RESEARCH ARTICLE | JULY 21 2016

Performance comparison of intrusion detection system based anomaly detection using artificial neural network and support vector machine FREE

Aditya Nur Cahyo; Risanuri Hidayat; Dani Adhipta

AIP Conf. Proc. 1755, 070011 (2016)
https://doi.org/10.1063/1.4958506

Performance Comparison of Intrusion Detection System based Anomaly Detection using Artificial Neural Network and Support Vector Machine

Aditya Nur Cahyo[1, a)], Risanuri Hidayat[1, b)], Dani Adhipta[1, c)]

[1]*Electrical and Information Technology Dept., Gadjah Mada University, Jl. Grafika No. 2 Yogyakarta, Indonesia*

[a)]Corresponding author: aditnc.mti13@mail.ugm.ac.id
[b)]risanuri@ugm.ac.id
[c)]dani@te.ugm.ac.id

Abstract. This study presents a comparison of the detection accuracy of ANN and SVM on the anomaly-based IDS and uses all the features in the dataset. The experiments were performed on two algorithms using KDDCup99 dataset, preprocessing performed on datasets for normalization and scaling attributes which consist of four categories in the dataset. Artificial Neural Network managed to obtain high accuracy in all categories outperformed SVM with accuracy, DoS 92.20%, 90.60% Probe, R2L 89%, and U2R 90.80%. According to the results obtained from experiments using all the features dataset showed that ANN has better performance than SVM in attack detection accuracy.

INTRODUCTION

The progress of Internet technology to bring a positive impact to a wide range of industries, this addition can help the growth of the industry but with the transfer of all the processes to the internet, of course, this raises security vulnerabilities [1]. In recent years, the increasing cyber-attacks, especially attacking the industry and companies that have a service on the Internet [2]. This is an important issue in data security because data security is very important for companies to safeguard their assets and user data. Companies must ensure the confidentiality, integrity and availability of the data, in other words, the data must be stored, managed and maintained properly to protect it from unauthorized access.

There are several ways to prevent cyber-attacks, one of which is by using Intrusion Detection Systems (IDS). IDS is one component of network security that protects data and information security, by monitoring the traffic on a packet of data to detect an intrusion [3]. There are many studies about the anomaly-based IDS using a variety of algorithms, such as machine learning algorithms and data mining [3].

Anomaly-based IDS, in principle, make the detection of the data packets in the network traffic, analyze packets of data that do not fit the normal profile that has been created [4]. Artificial Neural Network (ANN) is an algorithm that uses a neural network to filter the incoming data and forwarded to the expert system [5]. Support Vector Machine (SVM) is a type of learning method that attempts to find a global solution to the optimal value of non-linear classification problems [6]. From several studies showed that ANN and SVM has a good detection accuracy on anomaly-based IDS [5-8].

But there is no direct comparison between the two algorithms are using all the attributes of a dataset and using the proposed architecture. In addition to some research on ANN and SVM, have different parameters and dataset. Some researchers use the method of feature selection from the dataset, which, if too many attributes are eliminated will be less describe reality because it will affect the network traffic to detect attacks in real terms. This study will compare the performance of the detection accuracy of ANN and SVM using the same parameters, using datasets KDDCup99 for training and testing, and use all the attributes of the dataset to better describe the traffic in real, to know which algorithm is better in terms of detection accuracy by using all attribute datasets.

Advances of Science and Technology for Society
AIP Conf. Proc. 1755, 070011-1–070011-7; doi: 10.1063/1.4958506
Published by AIP Publishing. 978-0-7354-1413-6/$30.00

INTRUSION DETECTION SYSTEM

Intrusion Detection System is one component of network security that is used for monitoring data traffic and detect if there is a specific activity which is recognized as an intrusion [3].

IDS can be broadly differentiated on the basis of data sources and detection approaches. Based Data Sources IDS is divided into three models: Network-Based, Host-Based IDS and Hybrid IDS [9]. Host-Based IDS (HIDS) is a model of IDS placed on the host, either the user or the DMZ, ways of working by analyzing the packets to specific hosts, for example, Web Server so that later can be analyzed in the form of traffic and logs of the host. Network-Based IDS (NIDS) is an IDS that is placed at the entrance of network traffic or at the gate of the internet channel. IDS analyzes traffic data packets that occur on the network TCP / IP.

