Ilkogretim Online - Elementary Education Online, 2020; Vol 19 (Issue 4): pp. 1054-1061
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2020.04.214

Analysis of Machine Learning based Techniques for Intrusion Detection System

Shriram V. Wasule, M.Tech Scholar, Department of Computer Science and Engineering, Shri Ramdeobaba college of engineering and management, Nagpur, Maharashtra, India.
Dr. Manoj B. Chandak, Professor, Department of Computer Science and Engineering, Shri Ramdeobaba college of engineering and management, Nagpur, Maharashtra, India.
Mr. Mohan Bihari, Senior Executive- Solar industries India Pvt Ltd, Nagpur, Maharashtra, India.

Abstract— One of the solutions used against malicious threats is the Intrusion Detection System (IDS). In addition, attackers still keep adjusting their instruments and tactics. It is still a difficult job to incorporate an agreed IDS scheme, however. Several studies have been carried out and tested in this paper to test different machine learning classifiers based on the KDD intrusion dataset. In order to test the chosen classifiers, multiple output parameters were successfully computed. In order to increase the intrusion detection system's detection rate, the emphasis was on false negative and false positive performance indicators. The tests carried out found that the decision table classifier obtained the lowest false negative score, while the highest average accuracy rating was achieved by the random forest classifier.

Index Terms: Intrusion detection, Machine Learning, Deep Learning, Network Based Attacks, Denial of Service Attack

I. INTRODUCTION

In the last two decades, information technology has been evolving rapidly. Computer networks are commonly used in the areas of human life and business. Therefore it is very important for IT managers to create trustworthy networks. In turn, rapid IT growth has provided many problems, which is a very challenging challenge, in developing stable networks. Many kinds of attacks endanger computer network usability, credibility and privacy. DOS is one of the most common hailing attacks. Denial of service attack is known to be one.

The purpose of DOS attacks is to deny many end user facilities temporarily. Generally, network services are overwhelmed and unnecessary demands overwhelm the system. This is why DOS serves as a big reference for all forms of threats directed at using computers and communications networks. [1] Yahoo was the first DOS victim in 2000 and DOS also reported the first public attack on the same day. Currently, DOS attacks are being targeted by online services and social websites [2]. In other ways, remote to local (R2L) attacks represent another umbrella for all types of attacks that are intended to be allowed by the local right since certain network services are only only open to local users for example the file server. Any forms of R2L attacks exist, for example. These forms of attacks are intended by SPY and PHF for the planning of unauthorized network access [3].

Regarding unauthorized access to network and device services, attacks against User to Root (U2R) seek to transform the root user's authorization of an attacker to root, and that user has complete rights to access to technology and device resources [4]. The biggest challenge is to hold attackers novel in tools and tactics to hack bugs of all sorts. Therefore all sorts of attacks on the basis of single fixed solutions are very difficult to detect. For this method of intrusion detection (IDS), network protection has become important. It is used to track network traffic so that warnings are created when attacks occur. IDS may be used to track network traffic on a single computer, or to monitor all network traffics, which is the most popular method used for intrusion network detection systems.

Two forms of IDS are commonly available (anomaly base or misuse base). Anomaly monitoring system for intruder detected on the basis of normal behaviour. This method of intrusion detection is used commonly because of their ability to identify new type of intrusions, by compared the actual real-time traffics with the previous reported normal real-time traffic. From a different point of view, however the most high values of false positive alert are registered, which means that several regular packets are assumed to be attack packets. Intrusion monitoring mechanisms for misuse are however deployed to detect threats based on the threat signature registry. The new form of attack (new signature) will move through it but at the same time it does not have a false warning.

With respect to literature [5], the identification of attacks is called a classification issue where the goal is to clarify whether the packet is either a regular packet or an attack packet. Therefore, based on essential machine learning algorithms, the architecture of an agreed intrusion detection system can be applied. The following machine learning algorithms were applied in this paper (J48, Random Forest, Random Tree, Decision Table, MLP, Naive Bayes, and Bayes Network) to test and correctly assess the intrusion detection method model based on an Information Discovery in Databases (KDD) bench market dataset that includes the following types of attacks (DOS, R2L, U2R, and PROBE).

