International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 6, No. 1, 2019, pp. 167-177.
ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Intrusion detection system with decision tree and combine method algorithm

Saeed Mazraeh [a]*, Maryam Ghanavati[b], Sajedeh Hasan Nejad Neysi[b]

[a]*Sousangerd Branch, Islamic Azad University, Sousangerd, Iran*

[a]*Sousangerd Branch, Islamic Azad University, Sousangerd, Iran*

[a]*Sousangerd Branch, Islamic Azad University, Sousangerd, Iran*

Abstract

By developments in social networks, the number of attacks and intrusions into these networks are increasing in different ways. Hence an intrusion detection system could play an important role in protecting the systems against the hackers. In order to protect computer networks and systems, different solutions introduced which are called intrusion detection system. The purpose of such systems is to protect the computers and networks from prohibited uses, damages and misuse of such systems against internal and external intruders and hackers. Each intrusion detection system may use two different approaches to identify any suspicious (abnormal) activity or intrusion. In order to identify the misuse of systems, intrusion patterns are used; however those approaches used to identify abnormal activities, the normal manner of users are used as the pattern. Decision tree is one of the most famous and oldest methods in data analysis to make categorizing model. In those algorithms based on decision tree, the output is represented as a tree which is made of different states of values. Since the results of them could be interpreted and do not need all input parameters, decision trees are in scope their structures are easy to interpret too. Efficiency of a system strictly depends on selecting the properties. Since by an increase in properties, the computational cost also increases, it seems that design and implementation of systems needs the minimum number of them. The proposed method uses a training set of KDD-Cup99. The proposed method uses three main learning algorithms, SVM, Naive Bayes and J48 decision tree is implemented and evaluated separately. These algorithms are also implemented and evaluated individually as well. The results show the superiority of the proposed method with 97% efficiency using J48 learning algorithm and ADABOOST classification by reducing the dimension IG method.

Keywords: intrusion, intrusion detection systems, machine learning, attack

Received: 27 March 2019/Revised: 22 April 2019/Accepted: 28 May 2019/Published: 28 June 2019

Introduction:

Intrusion Detection Systems is one of the main elements of the security infrastructure in many organizations. These systems are hardware and software models and patterns that automate the process which processes for monitoring the events involved in the network or computer systems. Network-based Computer systems has growing role in modern societies and are targeted and influence by more attacks by enemies and criminals. In addition, intrusion prevention methods such as user authentication such as using passwords, using fire walls, data protection, such as encryption, intrusion detection as well as other wall is used to protect computer networks. The goal of intrusion detection is to identify unauthorized use, abuse or damage to computer systems and networks by both internal users and external attackers (Altwaijry, 2013).

In order to implement intrusion detection methods, several systems as intrusion Detection Systems are designed and manufactured. In the field of computer security, intrusion detection systems play a role in warning and it is announced every time that site security is at risk. Other entity that is responsible for site security is called Site Security Officer that can answer warnings and do appropriate provisions (Hanguang& Yu, 2012).

Intrusion Detection System

Intrusion Detection Systems play the role of warning whenever site security compromises, it can be stated. Another organ that is responsible for Site Security Officer can respond to these warnings and make proper rules. Over the past few years, several different types of intrusion detection systems are built. The first intrusion detection system worked by analysing the Log files that created by the operating system and applications. However, over time these systems were complex they didn't access enough results to conclusively identify an attacker. Thus, the intrusion detection methods turned based on the analysis of more complex data networks or hosts. According to the information source, intrusion detection systems can be divided into two categories: In some cases, they use these two approaches to abnormality detection and misuse detection along each other. In such systems, the abnormality detection method has the task of detecting new and unknown attacks and abuse detection systems, anomaly detection takes on the responsibility to protect. With this act to ensure that information and collected patterns safe for the abnormality detection system.

Over the past few years, different types of intrusion detection systems have been built. The initial intrusion detection systems worked by analysing Log files by operating system and created application files. However, these complex systems were reviewed; they did not have access to identify the attacked data. Thus, concentrations were shifted to the more sophisticated data analysis or host-based methods of intrusion detection.

