2019 5th International Conference for Convergence in Technology (I2CT)
Pune, India. Mar 29-31, 2019

Analysis of Network log data using Machine Learning

Shridhar Allagi
Assistant Professor
Department of Computer Science and Engineering
Jain College of Engineering
Belagavi,Karnataka, India
shridharallagi1@gmail.com

Rashmi Rachh
Associate Professor
Department of Computer Science
Center of PG Studies,VTU
Belagavi, Karnataka, India
rashmirachh@gmail.com

***Abstract*: The proliferation of web base usage has also resulted in an escalation in unauthorized network access. In this scenario, it is imperative to periodically analyze log records of the network so that malicious users can be identified. This process can be automated using machine learning techniques. In this paper, analysis of log records of a network is carried out using supervised machine learning techniques. K-means and Self organizing feature map (SOFM) algorithms have been used with the data set obtained from the UCI machine learning repository. An accuracy of 97.2% has been archived.**

***Keywords:* Machine Learning, Supervised Learning, K-means, SOFM.**

I. INTRODUCTION

Proliferation of web base usage has also resulted into escalation in unauthorized net-work access. In this scenario, it is imperative to periodically monitor log records of the network. The web servers are prone to various cyber-attacks so as to extract the in-formation about the internal architecture and various other details including the user access pattern and internal administrator account details. The periodic monitoring of the log recording is tedious and complex process which could be automated. Machine learning techniques are getting popularity recently because of their ability to make better decisions without human intervention Thus, job of analyzing the logs at large scale can be done more efficiently using machine learning techniques. The machine learning classification and clustering algorithms can be used to label a particular entry in log as normal or abnormal. The most challenging task is to get genuine log data set for analysis. Most often, the log data set are huge number of text messages in un-structured format. Even in the medium size firm, periodical log data set involves mil-lions of entries. Secondly, the log data format varies from service to service and organizations. The log data set entries are very diverse [1], they have a huge information ranging from one login event to critical system failures. Hence, in order to analyze logs, it is essential to generate a static template for analysis.

In our study, we are using company generated network log data set. The log data set contains several million entries. Only 5% of which resembles as abnormal behavior. This imbibes our system to class imbalance problem where one particular type of cluster is very minor compared to other type class. We propose to use k-means (k=2) resulting in two cluster normal and abnormal along with self-organizing feature map (SOFM) of artificial neural network consisting of neurons in input layer and output layer. Our focus is based on textual information in log data and majorly searching of bigram words (Ex: “ACCESS DENIED”, “401 ERROR”, INVALID PASSCODE” etc). We applied supervised machine learning approach and validated system against large volume of another network log dataset. The experimental results showed the accuracy of predicting the abnormal behavior was high and acceptable.

This paper is organized as follows: Section 2 provides state-of the art review of the work carried out in area of network security using various approaches. Section pro-vides the architecture of the proposed system. In section 4, we provides the experimental results. Section concludes the paper.

II. LITREATURE REVIEW

The following sections describes various work carried out by researchers using various machine learning algorithms and methodology implemented. Chaofei Wang et.al [1] focused on reconstruction of log data using association relation. Their approach consist of IP association, behavior association and glossary association. Authors used deep log analysis (DLA) approach in conjunction with various associations of the user’s behavior with the network interaction. Qimin Cao et.al [2] attempted to enhance traditional log analysis methods by using two level machine learning algorithms. The architecture in first level used the decision tree model to classify the normal and analogous data. The second level included the Hidden Markov model (HMM) to classify the anomalies. Biplob Debnath et.al [3] proposed the concept of LogLens. Authors used the unsupervised machine learning algorithms to determine anomalies with respect to stateful and stateless detections. The system typically uses log parser, which works on principal of pattern evaluation in dataset. JIA Zhanpei et.al [4] implemented spark-based log analyzer to analyze large scale log data on big data platforms. They combined machine learning, statistical learning and data mining techniques. Filtering approach before processing was initiated. Tatsuaki Kimura et.al [5] determined the abnormality in log using key word search and the patterns for burstiness. Their system worked on three phases, first converting in to static template, extract features and patterns and later applying supervised ma-chine method.

There are various machine learning algorithms for clustering and classification. Table 1 gives the abstract ideas of some renowned methods.

978-1-5386-8075-9/19/$31.00 ©2019 IEEE

TABLE I. DESCRIPTION OF VARIOUS MACHINE LEARNING CLASSIFICATION AND CLUSTERING METHODS

Machine Learning Algorithm	Description
Support Vector Machine(SVM)	It inherits the kernel level functions to process variety of data.
K-Means	Clusters the data in to various classes based on the computation of their centroids
Decision Tree (DT)	If- then rules are applied to determine the concept.
Naïve Bayes (NB)	Derived from theory of probability and is used for prediction.

