JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-8, Aug 2019 E-ISSN: 2347-2693

Classification of Firewall Logs Using Supervised Machine Learning Algorithms

Hajar Esmaeil As-Suhbani[1*], S.D. Khamitkar[2]

[1, 2] Department of Computational Sciences and Technology, S.R.T.M University, Nanded, India

[*]Corresponding Author: hajar.esmaeil@gmail.com, Tel.: +991 8208332074

DOI: https://doi.org/10.26438/ijcse/v7i8.301304 | Available online at: www.ijcseonline.org

Accepted: 16/Aug/2019, Published: 31/Aug/2019

***Abstract*— Most operating systems services and network devices, such as Firewalls, generate huge amounts of network data in the form of logs and alarms. Theses log files can be used for network supervision and debugging. One important function of log files is logging security related or debug information, for example logging error logging and unsuccessful authentication. In this study, 500,000 instances, which have been generated from Snort and TWIDS, have been examined using 6 features. The Action attribute was selected as the class attribute. The “Allow” and “Drop” parameters have been specified for Action class. The firewall logs dataset is analyzed and the features are inserted to machine learning classifiers including Naive Bayes, kNN, One R and J48 using Spark in Weka tool. In addition, we compared the classification performance of these algorithms in terms of measurement metrics including Accuracy, F-measure and ROC values.

***Keywords*— Machine Learning Algorithms, Classification, log analysis, firewall, Spark.

I. INTRODUCTION

In the era of information technology, computer networks constantly change human society, with the accompanying information and network security issues. The networks’ administrators spend a lot of money to purchase network security tools such as anti-virus software and firewalls, etc. Also, they spend a lot of time to ensure the availability and integrity of the networks and retain the confidentiality of network information to prevent attacks from outside or inside the network [1].

With the advent of internet technology, day to day work has been shifted over the internet and thus network security has become a global focus in the world. For that, there is a challenge to the traditional security solutions such as Firewall and VPN to detect security breach against attacks [2].

Firewall is the most important element in computer networks which play an important role in network security. It can either allow or prevent traffic according to a predefined policy by examining the ingoing-outgoing packets. The configuration of firewalls is necessary for computer and communication networks to work securely and properly [3]. Firewalls function as gateway for computer networks. Obviously, firewall is one of the most important components of the network and it should be no inconsistency in the security policies used and should not cause security vulnerabilities [4]. However, managing firewall rules have become complicated, complex, and error-prone [5].

Recently, data mining and machine learning, a sub-branch of artificial intelligence, have gained great and much significant attentions in the information technology, network security and also in information industry mainly due to the existence of a massive amount of data that can be widely used, and the ability to draw on and utilize the useful knowledge and information behind these data [5]. In addition, machine learning approaches are often used to analyze, discover hidden relationships in datasets [4]. Data mining and machine learning can implement classification, clustering, prediction, and association rule mining from the data items.
It is very important to analyze firewall logs on the Firewall devices and control the internet traffic according to the results of these analyzes. In this paper, we proposed an approach using machine learning classifiers including Naive Bayes, kNN, One R and J48 in the parallel processing framework using open-source Apache Spark in Weka. We performed our experiment on firewall dataset generated using Snort and TWIDS in our department.

This paper is organized as follows Section II the related work. Section III illustrates the overall methodology of the experimental studies. The results obtained and discussion are explained and discussed in section IV. The conclusion of the paper is presented in the section V.

II. RELATED WORK

Analyzes of firewall logs using data mining and machine learning approaches are implemented in many literatures and studies. One of these is Golnabi et al. [5] they proposed an approach based on data mining to reduce active firewall rules by analysing around 33,000 log instances. In onther study, Breier and Branišová [6] have proposed a method using Apache Hadoop for anomaly detection in log files based on data mining techniques to create dynamic rules. Ucar, E. & Ozhan, E [7], proposed a model based machine learning to detect anomalies in Firewall rules file by analysing a large-scale log files, approximately 5,000,000 data instances taken from a Firewall. In this experiment it was observed that, the maximum learning performance has been achieved when they reached 1,500,000 data instances of the training data value. However, their performance level began to decrease after this point. In paper [8], Al-Shaer and Hamed, developed a Java-based program called Firewall Policy Advisory in order to detect the rules which caused anomalies in the firewall rule repository, and made it possible to reorder and remove anomalous rules. In a follow up study [9], Al-Shaer et al. developed the same software in [8] to detect more interrelated overlaps between the rules.