The rest of this paper is structured in: Section (II) reveals how the KDD data set is incredibly useful in the application of machine learning algorithms. The details of the KDD reprocessed data set provided in section (III), and the chosen classifiers for the machine learning that are used in sectional experiments (IV). The first stage of studies on construction training models is discussed in the section (V). The assessment measures used to assess the efficiency of selected classifiers were defined in section (VI); the tests and the results obtained were also discussed. The article is concluded by Section (VII).

II. Relevant Works To The KDD Dataset

This section discusses the associated works for the application of machine learning algorithms using the KDD data collection. It offers a short description of the various machine study algorithms and illustrates how the DDC dataset is very helpful in determining and evaluating different types of machine study algorithms. In a detailed analysis of intrusion detection systems and KDD datasets, the classifier selection model suggested by [2] the author. 49596 instances of KDD data sets were collected for implementation of many algorithms of machine learning. Multi-layer perceptron and Naive Bayes. Authors also succeeded in suggesting two types of KDD dataset intrusions.

The authors used MATLAB tools to enforce the support vector machine (SVM) algorithm against network intrusions in [6]. To manipulate information, they was using the KDD dataset as a bench business dataset. They noticed that the SVM algorithm takes a long training period and that the usefulness of the SVM is also limited.

The Authors imported the KDD dataset and implemented the preprocess step, such as normalizing the region of attributes to [-1, 1], and translating symbols, according to a further analysis.

According to this study [7]. In two tests, Neural Network Feed Forward was applied. The authors concluded that the neural network is not ideal for R2L and U2R attacks, although the accuracy rating for DOS and PROBE attacks has been reasonable. The authors were able to implement the four following algorithms, namely Fuzzy ARTMAP, Radial Dependent Function, Back Propagation (BP) and Perceptron back propagation hybrid [8], in order to implement the neural network against KDD intrusion (PBH). The BP and PBH algorithms achieved the highest accuracy rate of the four algorithms evaluated and tested for intrusion detection.

From another angle, some of the researchers concentrate on attributes collection algorithms in order to reduce the cost of computing time. In [9] the authors concentrate on choosing the most important IDS attributes with the greatest precision and low time. For preparation and research, 10% of KDD was used. They used an expanded classifier and nerve network-based identification system to eliminate false positive warnings to the fullest degree possible. In the other hand, the knowledge gain algorithm was applied as one of the most powerful attribute choices. They also used multivariate technique to detect denial of service intrusions as a linear system method.

Furthermore to improve detection of multiple forms of intrusions, the genetic algorithm was applied. Meanwhile a technique to identify multiple forms of intrusions inside the KDD is suggested in [3]. The suggested technique aims to extract the optimal detection rate for intrusion forms, at the same time obtained the lowest false positive rate. The GA algorithm used to produce a set of efficient rules to detect intrusions. Based on this technique, they succeeded in recording 97 percent as a precision score. In some cases, if a single isolated machine learning algorithm is used to handle all types of intrusions, an unaccepted detection rate would be derived. In [11], the author used the Naive Bayes algorithm to classify all types of KDD intrusions. He showed that, based on a single machine learning algorithm, the detection rate was not acceptable.

Some researchers concentrate on a single form of attack, such as[12] the authors suggested a method for gathering new distributed denial of service data set that contains the following types of attacks (http flood, smurf, siddos and udp flood) following the proposed new DDOS dataset. Several machine learning algorithms were introduced to detect DOS intrusions, the MLP algorithm achieved a high accuracy rate of 98.36 per cent.

All previous research papers had a valued contribution and presented at the same time how the KDD dataset provides the appropriate environment for testing and evaluating different algorithms in machine learning. The previous works also demonstrate that the algorithm of single independent machine learning does not suggest the agreed detection rate. In this study based on the KDD dataset, the following machine learning classifiers (J48, Random Forest, Random Tree, Decision Table, MLP, Naive Bayes, and Bayes Network) were added, tested and evaluated. In order to test the preferred classifiers, the interest is in the most relevant output metrics, e.g. false negative and false positive. The emphasis would be on selecting the efficacy of the machine learning classifier, which met the agreed accuracy rate with the lowest false negative score, as a result of the applied tests.