Detection Approaches based IDS can be divided into two modules, called misuse Detection (Signature Based) and Anomaly Detection Based [9]. Signature-Based Detection techniques to detect an intrusion by matching patterns or signatures that already exists in the database with the event (either intrusion or attack) that was going on, the concept if the intrusion using the same pattern will be detected. Anomaly-Based Detection, is a technique for detecting an intrusion event-based anomaly that occurs on the network or host, must first create a profile normal to the IDS to define what normal activity in the host or network so that if circumstances were different premises profile that was created on IDS meal will be recognized as an intrusion or intrusion.

The detection method above is a detection method commonly used in the IDS, except that there is one more method that is Stateful Protocol Analysis, which works by comparing profiles had already made are considered safe for each protocol and then observing any protocol earlier on data traffic and a package [9].

RELATED WORK

IDS has gone through a lot of technological developments, utilizing a variety of techniques and methods, particularly using Machine Learning and Data Mining [3]. The use of data mining and machine learning aims to improve IDS capabilities, so it can detect or identify attacks that have certain patterns and can recognize a new pattern of intrusion.

From several sources of literature, Artificial Neural Network is used for classification of IDS on research [5], where the research using the MLP-based neural network for training, perform preprocessing on the dataset and attribute reduction using Information Gain. In the study [10], proposed the MLP Neural Network-based IDS to training and classification, while [11] using a classification approach Evolving Spiking Neural Network and Multi-Layer Feed Forward ANN in order to classify the exact type of the intrusion or anomaly.

While the Support Vector Machine used in the study [6], which proposes IDS using SVM for classification and also for feature selection. These studies use multiclass SVM is optimized by particle swarm optimization to perform classification at IDS, the research also uses feature selection methods to reduce the number of attributes in the dataset.

The research [12], also using SVM, which at the IDS using SVM classifiers, either to conduct training DoS attacks and for the classification of IDS, which datasets are used somewhat less familiar, namely PMU 2014, this study also applies feature selection from the dataset using Genetic Algorithms. Another study using SVM [13], the proposed IDS using SVM-based radial basis function, where applicable dataset feature selection methods for reducing attributes by using PCA.

ARTIFICIAL NEURAL NETWORK AND SUPPORT VECTOR MACHINE

ANN is almost similar to SVM in terms of functionality and condition problems can be solved. Both are part of the supervised learning classes. From many implementations in various areas of SVM gives a better result than ANN on average, especially in terms of the solution reached. ANN work in finding optimal solutions in the form of local SVM while finding the optimal global solution.

Artificial Neural Network

ANN is a mathematical algorithm to solve the problem inspired by the way biological neurons system, which is presented in the brain in exploring and studying the information. ANN is an adaptive system that contains and consists of interconnected processing elements (neurons) that work together to solve certain problems, such as the

human brain [5]. In a neural network algorithm, the algorithm structure consists of a network with many Processor (neurons). The processor receives data from outside or from other neurons. Then the data received earlier will be accepted through a link called weights. ANN algorithms need to be trained (to learn) to achieve the best results, in other words, the data input received from outside need training or learning to be processed or predicted well. Learning is a process wherein a parameter (synaptic weights and bias levels) of ANN adjusted / adapted through a process of continuous stimulation by the network environment. This type of learning is determined by the parameters of the ongoing changes.

Support Vector Machine

Support Vector Machine (SVM) is a model of learning algorithms that analyze the data and identify a pattern. Training algorithm of SVM can build a model that gives a new example in one category or the other [13]. SVM regression and classification were done by constructing nonlinear decision boundaries. SVM has a high degree of flexibility in classification and regression with varying complexity binary classification by building hyperplane to represent the boundary between the two classes [13]. The best hyperplane is a hyperplane which is located in the middle of the two sets of objects of two classes. Looking for the best hyperplane is equivalent to maximizing the margins or the distance between two sets of objects from different classes. SVMs can learn a pattern of large datasets and can scale better because the complexity of the SVM classification does not depend on the dimensions of the features. SVMs also has the ability to dynamically update the training pattern every time there is a new pattern for the classification [14].

PROPOSED METHOD

This paper compares the ANN and SVM algorithm on a Network Intrusion Detection System (NIDS) based anomaly. Overall the design proposed in this paper uses the type of network IDS or NIDS, using anomaly-based approach, as seen in Fig. 1.