Intrusion detection systems try to identify the user activities either normal or anomalous transaction which are compared with the network connection and based on known patterns of intrusion that have been designed by experts to examine the intrusion. Traditional methods can't be useful in discovering previously unknown intrusion patterns because manpower during analysis of intrusion detection systems will face computer networks with high speed and complexity. Intelligent decisions about the techniques

and technologies that are based on data mining are used to identify effective and efficient intrusion detection pattern or patterns (Hanguang& Yu, 2012).

Machine learning-based intrusion detection systems are divided into two parts. First is the discovery of abuse and the second is anomaly discovering. Detection systems build intrusion models using learning from labeled data. In this model, system can't identify new attacks. In contrast, anomaly detection systems can detect new and unknown strike (Hanguang& Yu, 2012).

Related Works

So far, a number of researchers inside and outside the country have studied intrusion using data mining techniques, artificial neural network and machine learning algorithms for Intrusion detection systems. Each of the studies is looking at achieving better results in intrusion detection systems. In reference (Altwaijry, 2013) presented a method by the selection of feature to improve the performance of the SVM algorithm. The results on KDD99 in intrusion detection system was evaluated that the results showed the model has high accuracy rate of intrusion detection and higher performance.

In reference (Zhang & Chen, 2012) presented a model using the algorithm of association rules in intrusion detection. The method of using association rules algorithm many of the attacks were identified and the use of coarse-grained algorithm sets theory for systems, intrusion detection and intrusion detection system overall performance was also mentioned.

In reference (Panda et al, 2012) discussed using a combination of class rather than an intrusion detection system in class.

(Srinivasu & Avadhani, 2012) presented a method for an intrusion detection system using genetic algorithms for deriving the weights of the neural network method.

(Khan et al., 2007) presented an intrusion detection system using support vector machine and hierarchical clustering. In this analysis, hierarchical clustering and DGSOT to shorten the time of learning model SVM is used in large data sets. Combination of SVM and DGSOT increases accuracy and decreases false positive and false negative rates.

(Yang et al., 2007) is used LVQ and neural network model to detect network intrusion. The plan is to start the process of feature selection and data normalization. Then, the model is used to the intrusion detection model.

(Horng et al., 2011) presented a combined approach based on hierarchical clustering algorithms and the function selecting a series of important and simple features and combined them with a series of vector machine techniques. The method reduced training time and increased efficiency and also in this paper KDD99 data is used to examine the functionality. The efficiency of this method was in the detection of attacks of Dos and Prob.

(Li et al., 2014) have created a method for the removal of features that seemed to make more efficient of intrusion detection system. This method is based on techniques of removing features and support

vector machine. The method is a combination of the clustering methods, ant colony algorithm, and support vector machines and the results show that the algorithm has a very high accuracy.

(Salama et al., 2011) reduced feature space in a deep belief networks then used Support Vector Machines which tried to divide the sample into 5 categories. Researchers used NSL-KDD data set to train and test the proposed system and the accuracy and speed of the proposed method were calculated. (Muda et al., 2011) proposed Network Intrusion Detection combining with supervised learning methods. K-Means algorithm is used for unsupervised learning and Naive Bayes algorithm for supervised learning. The first step is the using the K-Means algorithm for dividing the data into normal mode or attack types.

Then, using Naive Bayes algorithm the results classification obtained to evaluate the attack of KDD99 data set. Detection rate improved to 99.6 percent. However, this solution is not practical for real networks because the K-Means algorithm requires more time to process huge data in real networks which can lead to bottlenecks and problems encountered in the system.

(Sangkatsanee et al., 2011) is presented in real-time intrusion detection method using a supervised learning technique which is a simple efficient approach and can be used in many machine learning techniques. Therefore, we further develop an intrusion detection system in real time (RT-IDS) using decision tree to classify data on the network as normal data or attack. Finally, a new post-processing method was developed to reduce false rate and also increase the reliability and accuracy of intrusion detection systems.

Research Background

In this method, a pre-defined target variable exists. In this paper, our target variable groups are our attack categories. There are many examples of ways that the amount of the target variable is specified for them, thus, the algorithm can be trained with the help of them that which goal variables corresponding to the predictor variables. Data mining analyzed an extensive set of samples and it is created with the help of the training set. Then this model is evaluated by a test suite and its accuracy is assessed. Various methods are presented in order to classify such as decision trees in the following.