III. Machine Learning Classifiers Overview

This section offers a brief description of the various algorithms for machine learning and explains the criteria for the application of machine learning algorithms in different fields, such as the detection of intrusions. The effect of continued technology advancement makes it more important for machine learning algorithms to evaluate and extract information from a large number of datasets generated. Overall, machine learning algorithms can be classified as algorithms that are supervised and algorithms that are not supervised [14]. Supervised algorithms learn from pre-labeled (classified objects in order to predict the object type. The unsupervised algorithm, however, considers the normal grouping of objects as unlabeled data provided. The following supervised learning algorithms are of interest in this analysis, since the imported KDD dataset contains the predefined groups.

Multi-layer Perceptron (MLP) Classifier: is among the most popular classifier functions that demonstrates its usefulness in dealing with many domain fields, such as time series, problems with classification and regression. [15] It is possible to incorporate the testing process within a limited amount of time. The training process, on the other hand, is usually conducted over a long period of time. It is possible to implement the MLP algorithm with different transfer functions, e.g. Sigmoid, Hyperbolic and Linear.

Relevant architecture considerations of the MLP algorithm implementations are the number of outputs or planned groups and the number of hidden layers. At the start, each node in the neural network has its random network weights, the significant weight numbers had the most efficient attributes in a dataset, and the significant weight values would have the most powerful parameters in a dataset.

Random Tree Classifier: It is one of the tree groups that should be defined before deployment by this classifying the number of trees. Every tree is a single tree of decision. Each tree has randomly chosen data set attributes. The random tree grouping may then be treated as a finite group of decision-making trees. The prediction protocol of the entire dataset is for multiple decision-making outputs to be migrated and the predicted winner class based on total votes [16] chosen.

Random Forest Classifier: It is one of the algorithms of the classification of trees, the main purpose of which is to improve the classification of trees based on the forest principle. The random forest classifiers manufactured by the research referred to [17] were approved for accuracy rates and can be used to manage data set noise values. During the classification stage, there is no re-change process. To apply this algorithm, it is important to define the number of trees in the forest, since each individual tree predicts the predicted performance and the voting method used to pick the desired outcome with the highest number of votes [17].

J48 Classifier: This classification is planned to enhance the implementation by Ross Quilan [18] of the 1993 C 4.5 algorithm. This classifier is the predicted performance as binary judgments, but with more stability between time and accuracy of computation [19]. The leaf node had a decision of anticipated production with respect to the tree layout of the decision.

Naive Bayes Classifier: The category of probabilistic classifiers corresponds to this classifier. It uses Baye's classification problem theorem. The first step of the classification method of Naive Bayes is to determine the total number of groups (outputs) and the likelihood of each data set class. Then the conditional likelihood for each attribute will be determined. The Naive Bayes basic format can be found in the above inquiry [8]. In addition, it can also be implemented in a short time with discrete and continuous features, unlike MLP classifier Naive Bayes [11]. In the meanwhile, the Naive Bayes network (BN) or the Confidence network may be portrayed. BN encourages the presentation of autonomous, knowing framework-based conditional probability. Generally BN is an acyclic chart between the expected output class and several attributes [20].

Decision Table Classifier: This classifier's principal idea is to create a search table to help identify the predicted output class. Several searching algorithms can be used for the efficiency of the decision table [21], for instance, breadth of first search, the genetic algorithm and cross-validation. The search table comprises a variety of conditions and the behaviour required apply to the predefined conditions. To put it another way, there are important rules to predict new inputs as a result of the decision table classifier [22]. The search table of the decision table may be used in other fields for example when the system is complex and lacks an expert basis for presenting the significant rules for the established a global system.