Classification of Firewall Logs Using Supervised Machine Learning Algorithms has been implemented in Weka tool with Spark to spped up the overall processing of the system. 500,000 instances, which have been collected from Snort and TWIDS in our department, have been examined using only the 6 major features. The Action attribute was selected as the class attribute. The firewall logs dataset is analyzed and the features are inserted to machine learning classifiers including Naive Bayes, kNN, One R and J48. The proposed experiment has shown a good performance in terms of processing time and accuracy. In addition, the experiment results show that the proposed multi-classifier model is more accurate as compare to other techniques.

III. METHODOLOGY

The main purpose of our model is to use data mining techniques such as feature selection and removing redundancies in order to analyze firewall log dataset. The second purpose is to use machine learning classifiers including Naive Bayes, kNN, One R and J48 using Spark in WEKA [13] to speed up the process. In addition, we compared the classification performance of these algorithms in terms of Accuracy and F-measure. In addition, we used 10-fold cross-validation test. This model has shown a great performance in terms of accuracy. Lastly, we prove that our integrated model based on Supervised Machine Learning Algorithms and Spark is more powerful in classifying firewall logs and processing accuracy than others models. The block diagram of the proposed model is shown in Figure1.

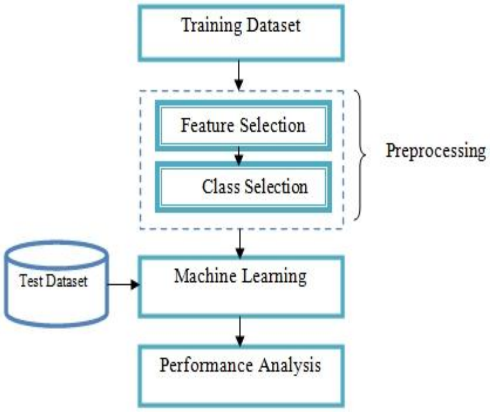

Figure 1 Model Block Diagram

3.1 Dataset

Firstly, log records are collected from the firewall. The Log records used were taken from a firewall implemented by Snort [10] and TWIDS [11] at a lab in our department, as explained in our previous work [12]. The receiving log record consists of 500,000 instances.

3.2 Preprocessing

A preprocessing must be performed before any analysis because the log data could not be used in the form as it is stored in the log files. Preprocessing involves cleaning, loading dataset and manipulating the data into a form that you want to work with.

3.2.1 Feature Selection and class selection

In each log line, the attributes are taken with importance to source and destination IP addresses, source and destination ports, and protocol (TCP or UDP). In addition, we considered only the major attributes as in the packet header and firewall log file. In order to classify the firewall log dataset, only 6 major features were selected: Action (Allow, Deny), Source IP, Source port, Destination IP, Destination port, and Protocol (TCP/UDP). The action attribute with "Allow", "Drop", which is a nominal attributes, has been selected as the class attribute. The "Allow" and "Drop" parameters have been specified for the Action class.

3.3 Classification Phase

The last phase in the study is classification. In this phase the firewall logs dataset is analyzed and the features were inserted to machine learning classifiers including Naive Bayes, KNN, One R and J48 using Spark in Weka tool to speed up the overall process. In addition, we compared the classification performance of these algorithms in terms of Accuracy, F-measure and ROC values.

IV. RESULTS AND DISCUSSION

Experiment results show that the proposed multi-classifier model is more accurate as compare to other techniques. To evaluate the results of the used classification algorithms, we have used standard metrics such as classifiers accuracy, F-measure and ROC values. The performance matrices of all proposed algorithms are listed in Table 1. The accuracy of a proposed model is near about 100% as shown in Figure 2. It was observed that the highest Accuracy value was obtained in the KNN classifier with 99.8736%, and the F-Measure value, it was observed that the best result was achieved with the same classifier with 0.999%.