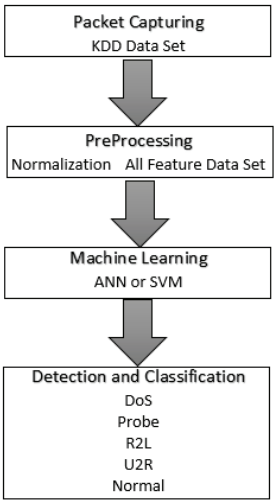

FIGURE 1. Proposed Architecture for Testing IDS

The system design as shown in Fig. 1. is an architecture proposed on testing to compare the two algorithms that ANN and SVM. From the design, the placement of the data is captured by the packet capturing (this study uses KDDCUP99), then after getting the data, the next process is encoded into a form that can be read by the system, then the data is subsequently entered the preprocessing stage, and will be processed by the algorithm applied. In this algorithm consists of four layers that serve to detect any kind of attack.

ANN algorithm in this method works by identifying activities that vary from patterns of users, or groups of users. In this study, using the training function Conjugate Gradient Scale, with the number of ANN input by 41 (the number of features selected), then use 12 hidden layer and one output. The training process to find the best weight preference for the detection performed repeatedly with the same dataset. The training process using Conjugate Gradient Scale as follows: In the first stage, performed initialization of initial weight (w) randomly, initialization of

epoch 0, MSE ≠ 0. If the termination condition has not been met (age <maximum age or MSE> error target), then, age = epoch + 1. Then we carried calculation of total error (SSE), Jacobian matrix and Hessian matrix for the weight vector and bias connections. Then performed the calculation of a new weight vector and bias then calculate the weight difference. Then we conducted calculation error that occurs with weights and biases of the new connection. Once finished, later compared between E(w) with SE(w(new)). If E(w) <= E(w(new)) be repeated in step 5. If E (w)> E (w (new)) then be repeated through step 2. In this study, repetition is done until the total fault current smaller than the required value.

SVM in this research is used for training and classification after data through preprocessing. In this case, SVM classifier will divide the data into two classes, namely -1 (normal, do not detect attacks) and +1 (attack). In this case, the data will be separated into two dimensions, i.e. normal and attack with a hyperplane. SVM used in this research uses a linear kernel function. In each layer of detection, SVM classification was performed by separating between normal and attack with hyperplane with separate data linearly.

$$\begin{aligned} H_1 &: W_0 + W_1 X_1 + W_2 X_2 \geq 1, \textit{for } Yi = +1 \\ H_2 &: W_0 + W_1 X_1 + W_2 X_2 \leq 1, \textit{for } Yi = -1 \end{aligned} \quad (1)$$

So each set of data that falls at or above grade H1 is +1 (attack). While each set of data that is at or below H2 is a class 1 (normal). The detail of the layer on ANN and SVM is shown in Fig. 2.

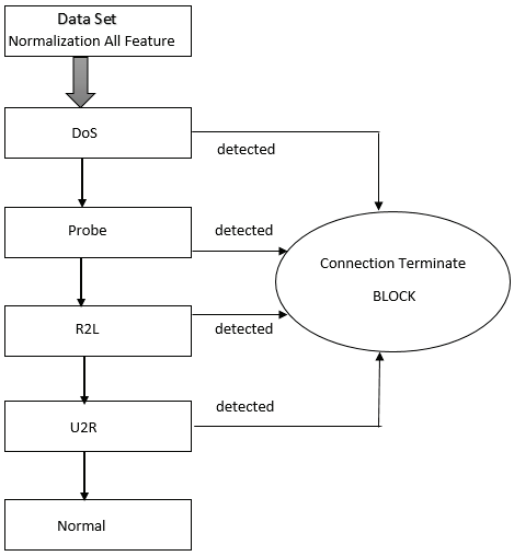

FIGURE 2. Artificial Neural Network and Support Vector Machine Detection Layer

At this stage of the algorithm is applied to both of which includes the process of training and classification consisting of 4 layers, composed of 4 main groups namely DoS attack, Probe, R2L, and U2R. Furthermore, do a comparison between the performances of each algorithm for level detection of attacks using test data.

EXPERIMENTAL SETUP AND RESULTS

In this paper, the research was conducted by using a dataset KDDCUP99 10% were run using Intel Pentium B950 2.1GHz, 4GB RAM with operating system Windows 7 64bit.

Dataset

KDDCUP99 created by DARPA in 1998 consisted of 4,900,000 connections, each connection consists of 41 attributes and labels for this type of attack and are divided into four categories, namely attacks Denial of Services

(DoS), Probe / Scan, Remote to User (R2L) and User to Root (U2R) [18]. KDDCUP99 a dataset that is extensively used for training as well as to evaluate the performance of IDS implemented by researchers.

