Comparison of Machine Learning Algorithms Performance in Detecting Network Intrusion

Kamarularifin Abd Jalil[1],
Muhammad Hilmi Kamarudin[2]
Faculty of Computer & Mathematical Sciences,
Universiti Teknologi MARA,
Shah Alam, Malaysia
kamarul@tmsk.uitm.edu.my 1
hilmi_kamarudin@yahoo.com[2]

Mohamad Noorman Masrek
Faculty of Information Management,
Universiti Teknologi MARA,
Shah Alam, Malaysia
mnoorman@salam.uitm.edu.my

***Abstract*— Organization has come to realize that network security technology has become very important in protecting its information. With tremendous growth of internet, attack cases are increasing each day along with the modern attack method. One of the solutions to this problem is by using Intrusion Detection System (IDS). Machine Learning is one of the methods used in the IDS. In recent years, Machine Learning Intrusion Detection system has been giving high accuracy and good detection on novel attacks. In this paper the performance of a Machine Learning algorithm called Decision Tree (J48) is evaluated and compared with two other Machine Learning algorithms namely Neural Network and Support Vector Machines which has been conducted by A. Osareh [1] for detecting intrusion. The algorithms were tested based on accuracy, detection rate, false alarm rate and accuracy of four categories of attacks. From the experiments conducted, it was found that the Decision tree (J48) algorithm outperformed the other two algorithms.**

Keywords-Machine Learning; Neural Network; Support Vector Machines; Decision Tree; KDD 99

I. INTRODUCTION

The importance of securing information in an organization has been very immense lately. Any organization would not want any of its information to be known by other organizations especially its competitor. In order to protect their information, organizations are willing to spent millions of dollars, showing how important the issue is. One of the solutions to securing the information is by using the Intrusion Detection System (IDS). Simply put, IDS which can be in the form of an application or device, just like firewall, would detect malicious or suspicious activities in the network. IDS was first introduced in 1980 by Anderson [2] and then improved by Denning [3] in 1987.

Basically, there are two Intrusion Detection techniques, i.e. Anomaly Detection and Misuse Detection [4]. Anomaly Detection [4] is basically based on assumption that attacker behavior is different from normal user's behavior. The strategy is to look for unusual or abnormal activities in a system or network. When the detection is performed, the normal behavior data will be compared to actual user's data. If the offset is below threshold value, then user's behavior can be considered as normal without any intention of attack. One of the advantages of this detection is that it has high detection rate and able to detect novel attack.

On the other hand, Misuse Detection [4] (also known as signature-based detection), uses pattern matching. In order to determine an attack, it will compare the data with the signature in signature database, and if the data match with the pattern as in signature database, then it will define as attack. This type of detection has high detection rate with low false alarm. In this research, the first technique i.e. the Anomaly Detection is used. The Anomaly Detection can be implemented using different other techniques such as Statistical Model, Computer Immunological Approach and Machine Learning. And to complicate things further, Machine Learning technique for example can employ a number of other algorithms. In this research, we will only be looking at three Machine Learning algorithms namely; Neural Network, Support Vector Machines and Decision Tree.

A. Neural Network

Neural Network is a mathematical model or computational model that simulates the structure functional aspect of biological Neural Network. It processes information using a connectionist approach and consist of artificial neurons. Modern Neural Network are non-linear statistical modeling tools, which is usually used for modelling complex relation ship between input and output in finding pattern in data [5]. The technique of Neural Networks follows the same theories of how the human brains works. In human brain, there is a large collection of interconnected neurons that connect both sensory and motor nerves. According to what most scientists believed neuron in the brains work by emitting fire an electrical impulse across the synapse to other neurons.

B. Support Vector Machines

In classification and regression, Support Vector Machines are the most common and popular method for machine learning tasks [6]. In this method, a set of training examples is given with each example is marked belonging into one of two categories. Then, by using the Support Vector Machines algorithm, a model that can predict

978-1-4244-7578-0/$26.00 © 2010 IEEE

whether a new example falls into one categories or other is built [7].

