International Journal of Computer Engineering & Technology (IJCET)
Volume 9, Issue 5, September-October 2018, pp. 181–189, Article ID: IJCET_09_05_021
Available online at
http://www.iaeme.com/ijcet/issues.asp?JType=IJCET&VType=9&IType=5
Journal Impact Factor (2016): 9.3590(Calculated by GISI) www.jifactor.com
ISSN Print: 0976-6367 and ISSN Online: 0976–6375

PERFORMANCE EVOLUTION OF MACHINE LEARNING ALGORITHMS FOR NETWORK INTRUSION DETECTION SYSTEM

Kushal Jani
M.E Student, GTU Cyber Security, School of Engineering & Technology
Gandhinagar 382028, Gujarat, India

Punit Lalwani
Project Scientist, Bhaskaracharya Institute for Space Applications and Geo-Informatics,
Gandhinagar 382007, India

Deepak Upadhyay
Assistant Professor, GTU Cyber Security, School of Engineering & Technology
Gandhinagar 382028, Gujarat, India

Dr. M. B. Potdar
Project Director, Bhaskaracharya Institute for Space Applications and Geo-Informatics,
Gandhinagar 382007, India

ABSTRACT

Network Intrusion Detection System (NIDS) is one of the best solutions against network attacks. Attackers also dynamically change tools and technologies. However, implementing an associated accepted NIDS system is an additional challenge. This paper conducts and analyzes many experiments to evaluate numerous machine learning techniques that support the NSL-KDD intrusion data set. We have succeeded in identifying a number of performance metrics to judge the chosen technology. The main focus was on accuracy, precision and recall performance metrics to enhance the detection rate of network intrusion detection systems. Experimental results show that the deep learning approach achieves the highest accuracy and detection rate, while false negatives and false positives are rarely achieved.

Key words: Network Intrusion Detection System, Machine learning, Network Security, Performance Evolution.

Cite this Article: Kushal Jani, Punit Lalwani, Deepak Upadhyay, Dr. M.B. Potdar, Performance Evolution of Machine Learning Algorithms for Network Intrusion Detection System. *International Journal of Computer Engineering and Technology*, 9(5), 2018, pp. 181-189.
http://www.iaeme.com/IJCET/issues.asp?JType=IJCET&VType=9&IType=5

1. INTRODUCTION

The incredible growth of the internet over the past few years. The quantity of network traffic extent is also hastily increasing. Tracking network site visitors and network traffic for intrusion detection isn't always a brand new concept as there are many types of attacks besides the virus and malware. The network attacks can crash not only the host computer systems, but also the network performance considerably, or inside the worst case situation, it is able to completely prevent some network services [1]. Network intrusion detection is meant to detect and save you from network attack. Consequently, IDS needs to be very efficient and effective.

The early research painting on the anomaly detection was commonly signature-based totally. The issue with signature based approach is that database signature desires to be up to date as the new signatures come to be available and consequently it isn't always suitable for the real-time network intrusion detection. Therefore, more research on the network traffic anomaly detection by superior machine learning classification techniques are required to discover new form of anomalies. With the notable boom in net traffic, pleasant the requirement of the real time anomaly detection is a remarkable challenge [2].

Self-machine learning is going to famous in latest years, this is because of the exposure of many new computing technology as well as availability of greater statistics. Despite the fact that the machine learning strategies have been around for a long time, new trend is come up with a way to use them effectively and efficiently. As cited in advance, the usage of machine learning techniques to identify intrusion have been researched by many researcher [17]. However, to the great of our expertise, the industrial tools for intrusion detection do not have these strategies carried out; these days the prevailing strategies are signature based. Further, Researcher attempted many diverse techniques it is not always suitable for this application, Need analysis of techniques to finds best suitable technique. In addition, there should be common benchmark data on basis of that various techniques need to be evaluated.

On this paper, we use eight strategies on a NSL-KDD data set and examine the overall performance of those machine learning techniques in terms of accuracy, recall, and precision [2]. The eight techniques followed are: 1) Decision Tree, 2) Naive Bayes, 3) Support vector machines (SVMs), 4) Random Forest, 5) K-nearest neighbor, 6) Artificial Neural Network and 7) Deep Learning, comprising of the above mentioned algorithms. These algorithms are examined and evaluated based on their performance parameters [1][13][18].

The rest of the paper is organized as follows: Segment II of this paper describes Literature Review. Segment III discusses about Network intrusion detection system Segment IV explain basics about Machine Learning. Segment V discusses machine learning algorithms for Network Intrusion Detection. Segment VI discusses performance measures and experimental results. Conclusion and future work are drawn in segment VII.