Because the existing data in the database is composed by numeric and text, then Normalization was performed to convert it to numeric forms. As in protocol_type attributes, tcp to 0, udp to 1 and icmp to 2, then the attack attribute name each layer consists of two classes, 0 for normal, 1 to attack and for other attributes also done the same thing. There are several attributes that have very large numeric data, so it is necessary to scale, duration attribute (0-60000) was changed to (0.0-4.99), attributes src_bytes (0-693376000) was changed to (0.0-9.9), dst_bytes (0- 5204000) changed to (0.0-9.99).

Training

In this study used KDDCup99 dataset, the version used is KDDCup99 10%, consisting of 494.021 of data then grouped into 4 categories attack and use all the attributes of a dataset.

TABLE 1. List training Attack on Training Data

DoS	Probe	R2L	U2R
back	ipsweep	ftp_write	buffer_overflow
land	nmap	guess_passwd	loadmodule
neptune	portsweep	imap	perl
pod	satan	multihop	rootkit
smurf		phf	
teardrop		spy	
		warezclient	
		warezmaster	

In conducting the training in the algorithms are not using all the data on KDDCUP99 10%, the data used for training taken at random from the dataset with the following details: DoS 23.185 (7.908 DoS, 15.277 Non DoS), Probe 20.420 (4.107 Probe, 16.313 Non Probe), R2L 9.642 (1.126 R2L, 8.516 Non R2L), U2R 1.868 (68 U2R, 1.800 Non U2R).

Results and Discussion

In tests performed, the data set used randomly selected from the test data KDDCUP99 10% consists of 50.749 records covering Normal, DoS, Probe, and U2R R2L, where the data is different from the training data there are 14 types of new attacks. Experiments were done using Matlab. Comparison of ANN and SVM were done using all the features on KDDCUP99. The test results can be seen in Fig. 3.

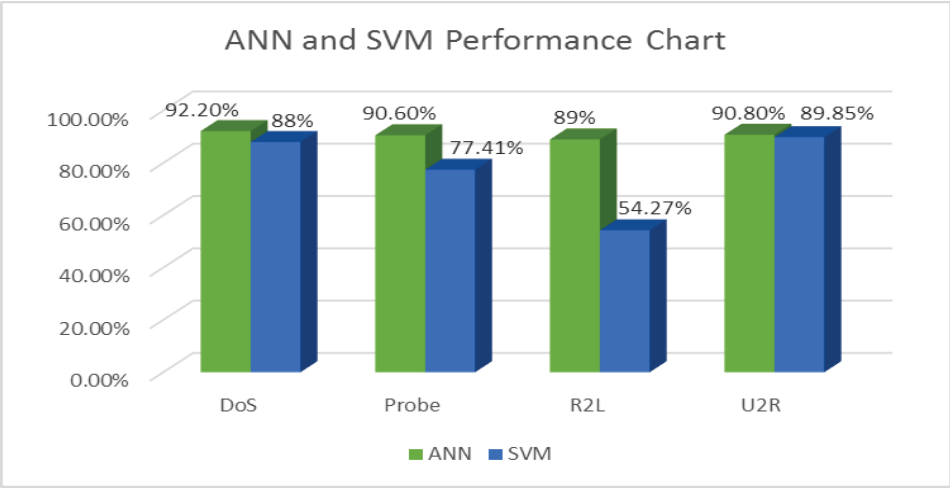

FIGURE 3. ANN and SVM Performance Chart

In the category of DoS, out of 50.749 test data, there are 15.108 new DoS attacks include attacks that apache2, mailbomb, processtable, and udpstorm. ANN algorithm is able to detect quite good at 92.20% while the SVM algorithm is able to detect attacks at 88%. DoS detection rates are quite high compared to other categories, this is in addition to the portion of the training data as well as many DoS attacks have the characteristics of the discrete attributes. Clearly, the ANN algorithm is superior in detecting DoS attacks on the category than SVM algorithm.

In the category of Probe, there are about 4,166 attacks Probe of all test data is performed, of which there is two new attacks that mscan and saint. ANN algorithm successfully detects 91%, while the SM detects 77.41%. The actual allocation of training data for categories does not probe as much as DoS category, it appears that SVM algorithm is only able to achieve an accuracy of 77.41%, while the ANN algorithm helped with the training of its features, so it can achieve the level of accuracy of 90.60%. Thus, the categories Probe ANN algorithm has an accuracy of better detection of attacks on the category Probe.

