JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

Performance Improving of Intrusion Detection System using ML Techniques

Nallu Keerthi Reddy[1], Mulkur Lilly Grace Reddy[2]

Received: 16 March 2020 Revised and Accepted: 17 June 2020

Abstract: Machine learning (ML) is for the detection of network intrusion due to its predictive ability to use relevant data after training. ML is a good way to detect unknown and new attacks. Due to the lack of public data sets for networked IDSs, this may not be a perfect representative of existing genuine networks, but can still be used as an effective bottom line to help researchers compare different intrusion detection methods. The NSL data detection and database is used as a benchmark for the assessment of mechanisms for intrusion detection. Random Forests (RF) was found to be 97.76% reliable, whereas the least accurate algorithm was a standard random tree of 96.80 %.

KEYWORDS: NSL–KDD, random forest, IntrusionDetection System, J48

I. INTRODUCTION

These days an ever increasing demand for robust security in advancing technology. Despite the high demand for network security, current solutions continue to be insufficient for fully securing computer networks and internet applications in the form of cyberattacks such as DOS attacks and many others against the ever-increasing threats from hackers [1]. Significant work has been carried out on the intrusion detection system (IDS), however, this problem will never be outdated as the design and growth of the network infrastructure continues to change. The challenge is to meet developments and threats with equally advanced and effective solutions. Compared to the current threats, data sets and classifications of types of attacks are obsolete [2].

There is, therefore, an immediate need to upgrade the IDS approaches to address the current scenario and effective work is necessary to ensure that these methods are not outdated. Machine Learning (ML) methods learn from past data patterns and predict current data. Because ML recognizes trends, it can be used for a hybrid-based approach that can identify small variations in documented attacks instead of complex signatures. It is always important when new attacks are produced and that NIDS must adapt to changes in order to detect both known and unknown attacks.DDoS is a form of NIDS that attacks from various sources or hosts. This is achieved through the control of multiple compromised hosts that act as the attacker's zombie host. The key targets for such threats are companies providing Internet services, such as Google, Facebook, and Amazon. Service loss for websites of this type leads to financial losses [3].Counter-measures are taken according to the information obtained from the observed assaults. The more the type of attack will be identified, the more successful countermeasures will be chosen and the less they can affect the proper functioning of the device or network. Furthermore, if we do not detect the exact type of attack, countermeasures can have more severe consequences than in some cases the attack itself. That is why we aim to create an intrusion detection model that correctly categorizes each type of attack. ML types and literature algorithms have shown that supervised algorithms work well in IDS, whereas previously known attacks do not have a supervised algorithm [4].

The main contributions and organization of this paper are summarized as follows: In section 2 we describe background details of work and places by the authors. Section 3 discusses the proposed work. Section 4 deliberates results and discussions. Finally, in section 5, we concluded the paper.

II. BACKGROUND WORKS

Computer information and network details are important for companies, as compromise information can cause a great deal of damage. This is why the intrusion detection system is of great importance. A number of algorithms have recently been suggested to be added to the KDD99 dataset. This data set can contribute to accuracies up to 98.3%, but the database has some constraints, such as minimum test data. The data set NSL-KDD [5] has the following advantages compared with the initial KDD data set that it does not include repetitive train records so the classifiers are not geared toward records that are more regular. NSL-KDD data set is based on the following advantages. There are therefore no redundant documents in the suggested sample sets; thus, student's output is not influenced by approaches of better detection levels in the regular data. The number of records selected from each problem level group corresponds inversely to the percentage of records of the original KDD data set. As a consequence, the classification rates of distinct master learning methods vary in a broader range, making it easier to evaluate different learning techniques accurately. The number of train records and test sets is reasonable so that it is easy to perform the complete experiments without randomly selecting a small part. The results of the assessment of various research works will, therefore, be consistent and comparable.

In [6], the researchers proposed an alternative name for an Extreme learning computer feed-in neural network that can be used to solve problems of grouping, clustering, regression and usability technology. It was found that

this algorithm is a very precise algorithm if the data size is huge. SVM can provide better results if the data size is smaller.

In [7], the authors proposed a hybrid IDS architecture based on neural systems using two methods: firstly the multi-layer neural perceptrons network (MLP), and secondly the RBF. The hybrid simulation approaches of bagging classifiers are used to improve robustness, precision, and generalization. Besides, UNM Sending-Mail Data is used in this study based on a University of New Mexico immune system. In terms of accuracy, IDS' output for normal and abnormal traffic was 98.88% and 94.31% respectively slightly better than the individual classifiers comprising it.

In [8] the researchers deal with decision-making problems by using the Intelligent Swarm-Based Rough Set (IDS-RS) to pick the features and simplify the Weighted Local Search Swarm Optimization (SSO-WLS) information classification technique. The thesis provides a complete process solution to boost SSO rule mining work by measuring three predetermined constants. The experimental results from the 1999 KDD CUP dataset demonstrate the good overall efficiency with 93.3 percent precision in an average of 20 runs in the proposed hybrid network intrusion detection systems using an intelligence dynamic, swarm-based rough set.