III. PROPOSED SYTEM

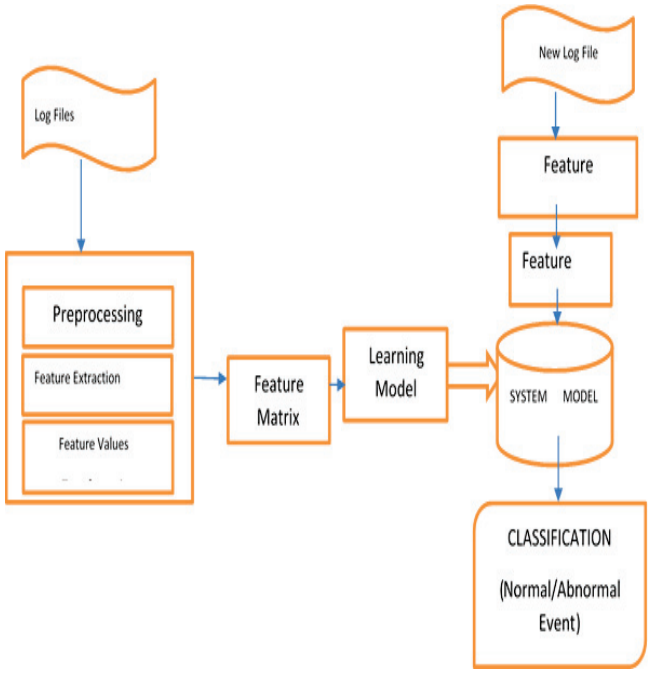

Fig. 1. System Architecture

In our model, we use the network log data set generated by the organization network which is publicly available in UCI Machine learning repository. The system architecture involves various phases which are data collection, data preprocessing, feature extraction and feature value transformations. The dataset involves both textual and numerical values. We mainly focus on the textual patterns in the log set. The initial log data set is an unstructured data, which is converted in to a structured template for efficient analysis. The various multiple attributes in the data set are timestamp, unique identifier of the request, origin IP header, destination IP, protocol, and port number, duration of connection and threat status.

In preprocessing stage, the raw data set is processed in multiple ways by replacing the timestamp with some special symbols. The data set comprises of both digits and characters, the digits are replaced and all the upper cases are converted to lower cases. In this phase, all the special symbols such as punctuations are also eliminated.

Feature extraction phase involves analysis of multiple attributes and extraction of features determining the behavior anomalies in to feature matrix. The attributes timestamp and n-gram are extracted as the feature. In n-gram, we mainly focus on character bigram and word bigram (t=2). The character bigram feature dictionary is created and checked in new log files for anomalies. Similar bigram word dictionary is created with t=2 such as ACCESS DENIED, INVALID PASSWORD, 401 ERROR etc. The attribute in the feature matrix is tf- idf which will determine the count of word in a collection of log imbibing its importance in log set.

In feature transformation, we apply normalization to the values as the values of some attributes may dominant. The values such as duration of session might be fewer seconds or few hours. Only the factors determining the anomalies are preserved and other non-important features such as duration of session in intermediate, IP type are eliminated through dimensionality reduction. The feature matrix created with attribute normalized values is generated where each row x1,x2, … xn represents each row in training data set and y1,y2,…yn represents the features in the training dataset. This feature matrix is used to make the model learn about the patterns for normal and abnormal activities in the log data set. In our system, we view the problem of log analysis as class imbalance problem as only less than 5% of log dataset resembles for anomaly, which is very negligible considered to entire log data set. Our problem can be viewed as classification with clustering domain. Here we apply the k-means (k=2) algorithm to cluster the trained data set to normal and abnormal behavior in log. The self-organizing map (SOM) of artificial network is combined with k means (K=2) to train the model with training data set and generate each log entry to normal or abnormal behavior.

$$FM = \begin{matrix} x1y1 & x1y2 & x1y3 \\ x2y1 & x2y2 & x2y3 \\ x3y1 & x3y2 & x3y3 \end{matrix} \quad (1)$$

The self-organizing feature map (SOFP) is dominant clustering process in artificial neural network which will update the centroids of clusters in topological organization. It has two layers input layer and output layer. Many structures in the map topology are possible such as triangular grid, hexagonal grid etc. The SOM hexagonal grid is shown in figure 2.