IV. Implementation Details

In general, the performance of machine learning algorithms depends on the nature of the data they are trained on. While standard training a deep neural networks can expanding the features involves manipulating numerous input parameters, the representation method, also called feature learning, is useful for explaining variations. Learning from real-world data, such as computer vision and natural language processing, should be straightforward, and interesting applications should become apparent very rapidly.

the most commonly used deep neural network models is the Multi-Scale Expand Since CNN doesn't need to look at information that has already been collected, it is best used as the raw data input into the network, and doesn't have many parameters. It has just a few factors that it must take in, which allows for fast incorporation. Since the work of CNNs has been developed, they have proven to be highly successful in image recognition. When it comes to those forms of network traffic, CNNs have stronger machine learning capabilities, so they are suitable for complete network expansion. However, this machine learning algorithm was found to be effective, which used a multilayer neural network (MLP), which effectively connected to the extraction layer, such as MLP, where it far outsider was used to create features that were far more accurate. CNN, on the other hand, is limited to looking at a single data set; it cannot do timing analysis on the way traffic moves. The word "single" here can be read to mean "on its own," or "just a single packet" in an attack." This packet can be dangerous since a large amounts of packets are being sent in fast succession or over a short time periods may overwhelm or interrupt other hosts. Under normal conditions, it's unknown whether or not the CNN would apply, which can lead to an alarm sounding way too often in this time of crisis.

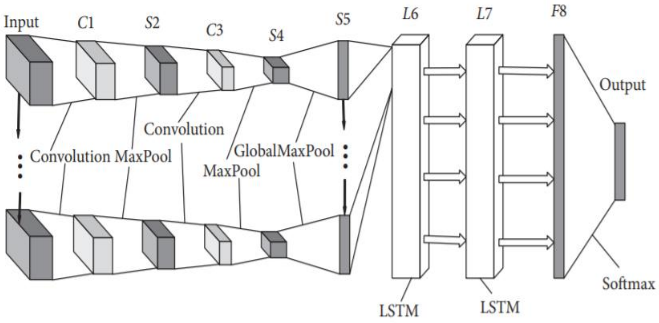

Figure 1. Architecture of DL based intrusion detection system.

This study makes the case for a DL-based intrusion detection system, DL-ITS, which blends the Convolutional Network long-N and Convolution Net (or Recurrent) Net in network and blends features to increase the efficiency and reduce prediction errors in detecting patterns in the long-term. During training, the DL-IDS applies long-memory with low-short term feature repetition neural network (LSTN) weights to aid with classification. This technique decreases the amount of unbalanced samples that can be present in a dataset and mitigates training and prediction errors by increasing robustness. Also, CNN is put to the test to categorize different types of networks traffic on the CICID2017 dataset to demonstrate if it works in more recent circumstances, and to show how well it works on other types of networks. Finally, we use CNN to identify networks with CIDDS (the brand new CID dataset) and attempt to show the accuracy in recent scenarios.

V. Result Analysis

Table-I recorded the highest value 0.999 based on ROC value by the Bayes network classifier, while the lowest value 0.953 was presented by the random tree classifier. In addition, based on the RMSE indicator, the random forest classifier had the lowest value of 0.0682, while the decision table presented the highest value of 0.0903. Testing and classifying 60000 instances of CICID2017 dataset records. The total number of incorrectly classified records is shown in Table-I for each selected classifier.

Table I. Average Accuracy Rate

Classification Algorithms	Instances Correctly Classified	Instances Incorrectly Classified	Accuracy (%)
CNN+LSTM	57654	2346	96.09
J48	55873	4127	93.12167
Random Forest	55628	4372	92.71333
Random Tree	53453	6547	89.08833
Decision Table	55646	4354	92.74333
MLP	55134	4866	91.89
Naive Bayes	54831	5169	91.385