2. LITERATURE REVIEW

A survey regarding network anomaly detection methodology is given in [15]. Different Machine Learning Techniques for Intrusion Detection and its Comparative Analysis is given in [3].Adaptive Hybrid method for Network Intrusion Detection and Comparison among Machine Learning Algorithms is mentioned in [14].Authors in [1] describe Evaluation of Machine Learning Techniques [17] for Network Intrusion Detection. Network threats detection system using fuzzy logic explained in [8].Authors in [10] have explained Classification of Attack Types for Intrusion Detection Systems using a Machine Learning Algorithm. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection explained in [6]. In [5] A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms was evaluated. Performance Evaluation

of Supervised Machine Learning Algorithms for Intrusion Detection is done in [7]. Network Intrusion Detection has different Classifiers, comparisons of them explain in this paper [2]. Deep Learning based Multi-channel intelligent attack detection for Data Security and A deep learning approach to network intrusion detection explained in [9],[13].

3. NETWORK INTRUSION DETECTION SYSTEM - NIDS

Network intrusion detection systems (NIDS) remain in active-mode all time for the system it monitored. They analyze and monitor the network packets coming into and leave the system. They are attempting to locate the anomalous behavior from the incoming packets, for example, if attacker scan ports of a machine, the NIDS can identify a large number of TCP-connection requests for distinct port scan within short time period. Then it could easily identify that a person is using port scans within the network. The intrusion detection system have to be smart enough to handle massive amount of packets (data) and differentiate amongst distinct patterns. The data may be continuous or discrete [12]. We are able to classify NIDS into two type: anomaly-based NIDS and misuse-based NIDS [14].Misuse based NIDS generally find for known intrusions pattern and anomaly based NIDS identify both known and unknown (new) patterns. Latest researches are widely concerned about anomaly based NIDS, our research concern with this area.

Anomaly-based systems study from the network traffic (packets) and prepare policies for classifying the activities as either normal or anomaly. This is contradict to signature-based systems that can detect attacks for which signature values have assigned previously. Before categories the network data as normal or anomalous, the intrusion detection system is required to learn about the normal behavior. This will be achieved in distinct ways, commonly using artificial intelligence (AI) and machine learning (ML) strategies. Different model of anomaly-based detection is strict anomaly detection, in which mathematical equations are used for identify deviation from normal network traffic. Previous researches display that NIDS has numerous flaws, particularly a high false positive rate and misleading true data packet as attack packet.

4. MACHINE LEARNING

Machine learning provides automated assessment techniques for large data sets. Work that is solved by machine learning techniques is generally divided into four types. These are characterized by the structure of the information analyzed through the corresponding learning algorithm [4].

4.1. Supervised Learning

A strategy that brings the link between input and output based on training practices in the way inputs are classified is the supervised learning methodology [16]. If the output data is categorized, the operation is called classification and the actual value of the output domain indicates a regression problem. Traditional examples of supervised learning activities include object reputation of pictures, machine translation, and unwanted mail filtering.

4.2. Unsupervised Learning

If a technique is given unlabeled input data, it is called unsupervised learning. In Unsupervised learning resolve problems with clustering elements, dimensional reduction of dimension subspace reduction data, and model pre-training, depending on the similarity measure. For example, you can perform clustering to find network anomaly detection [19].

4.3. Semi-Supervised Learning

Semi-supervised learning is a category of supervised learning activities and strategies that use unlabeled data for learning. It is usually unclassified data with a huge amount of unclassified data. Semi-supervised learning lies between un-supervised learning (without classified training data) and supervised learning (including absolutely classified training data) [17].

4.4. Reinforcement Learning

How to learn policies for a given behavior as part of a series of behavior, observation, and reward enhancement learning over the years [17]. Reinforcement learning can be viewed as a sub-discipline of machine learning related to creating plans and controls. It has recently been transformed into reinforcement learning by blending technology that has not been supervised and supervised so that PCs can beat human champions.

Machine learning allows evaluation of huge portions of data. At the same time as it usually offers quicker, more accurate results so one can pick out profitable opportunities or dangerous risks, it might also require extra time and sources to train it well. Combining machine learning with Artificial intelligent and cognitive technique will make it even extra effective in examine large volumes of data.

5. MACHINE LEARNING ALGORITHMS FOR NETWORK INTRUSION DETECTION

This segment describes the distinct machine learning techniques for cyber security. Every approach is described with detail, and references to formative works are furnished. Additionally, each technique, two to a three papers with their usages to cyber security domain are offered [6].