In the category of R2L, from all tests of data, there are about 16,347 R2L attack data, which include a new attack such as httptunnel, named, sendmail, snmpgetattack, snmpguess, worms, xlock, and xsnoop. ANN algorithm has an accuracy rate of 89% while SVM 54.27%. SVM performs below average with 54.27%, this is because R2L attack does not have the sequence of behavior, and generate a lot of connections in a short time.

In the category U2R, U2R attack data from all tests was 70 data. There is a new attack include ps, sqlattack, and xterm. ANN algorithm is able to detect attacks U2R with the level of 90.80% while SVM has an 89.85% accuracy rate. In this category between ANN and SVM algorithm has a high accuracy, the difference is only 0.95% which the ANN algorithm has better accuracy in detecting attacks on U2R category.

The results obtained are the result of several experiments performed. ANN algorithm has a stable performance in the classification, can be seen in the category of DoS, Probe, R2L and U2R. The results of detection accuracy in addition affected by the classification algorithms, preprocessing is also very influential, either normalization or scale attributes. The selected training data and variations in the chosen parameters are some of the factors that generate high accuracy classifier base. In addition, the use of the Selection Features in the dataset has a significant impact.

When viewed the results of the accuracy of the tests, still revolves around the figure of 90%, several papers have the result close to 100%, but in testing the paper implement feature selection methods to reduce dataset attributes so that the attributes or features reduced. Once again, this study focuses on the accuracy of anomaly-based intrusion detection in IDS, without reducing the attributes in the dataset that will illustrate the performance of IDS more like real network traffic. Data from tests performed on Fig. 3, Artificial Neural Network algorithm is better at detecting attacks than Support Vector Machine algorithm.

CONCLUSION

This paper gives an overview of the review and comparative performance of ANN and SVM. The performance comparison between ANN and SVM were implemented on IDS-based anomaly with using the proposed architecture. In the experiment found that of these two algorithms, overall ANN is superior in detection accuracy with the setup and the architecture of the proposed although fall in the DoS category. The proposed architecture uses a layer categories for classification and detection of attacks. For the future, the normal class and anomaly detection can be improved better by combining various machine learning algorithms, comparing with a variety of architectures and features extraction on normalization

REFERENCES

1. C. Perera, R. Ranjan, L. Wang, S. U. Khan, and A. Y. Zomaya, IT Prof. **17**, 32 (2015).
2. O. Arias, J. Wurm, K. Hoang, and Y. Jin, IEEE Trans. Multi-Scale Comput. Syst. **1**, 99 (2015).
3. S. Sharma and R.K. Gupta, Int. J. Secur. Its Appl. **9**, 69 (2015).
4. E. De La Hoz Franco, A.O. Garcia, J.O. Lopera, E. De La Hoz Correa, and F.M. Palechor, J. Theor. Appl. Inf. Technol. **71**, 324 (2015).
5. K. Jayakumar, T. Revathi, and S. Karpagam, Int. Arab J. Inf. Technol. **12**, 728 (2015).
6. G. Wang, S. Chen, and J. Liu, Int. J. Secur. Its Appl. **9**, 227 (2015).
7. R. Chitrakar and C. Huang, Comput. Secur. **45**, 231 (2014).
8. H. Igor, J. Bohuslava, J. Martin, and N. Martin, 24th DAAAM Int. Symp. Intell. Manuf. Autom. 2013 **69**, 1209 (2014).
9. M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, IEEE Commun. Surv. Tutor. **16**, 303 (2014).

10. Y. Abuadlla, G. Kvascev, S. Gajin, and Z. Jovanović, Comput. Sci. Inf. Syst. **11**, 601 (2014).
11. K. Demertzis and L. Iliadis, *A Hybrid Network Anomaly and Intrusion Detection Approach Based on Evolving Spiking Neural Network Classification* (2014).
12. K. Pradeep Mohan Kumar, M. Aramuthan, and T. Uthra Devi, Int. J. Appl. Eng. Res. **10**, 20081 (2015).
13. N. Kausar, B.B. Samir, I. Ahmad, and M. Hussain, J. Theor. Appl. Inf. Technol. **60**, 55 (2014).
14. F. Kuang, S. Zhang, Z. Jin, and W. Xu, Soft Comput. **19**, 1187 (2014).