In [9], the authors introduced the combination of misuse and anomaly detection into a hybrid framework based on two methods: the random algorithm of the forests and the K-mean algorithm of clustering. The random forest algorithm for the identification of misuse invasion and the k-means clustering algorithm for anomaly detection are used in this model. This system has poor design interpretability and output loss due to associated variables in random forests.

In [10] the authors integrate with a decomposition structure a misuse detection model and an anomaly detection model. This study uses the algorithm of the decision tree C4.5 and several SVM one-class versions. Experimental results on the NSLKDD dataset show that the proposed method of intrusion detection is in terms of detection performance, training time and time better than conventional methods.

III. SYSTEM MODEL

In general, with the same data set, the sub-data set of every class decreases, which results in a significant reduction in generalization capacity and an increase in classification errors, etc. The miss-classification is typical because the attacks are classified as normal behavior or other attacks in the same or dissimilar classification.

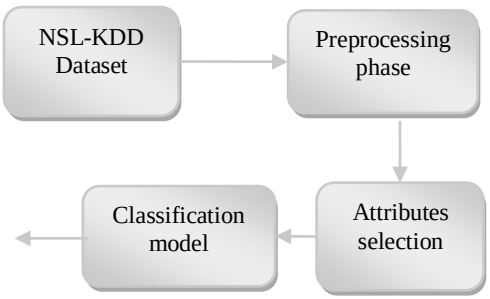

Fig.1.Block diagram of intrusion detection model

To reduce these classification errors, we propose the intrusion detection hierarchical model as illustrated in Fig.1 to increase the performance of the intrusion detection mechanism. The following are the main phases of our systems, each of which has its significance.

Dataset stage: Dataset selection is an important task because the system performance is based on the accuracy of a data set.

Pre-processing phase: Due to some of its symbolic characteristics, the classifier cannot handle the raw data array. Pre-processing is therefore necessary when non-numerical and symbolic elements are omitted and substituted because they do not signify an important role in the identification of invasion.

Selection of attributes: Also called variable selection, selection of attributes or variable subset selection is the process of selecting a subset of relevant features for the model's Classification stage: Data mining is the way to identify, analyze and simulate large volumes of data that can detect unknown movements or interactions that give the correct result.

Algorithm of the random forest: Random forest algorithm is a supervised algorithm for classification. It can somehow build a forest to render this random.

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

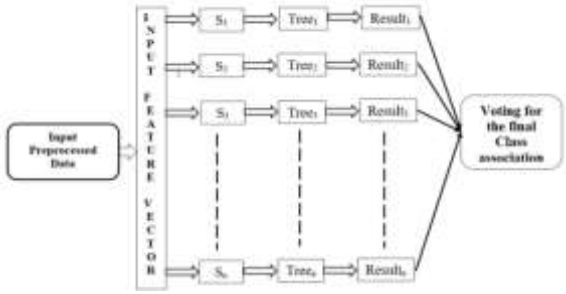

Fig.2 Framework the Random forest for intrusion detection model

The number of trees in this case has a direct relationship to the outcomes it can achieve: the higher the number of trees, the more precise the result.As stated, the Random Forest algorithm can be used for classification and regression tasks. Overfitting is a critical issue that may worsen the outcome, yet the classification system will not overfit the model for the Random Forest algorithm if there are sufficient trees in the forest. The third benefit is the identification of incomplete values from Random Forest, and the last benefit is that the classification of Random Forest can be based on group values. Fig.2 illustrates how the random forest classification method is applied in the proposed system's software classification. The random forest classification is fed a preprocessed sample of n samples. By using a variety of function subsets, RF generates n unique trees. Each tree provides a classification result, and the classification model results depend on the majority vote.

J48 Algorithm: Data must be separated into specific subclasses at each point depending on the judgment. J48 examines structured data that eventually leads to the division of information by selecting an item.

REP Tree: Classifier decreases error pruning Tree Classifier is a quick decision-making algorithm based on the principle of entropy computing the gain of information and reducing variance errors.

Random tree: This generates a collection of decision trees from a randomly selected learning group sub-set. The final class of the test object is then aggregated by votes from different decision trees.

IV. RESULTS AND DISCUSSION

The developed framework is tested by randomizing the generic NSL–KDD dataset in three parts: the full dataset, the quarter dataset, and the 1/4 dataset. The performance of each algorithm must be measured using performance metrics.

Mean absolute error: It is the average difference in all test cases between the predicted and the real value. It is a good measure for measuring performance.

Root Mean Square Error: This is used to scale differences between perceived values and the predicted values of the model. Taking the square root of the medium square error.