1) Decision Tree

Decision tree is similar like tree structure which has leaves, that describe branches and classifications, combination features prepare classification. An appropriate model is classified by checking attribute values against the decision tree's nodes. The most used techniques for dynamically constructing decision trees are J48 and C4.5 algorithms [10].

2) Naive Bayes

Naive Bayes is machine learning technique, it's based on Bayes' Theorem with an assumption of independence among predictors. In easy phrases, a Naive Bayes presume that the existence of a specific feature in class is independent to the existence of any other feature [1].

3) Support vector machines (SVMs)

The SVM is methodology in which locating a isolating hyperplane inside the feature space among two classes in the sort of way that the distance among the hyperplane and the closest data points of every class is maximized [7].

4) Random Forest

Random forests are primarily based on an easy idea: 'the knowledge of the group'. Combination of the output of more than one predictors offers a higher prediction than the fine individual predictor [7].To make a prediction, we simply achieve the predictions of all individuals' trees, then predict the class that receives the maximum votes. This method is referred to as Random forest.

5) K-nearest neighbour

k-nearest-neighbor is data classification technique that tries to decide what class a data point is by searching at the data points around it. This algorithm, look all point on data grid , and by

looking on data grid it try to figure out that a data point is in group X or Y , it check state of points that are nearby selected data point [11].

6) Artificial Neural Network

ANNs are innovated through the brain and made of connected artificial neurons able to do computations on given inputs [6]. The neurons of first layer are activated by input data, now output of first layer is consider as an input of second layer of neurons inside network. Same way each layer gives its output to next layer's input and output of last layer is final output called as result [6]. Layers among the input layer and output layer are define as hidden layer.

7) Deep Learning

Deep learning approach is enhanced version of neural network architectures, because of that deep learning referred as deep neural networks. Term deep used as number of hidden layer in neural networks. Deep learning provide facility of prepare model of complicated relationships and ideas using multiple levels of data representation [13].

6. PERFORMANCE MEASURES AND EXPERIMENTAL RESULTS

Seven machine learning algorithms are applied on classified network data. For every algorithm we prepared the confusion matrix(Table 1) which indicates the subsequent extensively used raw metrics [1].

True Positive (TP): Total number of network data flows predicted as anomaly and it is anomaly in reality.

False Positive (FP): Total number of network data flows predicted as anomaly but it is normal in reality.

True Negative (TN): Total number of network data flows predicted as normal and it is normal in reality.

False Negative (FN): Total number of network data flows predicted as normal but it is actually anomaly in reality.

Table 1 Confusion Matrix

		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

6.1. Description of Dataset

The NSL-KDD dataset is the newer redefined version of the KDD cup99 dataset. Many analysis had been achieved by way of many researchers on the NSL-KDD dataset using distinctive techniques and tools with an established objective to increase a powerful network intrusion detection system [5]. In depth evaluation on NSL-KDD data set by various ML strategies is performed within the WEKA (Waikato Environment for Knowledge Analysis) tool/WEKA API. Different clustering algorithm makes use of the NSL-KDD data set to train

and check various attacks including new definitions. Relative study at the NSL-KDD dataset with its previous version KDD99 cup data set is made through employing the Self corporation Map (SOM) ANN - Artificial Neural Network. An exhaustive analysis on various data sets like KDD99, GureKDD and NSLKDD are made using various data mining based machine learning algorithms like Depth analysis on numerous data sets like KDD99, GureKDD and NSLKDD are made in the use of various data mining based machine learning algorithms like, Decision Tree, Support Vector Machine (SVM), K-nearest neighbor, Fuzzy Logic, clustering algorithms, etc.

The dataset that we used for performance comparison of different machine learning algorithms has been made from NSLKDD dataset. [2]. NSL-KDD dataset is used for overall performance assessment in our paper. Network traffic is classified into two class normal and anomaly. Both training and testing data set are used with ARFF file extension. The training data set has of 42 attributes & 1166 instances at the same time as the testing data set have 42 attributes & 7456 instances [2]. NSL-KDD dataset have many advantages with the previous KDD data set, so we have selected NSL-KDD data set. List of attribute used for evaluation are mentioned in Table 2 [5].

