observed data are the original unlabeled data and the synthetic data are drawn from a reference distribution. Some supervised learning methods distinguishing observed from synthetic data yield a dissimilarity measure that can be used as input in subsequent unsupervised learning methods (Liu, Xia, and Yu 2000; Hastie, Tibshirani, and Friedman 2001; Breiman and Cutler 2003). Breiman and Cutler (2003) proposed using random forest (RF) predictors to distinguish observed data from synthetic data. When the resulting RF dissimilarity is used as input in unsupervised learning methods (e.g., clustering), patterns can be found which may or may not correspond to clusters in the Euclidean sense of the word. The RF dissimilarity has been successfully used in several unsupervised learning tasks involving genomic data: Breiman and Cutler (2003) applied RF clustering to DNA microarray data; Allen et al. (2003) applied it to genomic sequence data; and Shi et al. (2005) and Seligson et al. (2005) applied it to tumor marker data. In these real data applications, the resulting clusters often made biological sense, which provides indirect empirical evidence that the method works well in practice.

3. Random Forest Pseudo Code

Precondition: A training set S := (x1, y1), . . . ,(xn, yn), features F, and number of trees in forest B.

1. function RandomForest(S , F)
2. H ← ∅
3. for i ∈ 1, . . . , B do
4. S (i) ← A bootstrap sample from S
5. hi ← RandomizedTreeLearn(S (i) , F)
6. H ← H ∪ {hi}
7. end for
8. return H
9. end function
10. function RandomizedTreeLearn(S , F)
11. At each node:
12. f ← very small subset of F
13. Split on best feature in f 14 return The learned tree 15 end function

4. Proposed Implementation

The random forest clustering method will be used on various log values in the pool of sys logs. They will be clustered to detect values that deviate from the norm. Various log values will be clustered in this method and values in these logs that deviate from the norm will be detected. The method will adapt to the values presented to it and appear to shift its threshold values appropriately.

5. Conclusion

In this paper, the Random Forest clustering method is used to analyze large amounts of unlabeled log data and detect values that deviate and may be cause of an anomaly. This data is then presented graphically for further analysis.

References

[1] https://nishanthu.github.io/articles/ClusteringUsingRandomForest.html
[2] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#intro
[3] https://mapr.com/ebooks/spark/08-unsupervised-anomaly-detection-apache-spark.html
[4] https://labs.genetics.ucla.edu/horvath/RFclustering/RFclustering.htm
[5] https://labs.genetics.ucla.edu/horvath/RFclustering/RFclustering/RandomForestHorvath.pdf
[6] http://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/ensembles/RandomForests.pdf