C. Decision Tree

Decision Tree algorithm is normally used for classification problem. In this algorithm, the data set is learnt and modeled. Therefore, whenever a new data item is evaluated, it will be classified accordingly. Decision Tree algorithm can also be used for Intrusion Detection. For this reason, the algorithm will also learn and models data based on the training data. As a result, the model can classify which attack types does a future data belongs to based on the model built. One of the strength of Decision Tree is it can works well with huge data sets. This is important as large amount of data flow can be found across the computer networks. In real-time Intrusion Detection, it works well because Decision Tree gives the highest detection performance and can construct and interprete model easily. Another useful property of Decision Tree in Intrusion Detection model is its generalization accuracy. This is due to the trends in the future where there will always be some new attacks, and by having generalized accuracy provided by Decision Tree, these attacks can be detected.

II. RELATED WORKS

In the Intrusion Detection field of studies, there are several approaches in solving Intrusion Detection problem. In his paper entitled "Identifying False Alarm for Network Intrusion Detection System using HYBRID Data Mining and Decision Tree", Anuar [8] has proposed a hybrid statistical method using data mining and decision tree. The author believes that by using this approach, the misclassification of false positives and to distinguish between attacks and false positives can be reduced.

Shamsuddin [9] in his paper entitled "Applying Knowledge Discovery in Database Techniques in Modeling Packet Header Anomaly Intrusion Detection Systems" introduced the model called Protocol based Packet Header Anomaly Detector (PbPHAD). The model was designed to detect anomalous behavior of network traffic packet based on three specific network and transport layer protocols namely; UDP, TCP and ICMP. He said in his paper, one of the keys of having a good classification results is to have secondary attributes intelligently, which would greatly assists the classifier algorithm to produce the good production rules in IDS model.

In his paper entitled "Anomaly Intrusion Detection System Using Information Theory, K-NN and KMC Algorithms", Shirazi [10] has proposed an anomaly detection engine based on K-NN, K-Means Clustering (KMC) algorithm. Shirazi has proved his finding by enhancing the efficiency of detecting dangerous attack like R2L and U2R.

Makkithaya, Reddy and Acharya in their paper entitled "Improved C-Fuzzy Decision Tree for Intrusion Detection" [11], studied and proposed a fragmentation based on c-fuzzy decision tree model for Intrusion Detection system by focusing on improving the performance of detection. The authors used c-fuzzy decision tree to select the best feature of the data set, which would built an effective IDS.

Lee [12] in his paper entitled "A Data Mining Framework for Building Intrusion Detection Models" has proposed using data mining framewrok to build an Intrusion Detection model. The author has suggested that the learning rules that accurately capture the behavior of intrusions and normal activities can be used for misuse detection and anomaly detection.

Barbara, Jajodia and Wu [13] in their paper entitled "ADAM" (A testbed for exploring the use of data mining in Intrusion Detection) discussed the uses of data mining approach in Intrusion Detection. This data mining technique works by learning training data know to be free of attacks (normal) and then uses an algorithm group an attack from the data. It also uses associates rules to store knowledge data about the nature of pattern about individual records that can improve the classification efficiency. In ADAM [13], association rules and classification algorithm has been combined to discover attacks in audit data.

III. EXPERIMENT

In this research, we intend to compare the efficiency of Neural Networks, Support Vector Machines and Decision Ttree (J48) algorithms against KDD-cup dataset. The dataset is very large, and it is almost impossible to test the data using ordinary computer. The large dataset requires high performance machines. Thus in this research, we need to resemble the data set into smaller dataset, and run it using an ordinary computer. The process of testing the data started with the selection of 10,000 data from the training and testing dataset with the normal data proportion of 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90% while the remaining data proportion are left for the attack data which were evenly distributed and sampled. The two datasets were then resembleed into one dataset.

What the proportion means is that, for example, when the normal proportion of the data is 10%, the remaining data (which are the attack data) will be 90%. These remaining data will consists of 39 attacks types which are distributed evenly. The behaviour of attacks and normal data can be described as below:

- True Positive (TP): when amount of attack data detected is actually attack data.
- True Negative (TN): when amount of normal data detected is actually normal data.
- False Positive (FP): when normal data is detected as attack data.
- False Negative (FN): when attack data is detected as normal data.