Table 2 Attributes of NSL-KDD [5]

Sr. No.	Attribute Name	Description	Sr. No.	Attribute Name	Description
1	Duration	Length of time duration of the connection	2	Protocol_ type	Protocol used in the connection
3	Service	Destination network service used	4	Flag	Status of the connection
5	Src_ bytes	Number of data bytes Transferred from src	6	Dst_ bytes	Number of data bytes Transferred from dst
7	Land	if src and dst IP and port numbers are equal	8	Wrong_ fragment	Total number of wrong fragments in connection
9	Urgent	Number of urgent packets in this session.	10	Hot	Number of "hot" indicators in the content
11	Num_fail ed_logins	Count of failed login attempts	12	Logged_ in	Login Status: 1 if successfully logged in
13	Num_com promised	Number of compromised conditions	14	Root_shell	1 if root shell is obtained; 0 otherwise
15	Su_attemp ted	1 if ``su root'' command attempted or used	16	Num_root	Number of "root" accesses
17	Num_file_ creations	Number of file creation in this connection	18	Num_ shells	Number of shell prompts
19	Num_acce ss_files	Number of operations on access control files	20	Num_outb ound_cmds	Number of outbound command in ftp session
21	Is_hot_ login	1 if the login belongs to the "hot" list	22	Is_guest_ login	1 if the login is a "guest" login;

23	Count	Number of connections same destination	24	Srv_count	Number of connections to the same service
25	Serror_ rate	Percentage of activated the flag s0, s1, s2 or s3	26	Srv_ser ror_rate	Connections aggregated in srv_count (24)
27	Rerror_ rate	Activated flag (4) REJ, aggregated in count (23)	28	Srv_rer ror_rate	Activated flag (4) REJ, aggregated in count (24)
29	Same_ srv_rate	Percentage of service, aggregated in count (23)	30	Diff_srv_ rate	Percentage of service, aggregated in count (33)
31	Srv_diff_ host_rate	Different destination, aggregated in srv_count	32	Dst_host_ count	Number of connections having the same dst
33	Dst_host_ srv_count	Number of connections having the same port	34	Dst_host_sa mesrv_rate	percentage of conn agg-regated dst_host_count
35	Dst_host_ diff_srv_ rate	Different services, connections aggregated in dst_host_count (32)	36	Dst_host_ samesrc_ port_rate	Different services, connections aggregated in dst_host_srv_count
37	Dst_host_ srv_diff_ host_rate	Different destination, connections aggregated in dst_host_srv_count	38	Dst_host_ serror_rate	The percentage of connections aggregated in dst_host_count (32)
39	Dst_host_ srv_serror_ rate	The activated connection in dst_host_srv_count (33)	40	Dst_host_ rerror_rate	The percentage of connections activated the in dst_host_count
41	Dst_host_ srv_rerror_ rate	Percentage of conn. activated REJ, in dst_host_srv_count(33)	42	Class	The various classes Example (anomaly or normal)

6.2. Evaluation Metrics

Using evaluation indicator we can evaluate performance, which is important to compare performance among numerous techniques or different datasets [10]. Accuracy is main parameter to evaluate performance of Network intrusion detection system. In this research we consider three evolution parameter which is describe bellow [10].

Accuracy defines the percentage of the total quantity of accurate classifications.

$$Accu = TP + TN / TP + TN + FP + FN \quad (1)$$

Precision defines the total quantity of data appropriately predicted positive(anomaly) data over the number of instance predicted as positive(anomaly).

$$Preci = P / TP + FP \quad (2)$$

Recall defines the percentage of appropriately predicted positive(anomaly) data out of the number of actual positive(anomaly)data[10]

$$Recall=TP/FN+TP \quad (3)$$

6.3. Experimental Results

All the experimental analyses were executed in Java language. Result of the analyses in which precision, recall and accuracy are shown in Table 3. By looking results of machine learning algorithms it is quite satisfactorily with the dataset used for assessment. The reduce NSL-KDD data set is considered to evaluate the performance of different machine learning algorithms. In machine learning different classifier and methodology exists for single machine learning algorithm. Values mentioned in bracket indicated which classifier or method used for analysis. From the table and chart, we can notice that the precision is maximum for Decision tree(J48) ie 96.89% and minimum for Naive Bayes i.e. 85.31%. We can also observe that the

recall is highest for Deep learning ie 86.98% and lowest for Random forest i.e. 61.37%. In terms of accuracy, which is very important parameter among this three parameter, Deep learning is highest accuracy of 86.75%.ANN also gives very good result of accuracy ie 84.87%.According to the above analysis, The Best machine learning methodology mentioned from the results are ANN and Deep Learning.