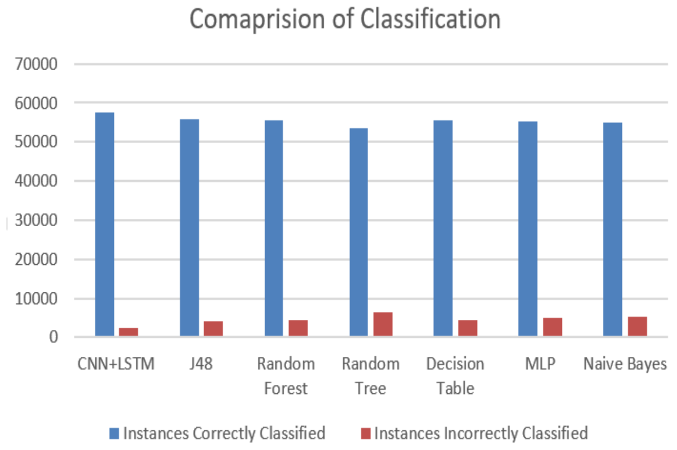

Figure 2. Comparison of Classification of Instance based on Correct and Incorrect

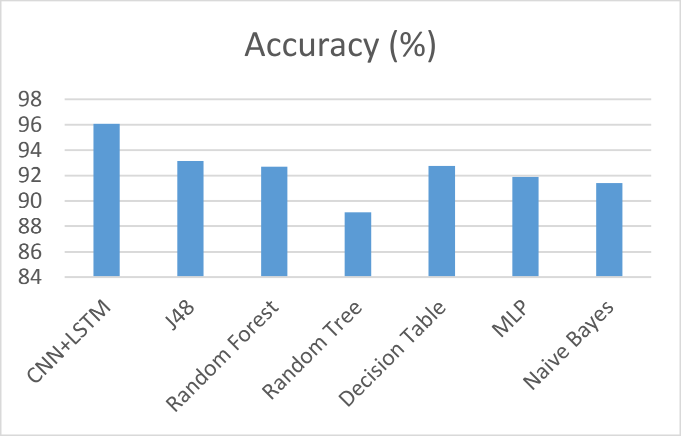

Figure 3. Accuracy of Classification

The results show that DL-IDS reached 96.09% in overall accuracy, and the accuracy of each attack type was above 98.50%, which achieved the best results in all models.

The time required for the construction of the classifier training models could be another problem. Random tree classifier built the training model in the fastest time based on the experiments, while MLP classifier built its model over 176 minutes; which is the longest time. The results of numerical examples can be summarized in the following paragraphs:

- The highest precision rate of 93.77 with the smallest RMSE value and false positive rate was achieved by the Random forest.
- The Random tree classifier, with the smallest ROC value, reached the lowest average precision rate of 90.73.
- There is no big difference between the MLP classifier and the Naive Bayes classifier with regard to the average precision rate.
- For the detection of normal packets, all machine learning classifiers present appropriate accuracy rates.
- The highest value for correctly detecting the normal packet was recorded by the Bayes network classifier.
- Based on FN parameters, there are no big distinctions between MLP and J48 classifiers.
- The decision table classifier has not achieved the highest accuracy rate, but it has the lowest FN rate and has a low time requirement for the training model to be built.
- In the accepted time period, all selected machine learning classifiers except MLP built their training models.
- It can be concluded that an acceptable accuracy rate with the lowest FN rate can be presented by the group of rules classifiers (the decision table), which also increases the confidentiality and availability of the network resources.

VI. Conclusion

Our proposed DL-based network traffic intrusion detection method, called DL-IDS, leveraged a Convolutional Neural Network (CNN) and an LSTM feature extraction technique to analyze the network traffic. Extract the features of a single spatio-temporal packet and then fuse the temporal ones to increase the efficiency of the overall system, in DL-IDS, on the other hand, and in turn, expand data stream temporal features in CNN. Additionally, the DL-IDS also utilizes category weights for optimization during the training process. This attack-type adjustment method reduced the number of unbalanced data points in the set, making it more accurate.

To see whether the proposed system would be viable, we ran the benchmark test on the CICIDS dataset, which is frequently used by researchers. Six data type files were selected to test DL-ability IDS's to identify various types of attackers, including well-known file-traffic data traffic and a command-traffic SSH traffic, all of these files and these files, a (patched)Dos-file-stealer, a (penetration test, dyedSFP), an obfuscation packer, and two classes of open ports. Also, we tested CNN models, the CNN-only model, and other machine learning models, and other models using LSTM data is fairly standard practice among ML practitioners. The results show that DL-IDS obtained a rate of 96.09% and were the best among all models.