The main objective of an Intrusion Detection System is to have high accuracy and detection rate with low false alarm. For this reason, in this paper, we will compare the accuracy, detection rate, false alarm rate and accuracy result of 4 different attacks categories using the Decision Tree (J48) algorithm with two other algorithms i.e. Neural Network and Support Vector Machines which has been conducted by A. Osareh [1].

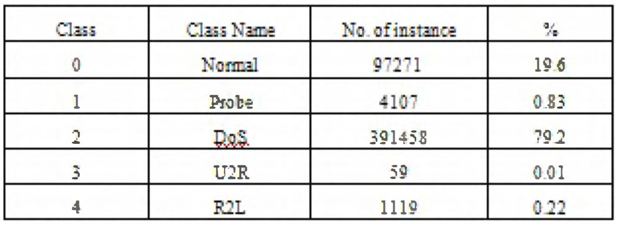
In order to compare the performance of these three machine learning algorithms, an experiment was carried out. In the experiment, the KDD 99 dataset was used (for training and testing data set). The dataset from KDD 99 was pre-processed first by selecting different percentage of normal data from both training and testing data (refer to Table I and Table II). Then the remaining data i.e. the attack data were organized evenly, and after that they are sampled into one data set. During the testing, we used the Decision Tree C 4.5 algorithm, which is known as J48 in WEKA. After the evaluation was completed, the results are then compared with the other two algorithms i.e. Neural Network and Support Vector Machine.

One of the objectives of this research is to have a fair comparison for detecting the different categories of attacks. In order to achieve this, 10,000 data for each attack category was used instead of using the whole KDD 99 dataset.

In this research, we used Weka 3.7.1. Weka requires dataset with arff format. After installing Weka 3.7.1, we also need to change the option in JRE (Java Runtime Environment) in Weka with –Xmx1600m. This was done to allocate 1600Mb of Ram memory. By doing this, we can increase the available memory of machine learning in Weka. Besides the option above, the Weka parameter of decision tree J48 algorithm such as reducedErrorPrunng, confidenceFactor and minNumObj were run using the default values. During the test, we used the 10 folds cross-validation as our test option. Cross-validation is a technique of analyzing independent data set and is used to predict accurately the predictive model. Generally, cross validation works by partitioning a data into complementary subsets and performing analysis on the subset. In our research, we perform 10 folds cross validation in order to obtain the optimum result from the dataset.

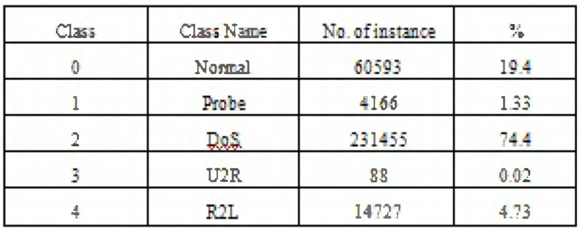
With 10 fold cross-validation, the data was randomly partitioned into 10 subsamples. From the 10 subsamples, one subsample was retained for validation and testing; while the remaining 9 subsamples were used for training data. Then, the cross validation process were repeated 10 times (folds), with each of the 10 subsamples are used once as the validation data. The 10 results from the folds are then averaged to produce a single estimation. The advantage of using this technique is, it will repeat the random sub-sampling i.e. all observations were used for both training and testing and each observation was used exactly once for validation.

TABLE I. TESTING DATA



Class	Class Name	No. of instance	%
0	Normal	97271	19.6
1	Probe	4107	0.83
2	DoS	391458	79.2
3	U2R	59	0.01
4	R2L	1119	0.22

TABLE II. TRAINING DATA



Class	Class Name	No. of instance	%
0	Normal	60593	19.4
1	Probe	4166	1.33
2	DoS	231455	74.4
3	U2R	88	0.02
4	R2L	14727	4.73

IV. RESULTS

Figure 1 shows the results obtained from the experiment using the original class label classification with the percentage of normal data of 10%-90%. From Figure 1, we can see that when there is an increase in the percentage of normal data the accuracy for these three algorithms also increases. This is because when the percentage of normal data is low, the percentage of attack data will be high. For this reason, all three algorithms had difficulty to classify all the 39 attacks when the percentage of normal data is low. In the experiment, when the percentage of normal data rised from 10% to 90%, the percentage of attack data fell from 90% to 10%. As a result, the number of data that were correctly classified also rises. Overall, the Decision Tree algorithm has outperformed both Neural Network and Support Vector Machines algorithms with the average accuracy of 99%.