Table 1 Evaluation Results in terms of performance metrics

Algorithm	Accuracy %	F-Measure	ROC Area
Naive Bayes	99.2582	0.993	0.975
KNN	99.8736	0.999	1.000
One R	99.5238	0.994	0.967
J48	99.8365	0.998	0.981

In the above experiment, the results show the average performance of 10-Cross validation. These models are compared on the basis of each individual fold or rounds. Figure 3 and 4 show the F-Measure and ROC analyzes of different models under 10 folds. In this proposed method, as shown in table 1, KNN algorithm performs better than the remaining classification algorithms proposed in the model.

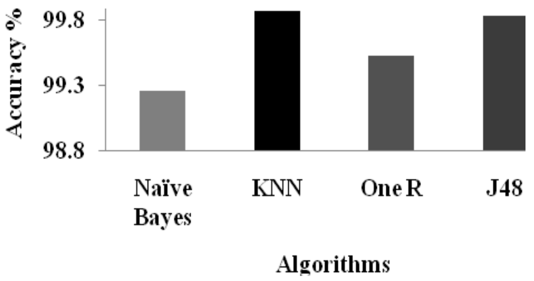

Figure 2 Classifier accuracy % for all classification algorithms

F-Measure is basically used to measure the efficiency of the classifiers. Basically, it is the harmonic mean/average of precision and recall values. It is also known as balanced F-score or traditional F-Measure [14]. Figure 3 show that the proposed multi-classifier model shows better results in every fold.

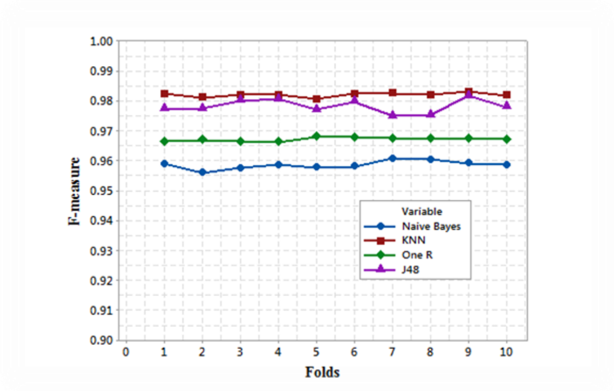

Figure 3 F-Measure values of different algorithms under 10 folds

In addition, (ROC) curves were created for each of the algorithms. Figure 4 shows the result of different algorithms under ROC Area in 10 folds. ROC Area defined the correctness of the classifier that how a normal or abnormal dataset is separated by using training dataset. Hence, more area under the ROC curve shows how the classifier is more accurate. Figure 4 shows that the proposed model covers the maximum area which means it result with the maximum accuracy while classification.

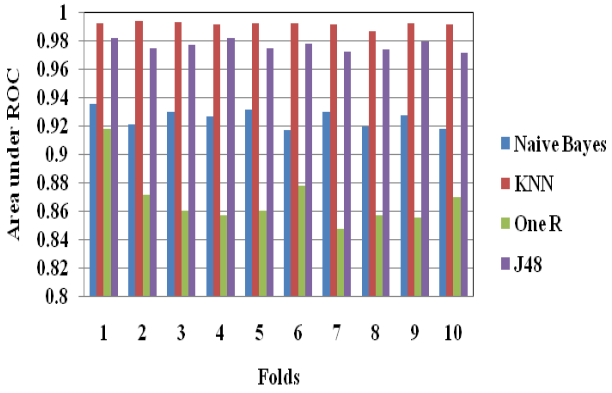

Figure 4 Comparisons of different algorithms under ROC Area in 10 folds

We can conclude that the result of our multi-classifier model was efficient and perfect in terms of accuracy, F-measure and ROC Area, based on four Machine Learning classification algorithms.