Table 3 Precision, Recall and Accuracy Metrics

	Precision	**Recall**	**Accuracy**
K-nearest neighbor	88.32	64.02	76.47
Random Forest	96.66	61.37	78.43
Decision Tree (J48)	96.89	66.13	80.95
Naive Bayes	85.31	79.89	82.07
Support vector machines	95.03	70.89	82.63
Artificial Neural Network (FNN)	91.41	78.83	84.87
Deep Learning (DNN)	87.14	86.98	86.75

7. CONCLUSION AND FUTURE WORK

This Paper presented seven commonly using machine learning techniques namely K-nearest neighbor, Random Forest, Decision Tree, Naive Bayes, Support vector machines, Artificial Neural Network, Deep Learning for network intrusion detection system. We compared those techniques and obtained the preliminary results using the sample data from NSL-KDD data set. The Deep learning worked the best with accuracy value of 86.75%, whereas the Artificial Neural Network also came close with accuracy of 84.87%.Although Deep learning used in this study did not give the best result, the technique has capability and further work in this area is worth-pursuing. There are multiple techniques in deep learning, we can merge those techniques to achieve best results. Using threat intelligence we can easily identify new threats. To device efficient network intrusion detection system we can merge threat intelligence to machine learning technique.

ACKNOWLEDGMENTS

We are thankful to Shri T. P. Singh, Director, and BISAG for providing us infrastructure, motivation, and permitting to carry out this project at BISAG.

REFERENCES

[1] lmseidin, Mohammad, et al. "Evaluation of machine learning algorithms for intrusion detection system." *Intelligent Systems and Informatics (SISY), 2017 IEEE 15th International Symposium on*. IEEE, 2017.

[2] Choudhury, Sumouli, and AnirbanBhowal. "Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection." *Smart technologies and management for computing, communication, controls, energy and materials (ICSTM), 2015 International conference on*. IEEE, 2015.

[3] Hamid, Yasir, M. Sugumaran, and LudovicJournaux. "Machine learning techniques for intrusion detection: a comparative analysis." *Proceedings of the International Conference on Informatics and Analytics*. ACM, 2016.

[4] Haq, Nutan Farah, et al. "Application of machine learning approaches in intrusion detection system: a survey." *IJARAI-International Journal of Advanced Research in Artificial Intelligence* 4.3 (2015): 9-18.

[5] Dhanabal, L., and S. P. Shantharajah. "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms." *International Journal of Advanced Research in Computer and Communication Engineering* 4.6 (2015): 446-452.

[6] Buczak, Anna L., and ErhanGuven. "A survey of data mining and machine learning methods for cyber security intrusion detection." *IEEE Communications Surveys & Tutorials* 18.2 (2016): 1153-1176.

[7] Belavagi, Manjula C., and BalachandraMuniyal. "Performance evaluation of supervised machine learning algorithms for intrusion detection." *Procedia Computer Science* 89 (2016): 117-123.

[8] Shanmugavadivu, R., and N. Nagarajan. "Network intrusion detection system using fuzzy logic." *Indian Journal of Computer Science and Engineering (IJCSE)* 2.1 (2011): 101-111.

[9] Shone, Nathan, et al. "A deep learning approach to network intrusion detection." *IEEE Transactions on Emerging Topics in Computational Intelligence* 2.1 (2018): 41-50.

[10] Park, Kinam, Youngrok Song, and Yun-Gyung Cheong. "Classification of Attack Types for Intrusion Detection Systems Using a Machine Learning Algorithm." *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 2018.

[11] Papernot, Nicolas, et al. "Towards the science of security and privacy in machine learning." *arXiv preprint arXiv:1611.03814*(2016).

[12] Lee, Chie-Hong, et al. "Machine learning based network intrusion detection." *Computational Intelligence and Applications (ICCIA), 2017 2nd IEEE International Conference on*. IEEE, 2017.

[13] Jiang, Feng, et al. "Deep Learning based Multi-channel intelligent attack detection for Data Security." *IEEE Transactions on Sustainable Computing* (2018).

[14] Haque, MdEnamul, and Talal M. Alkharobi. "Adaptive hybrid model for network intrusion detection and comparison among machine learning algorithms." *International Journal of Machine Learning and Computing* 5.1 (2015): 17.

[15] Ahmed, Mohiuddin, AbdunNaser Mahmood, and Jiankun Hu. "A survey of network anomaly detection techniques." *Journal of Network and Computer Applications* 60 (2016): 19-31.

[16] Liu, Qiang, et al. "A survey on security threats and defensive techniques of machine learning: a data driven view." *IEEE access* 6 (2018): 12103-12117.

[17] https://en.wikipedia.org/wiki/Machine_learning

[18] https://en.wikipedia.org/wiki/Random_forest

[19] https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

[20] http://mysoftheaven.com/service/machine-learning