In Figure 2, it shows that the detection rate result for the percentage of normal data from low to high. As shown in Figure 2, the result of detection rate for these three algorithms decline when the percentage of normal data increases. This happened because, when the percentage of normal data is higher i.e. 10%-90%, the proportion of attacks will be getting lower i.e. 90%-10%. As a result, the learning for attacks data will be more difficult as the amount of attacks data to be learnt are insufficient.

From Figure 3, we can see that, when the percentage of normal data increases, the result of false alarm rate (FAR) decreases. These scenarios happen because of the high percentage of normal data which helped the learning process of the algorithms in order to understand a normal behavior. As a result, it is easy for them to recognize a normal data and as a result the false alarm rate (FAR) decreases. Based on Figure 3, the result of false alarm rate (FAR) between the Neural Network and the Decision Tree algorithms are closed to each other. The Decision Tree algorithm is slightly lower (1.55%) than the Neural Network algorithm (1.62%). A good IDS should have a low false alarm rate (FAR) and between these three algorithms; Support Vector Machines (0.92%) produced a very low false alarm rate (FAR).

Given below are the comparisons of the three machine Learning Algorithms performance for each of the attack types.

Probe: From Table III, we can see that, the Decision Tree (J48) algorithm shows the highest result of accuracy when the percentage of normal data is at 10%. In addition, the Decision Tree (J48) algorithm also shows consistency by giving the average of (98.7%), outperformed the Neural Network and the Support Vector Machine algorithms. Overall these three algorithms still have a good detection on Probe attack.

DoS: From Table III, the Neural Network algorithm and the Support Vector Machine algorithm showed a bit low of accuracy with the average of (58.6% and 62.5% respectively), while the Decision Tree (J48) algorithm has the accuracy of 99.7%. This shows that the Neural Network algorithm and the Support Vector Machine algorithm had difficulty in detecting DoS attack, while the Decision Tree (J48) algorithm shows a good detection on this category of attack.

U2R: From Table III, the Neural Network and the Support Vector Machine algorithms shared almost the same level of accuracy with 65.4% and 65.5% respectively. It shows that both the Neural Network and the Support Vector Machine algorithms having the same capability of detecting U2R attack. On the other hand, the Decision Tree (J48) algorithm still shows a good result of detection with 99.6% for this class of attack.

R2L: From Table III, the Neural Network and the Support Vector Machine algorithms once again shared almost the same average of accuracy with 14.6% and 14.7% respectively. For both Neural Network and Support Vector Machine algorithms, R2L attack is their poorest detection performance compared to oher detection. The same goes with the Decision Tree (J48) algorithm. It shows, these three algorithms having difficulty on detecting this kind of attack. But for this category of attack, the Decision Tree (J48) algorithm still outperformed both Neural Network and Support Vector Machines algorithms with the average of detection is 96.2%, which is far good than the Neural Network and Support Vector Machines algorithms.