REFERENCES

[1] G. C. Kessler, “Defenses against distributed denial of service attacks,” SANS Institute, vol. 2002, 2000.View publication stats

[2] H. A. Nguyen and D. Choi, “Application of data mining to network intrusion detection: classifier selection model,” in Asia-Pacific Network Operations and Management Symposium. Springer, 2008, pp. 399–408.

[3] S. Paliwal and R. Gupta, “Denial-of-service, probing & remote to user (r2l) attack detection using genetic algorithm,” International Journal of Computer Applications, vol. 60, no. 19, pp. 57–62, 2012.

[4] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the kdd cup 99 data set,” in Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on. IEEE, 2009, pp. 1–6.

[5] P. Amudha, S. Karthik, and S. Sivakumari, “Classification techniques for intrusion detection-an overview,” International Journal of Computer Applications, vol. 76, no. 16, 2013.

[6] M. K. Lahre, M. T. Dhar, D. Suresh, K. Kashyap, and P. Agrawal, “Analyze different approaches for ids using kdd 99 data set,” International Journal on Recent and Innovation Trends in Computing and Communication, vol. 1, no. 8, pp. 645–651, 2013.

[7] F. Haddadi, S. Khanchi, M. Shetabi, and V. Derhami, “Intrusion detection and attack classification using feed-forward neural network,” in Computer and Network Technology (ICCNT), 2010 Second International Conference on. IEEE, 2010, pp. 262–266.

[8] Z. Zhang, J. Li, C. Manikopoulos, J. Jorgenson, and J. Ucles, “Hide: a hierarchical network intrusion detection system using statistical preprocessing and neural network classification,” in Proc. IEEE Workshop on Information Assurance and Security, 2001, pp. 85–90.

[9] W. Alsharafat, “Applying artificial neural network and extended classifier system for network intrusion detection.” International Arab Journal of Information Technology (IAJIT), vol. 10, no. 3, 2013.

[10] N. Bhargava, G. Sharma, R. Bhargava, and M. Mathuria, “Decision tree analysis on j48 algorithm for data mining,” Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 6, 2013.

[11] C. Fleizach and S. Fukushima, “A naive bayes classifier on 1998 kdd cup,” 1998.

[12] M. Alkasassbeh, G. Al-Naymat, A. B. Hassanat, and M. Almseidin, “Detecting distributed denial of service attacks using data mining techniques,” International Journal of Advanced Computer Science & Applications, vol. 1, no. 7, pp. 436–445.

[13] S. D. Bay, “The uci kdd archive [http://kdd. ics. uci. edu]. irvine, ca: University of california,” Department of Information and Computer Science, vol. 404, p. 405, 1999.

[14] M. Al-Kasassbeh, “Network intrusion detection with wiener filter-based agent,” World Appl. Sci. J, vol. 13, no. 11, pp. 2372–2384, 2011.

[15] S. K. Pal and S. Mitra, “Multilayer perceptron, fuzzy sets, and classification,” IEEE Transactions on neural networks, vol. 3, no. 5, pp. 683–697, 1992.

[16] A. Cutler and G. Zhao, “Pert-perfect random tree ensembles,” Computing Science and Statistics, vol. 33, pp. 490–497, 2001.

[17] L. Breiman, “Random forests,” Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[18] J. R. Quinlan, C4. 5: programs for machine learning. Elsevier, 2014.

[19] M. S. Bhullar and A. Kaur, “Use of data mining in education sector,” in Proceedings of the World Congress on Engineering and Computer Science, vol. 1, 2012, pp. 24–26.

[20] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian network classifiers,” Machine learning, vol. 29, no. 2-3, pp. 131–163, 1997.

[21] R. Kohavi and D. Sommerfield, “Targeting business users with decision table classifiers.” in KDD, 1998, pp. 249–253.

[22] P. Aditi and G. Hitesh, “A new approach of intrusion detection system using clustering, classification and decision table,” 2013.

[23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The weka data mining software: an update,” ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10–18, 2009.

[24]