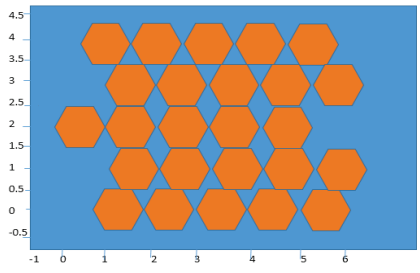

Fig. 2. 5 x 5 Hexagonal SOFM Neural Network

The output neurons are trained with extensive training input. For each of input neuron x, it is clustered in to nearest output neuron y. During the updating in various iterations, the assigned output neuron y is reassigned based on the learning factor in subsequent iterations. The mathematical expression for updating is as below:

$$Wy(t+1) = Wx(t) + \mu f(uj, uk)(pi(t) - wj.x(t)) \quad (2)$$

Where μ is the learning step which increases in increase in time. f (uj,uk) is neighborhood function which determine strength of neighbor uj based on neuron uk. The Gaussian function is used in f (uj,uk).

Once the model is trained with above iterations, the data set comprises of training data set and test data set. The new log data set is tested against trained model by extracting features from new log and using its feature matrix.

IV. EXPERIMENTAL RESULTS

We are using the log data set comprising of 22614256 records. We are using data set in the ratio of 8: 2 for training and testing i.e. out of 100, 80 are for training and 20 are for testing. The dataset comprises of various information such as timestamp, origin IP, intermediate IP, destination IP, port number, error status and other details associated with network incoming and outgoing traffic.

```
>TestOutput=netLog (test1$CLASS, out)
> TestOutput
```

	Actual	Predicated
1115454	1	1
1214545	0	0
1234566	0	1
1239856	1	1
1245686	1	1
1248799	1	1
1252265	0	0
1254777	1	1
1268946	0	0
1269994	0	0
1278546	1	1
1289646	0	0

Fig. 3. Test Results for Dataset

The model was tested for 500 records, and system predicted 486 class records correctly, giving accuracy up to 97.2 %. The figure 3 illustrates the test results for the dataset for clustering a record to normal and abnormal log event. Hence using our model we can do the network log data auditing and analysis in most optimal way.

V. CONCLUSION

In this paper, analysis of log data of a network has been carried to identify anomalies. In access pattern. The model has been trained k -means and SOFM clustering. The proposed model has been tested across sample dataset and has achieved 97.2 % accuracy and false rate of 2.7 %. In this work, only text data of the log records has been analyzed and supervised learning has been used, this work can be further extended for analysis of text and non-text real time data.

REFERENCES

[1] Chaofei Wang, Jing Chen, Xiaopeng Liu, Jinwei Zhao.: AN IMPROVED DEEP LOGANALYSIS METHOD BASED ON DATA RECONSTRUCTION. CCIS2014 978-1-4799-4719-5 /14/$31.00 ©2014 IEEE (2014).

[2] Qimin Cao , Yinrong Qiao, Zhong Lyu.: Machine Learning to Detect Anomalies in Web Log Analysis, 3rd IEEE International Conference on Computer and Communications (2017).

[3] Biplob Debnath, Mohiuddin Solaimani, Muhammad Ali Gulzar, Nipun Arora,Cristian Lumezanu, Jianwu Xu, Bo Zong, Hui Zhang, Guofei Jiang, and Latifur Khan, LogLens: A Real-time Log Analysis System, IEEE 38th International Conference on Distributed Computing Systems (2018).

[4] JIA Zhanpei, SHEN Chao, YI Xiao, CHEN Yufei, YU Tianwen, GUAN Xiaohong: Big-Data Analysis of Multi-Source Logs for Anomaly Detection on Network-based System In 13th IEEE Conference on Automation Science and Engineering (CASE) Xi'an, China (2017).

[5] Tatsuaki Kimura, Akio Watanabe, Tsuyoshi Toyono, and Keisuke Ishibashi, Proactive Failure Detection Learning Generation Patterns of Large-scale Network Logs.

[6] Chenn-Jung Huang, Ming-Chou Liu, San-Shine Chu, Chin-Lun Cheng, Application of Machine Learning Techniques to Web-Based Intelligent Learning Diagnosis System, Fourth International Conference on Hybrid Intelligent Systems (2014).

[7] Ho Chun Leung, Chi Sing Leung, Eric W. M. Wong, and Shuo Li, Extreme Learning Machine for Estimating Blocking Probability of Bufferless OBS/OPS Networks, IEEE Transactions , J. OPT. COMMUN. NETW./VOL. 9, NO. 8/AUGUST 2017.

[8] Fatih Ertam, Mustafa Kaya , Classification of Firewall Log Files with Multiclass Support Vector Machine , 978-1-5386-3449-3/18/$31.00 ©2018 IEEE

[9] Jiexiong Tang, Chenwei Deng, Guang-Bin Huang: Extreme Learning Machine for Multilayer Perceptron, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 27, NO. 4, APRIL 2016.

[10] Yanwei Pang, Manli Sun, Xiaoheng Jiang, Xuelong Li: Convolution in Convolution for Network in Network, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 5, MAY 2018.