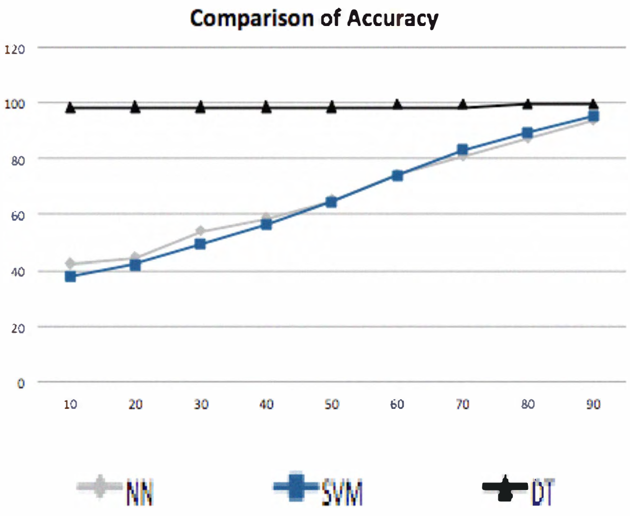


Figure 1. Acuracy result using KDD 99 dataset

Table III shows the detection rate for four different attacks consisting of (Probe, DoS, U2R and R2L) which are the common category of attack, with Decision Tree (J48) outperform both Neural Network and Support Vector Machines for all four different categories of attacks. Finally, according Table 7, the average result for these three different algorithms were compared in order to see which algorithm gives high detection rate. As summarized, from these four categories of attack (Probe, DoS, U2R, and R2L), Decision Tree (J48) has shown excellent results that outperform Neural Network and Support Vector Machines. Both Neural Network and Support Vector Machines shows good result on detecting category attack Probe and Dos but having difficulty on detecting category attack U2r and R2L. For Decision Tree, (J48) it shows a good detection for Probe (98.7%), DoS (99.7) and U2R (99.6%) but slightly poor detection for detecting category attack R2L (96.2%).

V. CONCLUSION

In this paper, three Machine Learning algorithms namely Neural Network (NN), Support Vector Machine (SVM) and Decision Tree (DT) have been compared in terms of accuracy, detection rate, false alarm rate and accuracy for four categories of attack under different percentage of normal data. The purpose of this research is to determine the best algorithm that can be used as a benchmark for research in Intrusion Detection by using KDD 99 dataset. KDD 99 dataset was the current benchmark dataset in Intrusion Detection. However, the dataset was distributed unevenly and might produce an error if only one set of dataset is used. Therefore in this research, the dataset was distributed evenly and the datasets from the training and testing were combined. The main reason we combine the dataset is to make sure that all 39 attacks in both datasets can be run simultaneously with different percentage of normal data in order to get an average value.

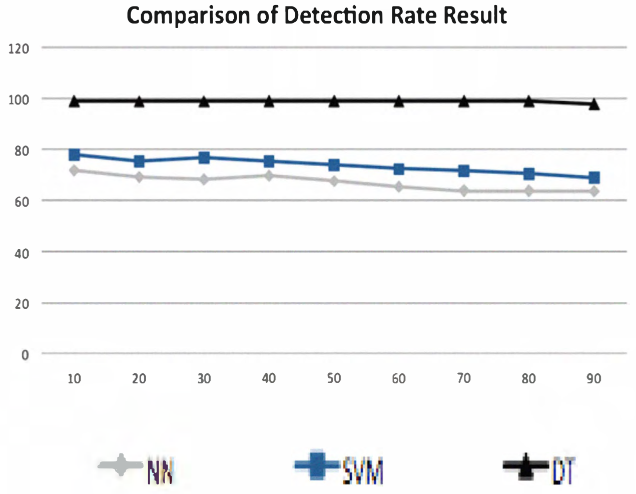


Figure 2. Detection rate result using KDD 99 dataset

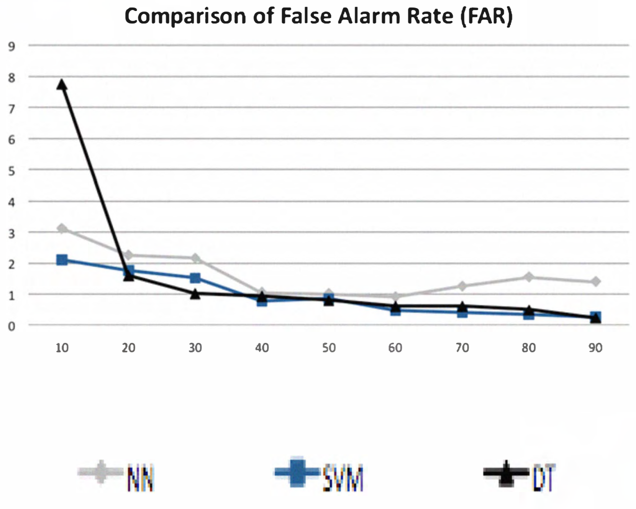


Figure 3. False alarm rate (FAR) result using KDD 99 dataset

The findings of this study suggested that these three machine-learning algorithms would misclassify an undesirably large amount of data when the number of attack is higher than the number of normal data. From the experiment conducted, we found that, in accuracy and detection rate, the Decision Tree (J48) was superior from others. However in determine the false alarm rate; Support Vector Machines has outperformed the others. On the other hand, the comparison result of accuracy for four categories of attack shows that, the Decision Tree algorithm performed much better than the other two algorithms.