V. CONCLUSION AND FUTURE SCOPE

Firewall is the most important components of the network and it should be no inconsistency in the security policies used and should not cause security vulnerabilities. However, managing firewall rules have become complicated, complex, and error-prone. In this experiment, we proposed a multi-classifier model using 4 classifiers including Naive Bayes, kNN, One R and J48 in the parallel processing using Spark in Weka tool. We performed our experiment on firewall dataset generated using Snort and TWIDS in our department. We considered only the 6 major features: Action (Allow, Deny), Source IP, Source port, Destination IP, Destination port, and Protocol (TCP/UDP). The action attribute with "Allow", "Drop has been selected as a class attribute. The experiment results showed that the proposed model is more accurate as compare to currently used techniques. The accuracy of a proposed model is near about 100%. We have used standard metrics such as classifiers accuracy, F-measure and ROC values to evaluate the performance of the model. KNN algorithm showed better performance than the remaining classification algorithms proposed in the model. It was observed that the highest Accuracy value was obtained in the KNN classifier with 99.8736%, and the F-Measure and ROC values, it was observed that the best result was achieved with the same classifier with 0.999% and 1 respectively.

ACKNOWLEDGMENT

We place our heart-felt gratitude to Dr. Khamitkar S.D for his most valuable guidance without which this work would have been impossible. The authors would sincerely thank the Director of the School of computer Sciences S. R. T. M. University, Nanded, India for his support and encouragement. This work is a part of the dissertation work in Doctor of Philosophy degree (Ph.D), by the first author at SRTM University, India.

REFERENCES

[1] Rizzardi, A.Security in Internet of Things: networked smart objects. (Doctoral Thesis, Università degli Studi dell'Insubria, 2016).

[2] Roesch, M. (1999, November). Snort: Lightweight intrusion detection for networks. In Lisa (Vol. 99, No. 1, pp. 229-238).

[3] F. Ertam and M. Kaya, "Classification of firewall log files with multiclass support vector machine," 2018 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, 2018, pp.1-4. doi: 10.1109/ISDFS.2018.8355382.

[4] R. Hunt, "Internet/Intranet firewall security - Policy, architecture and transaction services," Comput. Commun., vol. 21, no. 13, pp. 1107–1123, 1998.

[5] Golnabi, K., Min, R. K., Khan, L., & Al-Shaer, E. (2006). Analysis of firewall policy rules using data mining techniques. In 10th IEEE/IFIP Network Operations and Management Symposium NOMS 2006 (Vol. 5, pp. 305–315). IEEE. doi:10.1109/NOMS.2006.1687561.

[6] Breier, J., & Branišová, J. (2017). A dynamic rule creation based anomaly detection method for identifying security breaches in log records. Wireless Personal Communications, 94(3), 497-511.

[7] Ucar, E., Ozhan, E.: The analysis of firewall policy through machine learning and data mining. Wirel. Pers. Commun. 96, 2891 (2017). https://doi.org/10.1007/s11277-017-4330-0.

[8] Al-Shaer, E. S., & Hamed, H. H. (2003, March). Firewall policy advisor for anomaly discovery and rule editing. In International Symposium on Integrated Network Management (pp. 17-30). Springer, Boston, MA.

[9] Al-Shaer, E., Hamed, H., Boutaba, R., & Hasan, M. (2005). Conflict classification and analysis of distributed firewall policies. IEEE journal on selected areas in communications, 23(10), 2069-2084.

[10] Snort. An open source network intrusion detection system. http://www.Snort.org/.

[11] Link to download TWIDS tool: http://twids.cute.edu.tw/en.

[12] As-Suhbani, H., Khamitkar, S.D. (2017): Enhancing snort IDS performance using TWIDS for collecting network logs dataset. Int. J. Res. Adv. Eng. Technol. 42–45 (2017). https://doi.org/10.22271/engineering.

[13] Link to download Weka: http://www.cs.waikato.ac.nz/ml/weka/

[14] Z. C. Lipton, C. Elkan, and B. Naryanaswamy, "Optimal thresholding of classifiers to maximize F1 measure," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014, vol. 8725 LNAI, no. PART 2, pp. 225–239.

Authors Profile

Miss. Hajar Esmaeil As-Suhbani, Research Scholar At School of Computational Sciences, S.R.T.M University Nanded, Maharashtra.

Dr. S. D. Khamitkar, Professor At School of Computational Sciences, S.R.T.M University Nanded, Maharashtra.