Table 1: Classifiers parameters for evaluation

Classifier name	MAE	RMSE	Accuracy
J48	0.0389	0.1552	97.12%
REPtree	0.0369	0.1471	97.24%
Random tree	0.0321	0.1763	96.80%

Classification Accuracy: It determines that any classifier will have an error rate and cannot correctly categorize it. The precision of classified instances divided by the Total number of instances multiplied by 100 is calculated as correctly classified instances.

In this study, four approaches are used. Any form of classification properly rates properly classified instances.

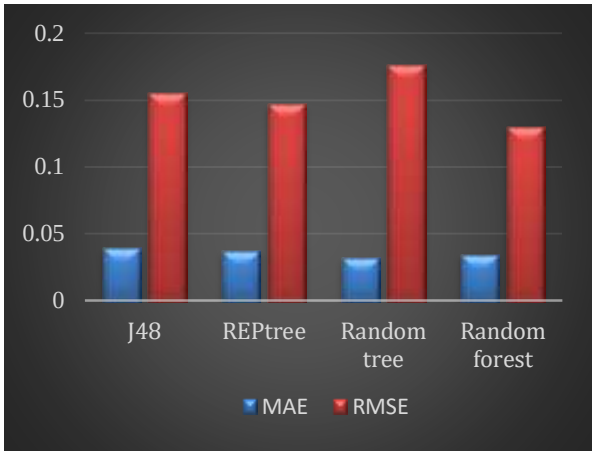

Fig. 3. Performance of the classifiers MAE, RMSE with cross validation of 10

Fig.3 illustrates that Random forest have least RMSE value as related to other methods due to less miss-classification in the instances.

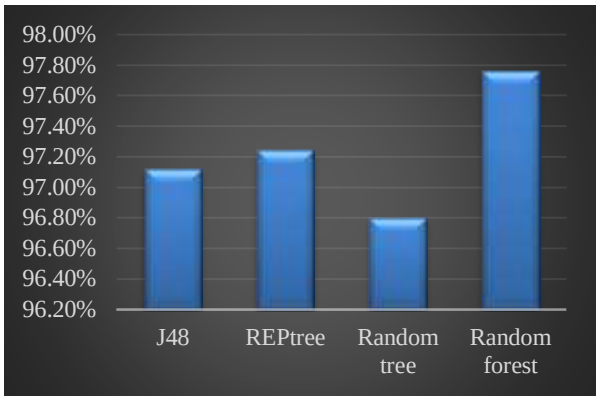

Fig. 4. Performance of the classifiers accuracy with cross validation of 10

Fig.4 illustrates that Random forest have highest accuracy of 97.76% as related to other methods due to high categorization ability of instances.

V. CONCLUSION

Intrusion detection and protection are important to present and future data networks and processes, although our daily routines depend heavily on them. Different machine learning methods have also been used but some techniques are better suited to the analysis of large data for networking and information systems intrusion detection. In this paper, they suggested a model of hierarchical intrusion detection, which was focused on the combination of J48, REP Tree, Random Tree Algorithm and Random Forest. Random Forest clearly outperforms other methods in accuracy, precision and recall the full data samples comprising 2^{16} records of normal and intrusive activities.

VI. REFERENCES

[1] Iftikhar Ahmad, Mohammad Basheri, Muhammad Javed Iqbal, Aneel Rahim, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection", IEEEAccess, 6, 33789-33795, 2018.

[2] Nisioti A, Mylonas A, Yoo PD, Member S, Katos V. From Intrusion Detection to Attacker Attribution : A Comprehensive Survey of Unsupervised Methods.IEEE Commun Surv Tutorials. 2018.

[3] Behal S, Kumar K. Detection of DDoS attacks and flash events using information theory metrics–An empirical investigation. Comput Commun [Internet].2017;103:18–28.

[4] Zamani M, Movahedi M. Machine Learning Techniques for Intrusion Detection: 1–11.

[5] Yan Zhang, Chong Di, Zhuoran Han, Yichen Li, Shenghong Li, "AnAdaptive Honey-pot Deployment Algorithm Based on Learning Automata ",in IEEE Second International Conference on Data Science in Cyberspace(DSC), 2017.

[6] Ahmad I, Basheri M, Iqbal MJ, Rahim A. Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection. 2018;33789–95.

[7] M. Govindarajan and R. Chandrasekaran, "Intrusion detection usingneural based hybrid classification methods," *Computer networks*, vol. 55,no. 8, pp. 1662–1671, 2011.

[8] Y. Y. Chung and N. Wahid, "A hybrid network intrusion detection systemusing simplified swarm optimization (SSO)," *Applied Soft Computing*,vol. 12, no. 9, pp. 3014–3022, 2012.

[9] R. M. Elbasiony, E. A. Sallam, T. E. Eltobely, and M. M. Fahmy, "Ahybrid network intrusion detection framework based on random forestsand weighted k-means," *Ain Shams Engineering Journal*, vol. 4, no. 4,pp. 753–762, 2013.

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

[10] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection methodintegrating anomaly detection with misuse detection," *Expert Systemswith Applications*, vol. 41, no. 4, pp. 1690–1700, 2014.