REFERENCES

[1] A. Osareh, B.Shadgar, "Intrusion Detection in Computer Network based on Machine Learning Algorithm", *IJCSNS International Journal of Computer Science and Network Security*, Vol.8 No.11, November 2008.

[2] J. P. Anderson (1980), "Computer Security Threat Monitoring and Surveillance, technical report, February 26, 1980

[3] D. Denning, "An Intrusion Detection Model, *IEEE Transaction on Software Engineering*, 13(2):222-232, 1987

[4] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, "Machine Learning: An Artificial Intel ligence Approach", Tioga Publishing Company, 1983.

[5] D. Rumelhart, G. Hinton and R Williams, "Learning internal representations by back-propagating errors," Parallel Distributed Processing: Explorations in the Microstructure of Cognition, D. Rumelhart and J. McClelland editors, vol. 1, pp. 318-362, MIT Press, 1986.

[6] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verleg, 1995.

[7] J. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, pp. 121-167, 1998.

[8] N. B. Anuar, H. Sallehudin, A. Gani, O. Zakari, "Identifying False Alarm for Network Intrusion Detection System using HYBRID Data Mining and Decision Tree", *Malaysian Journal of Computer Science*, Vol. 21(2), 2008, Faculty of Computer Science and Information Technology University of Malaya, Kuala Lumpur, Malaysia

[9] S. Shamsuddin and M. E. Woodward, "Applying Knowledge Discovery in Database Techniques in Modeling Packet Header Anomaly Intrusion Detection Systems, School of Informatics", University of Bradford, Bradford BD7 1DP, United Kingdom, 2008

[10] H. M. Shirazi, "Anomaly Intrusion Detection System Using Information Theory, K-NN and KMC Algorithms", Australian Journal of basic and Applied Sciences vol. 3(3), pp. 2581- 2597, INSInet Publication, 2009.

[11] Makkithaya, K., Subba Reddy, N.V., Dinesh Acharya, U."Improved C-Fuzzy Decision Tree for Intrusion Detection", World Academy of Science, Engineering and Technology, vol. 32, 2008.

[12] W. Lee, S. Stolfo, and K. Mok, "A Data Mining Framework for Building Intrusion Detection Models", Proc. Of the 1999 IEEE Symposium on Security and Privacy, Oakland, California, May 1999.

[13] Barbara, D., Couto, J., Jajodia, S., Popyack, L., Wu, N.: ADAM: Detecting intrusions by data mining. In: Proc. of the IEEE Workshop on Information Assurance and Security (June 2001)

TABLE III. DETECTION RATE AVERAGE RESULTS FOR FOUR DIFFERENT ATTACKS FOR NN, SVM AND DT

Percentage of normal data/ Percentage of attack data	Probe			Dos			U2R			R2L		
	NN	SV M	DT	NN	SVM	DT	NN	SVM	DT	NN	SVM	DT
10/90	75.0	70.3	99.5	55.3	60.8	99.9	58.3	62.4	100	8.3	14.5	96.7
20/80	81.2	72.5	99.1	56.0	61.2	99.8	60.1	66.0	100	16.8	12.3	94.3
30/70	83.4	78.9	99.3	56.4	61.9	99.9	67.2	70.9	100	17.4	12.9	96.4
40/60	80.1	78.5	99.5	54.2	56.7	99.8	72.0	74.1	100	10.6	16.8	96.9
50/50	85.8	90.1	99.3	53.1	58.0	99.8	70.3	72.2	100	18.0	19.2	97.0
60/40	82.0	86.4	98.7	60.7	65.1	99.8	59.2	64.1	99.9	20.1	17.6	96.8
70/30	84.0	89.0	98.8	66.5	66.2	99.8	65.7	61.3	99.9	11.3	13.7	97.2
80/20	85.3	91.2	97.7	64.0	67.4	99.5	67.4	58.9	99.8	12.5	10.0	97.1
90/10	85.7	92.5	96.4	61.2	65.5	98.7	69.1	60.0	96.6	16.8	15.9	93.6
Average	82.5	83.2	98.7	58.6	62.5	99.7	65.4	65.5	99.6	14.6	14.7	96.2