Classification Algorithm:

In this method, a pre-defined target variable exists. In this paper, our target variable groups are our attack categories. There are many examples of ways that the amount of the target variable is specified for them, thus, the algorithm can be trained with the help of them that which goal variables corresponding to the predictor variables. Data mining analyzed an extensive set of samples and it is created with the help of the training set. Then this model is evaluated by a test suite and its accuracy is assessed. Various methods are presented in order to classify such as decision trees in the following.

J48 algorithms

Decision tree classification based on the measure of characteristic selection in decision nodes are divided into two groups: CART (classification and regression tree) and C4.5. CART is the algorithm of classification and regression. J48 decision tree crates an improved tree (pruning) of C4.5 or not pruning C4.5 (Moore et al, 2009)

Naïve Bayes algorithms

Naïve Bayes algorithm is an algorithm based on Bayes theorem, which can be used for classification datasets. This algorithm is based on simplifying assumptions on which in it attribute values with condition of independent of the target variable is considered to be. Naïve Bayes is a probabilistic model and provides a systematic method for data analysis process (Panda & Patra, 2007).Process of this method always starts with the probability distribution given analyzed dataset. Bayes theory is a theoretical background for the statistical approach to classification problems of statistical inference (Panda & Patra, 2007).

SVM Classification

The theory of SVM is from statistics and the basic principle of SVM is finding the optimal linear hyper plane in the feature space that maximally separates the two target classes. Geometrically, the SVM modeling algorithm finds an optimal hyper plane with the maximal margin to separate two classes (Kim et al, 2003).

Hybrid classifiers:

In general, there are two methods for combining the outputs classification: The first group, methods that work directly on the output of the classifiers in order to get better results from a combination of classifiers. Those participating classification to combine these methods are usually heterogeneous. The aim of the second group of techniques is finding the ideal hybrid algorithm to combine the best set of classification. Number of categories used in these methods can be more and all of them have the same structure. In this way, a method of production category is often required with the use of training data collection. Two popular methods of production categories in this way are Bagging and Boosting (Bauer& Kohavi, 1999)..

Bagging Algorithm

The bagging algorithm is only one way of learning. With different sampling by replacement from the training set, several new training sets with the first size is achieved and thus each of them can be taught in a separate category. Bagging has been shown that produces excellent results, especially when the data set is small to achieve Bauer& Kohavi, 1999).

Boosting Algorithm

Boosting algorithm classifier algorithm, which are called a weak learning algorithm. Several times with different training sets are selected according to a previous run executed and finally the answer which will be repeated more are selected. Although this method is time-consuming, but the answer will be secure. This method assigns a weight to each sample of the training data set. First of all weights of samples is equal in the next iteration according to the classification performance, the weight of each sample is corrected, incorrectly classified examples are increased so that the weight of the samples were correctly classified their weight lowered. AdaBoost (Bauer& Kohavi, 1999) is the most well- known method of Boosting.

Boosting method for classification algorithms, which is called a weak learning algorithm; several times with different training sets are executed (selected according to a previous run) and finally answer the most frequently selected, although this method is time-consuming, but it sure would be the answer. AdaBoost (Adaptive Boosting) is an efficient boosting method (Bauer& Kohavi, 1999).

Statement of the Problem

First, a training set of data for the classification algorithm is created, that containing the values for the target variable. Each category of samples of the training set must be specified in advance. Then, with the help of training samples in which there was an educational data collection a provisional data mining model is created. The accuracy and efficiency of the provisional data mining model variable with integer values to the target values obtained from the interim model is evaluated and provisional data mining model is adjusted so that the error rate on the test set is minimal.

Classification Of Attacks

Below are four different categories of network attacks that include:

- Probing or PRB: Groups of an attack in which an attacker can search the network for the purpose of gathering information to find known vulnerabilities. Then, with a map of machines and services available in a network can use the information obtained and to identify behaviours.
- Denial of service attacks, or DOS: Groups of the attack in which the attacker so busy keeping any computational resources and memory resources that are so high they cannot afford legal requests resultant network. Thus, the legitimate user access to network machines is impossibility.
- User to root attack or U2R: Groups of the attack who attackers first, reach to a User ID and then exploit the vulnerabilities of a system until the root system is available. The continuous overflow memory is the most widely used network behaviour due to programming mistakes and environment assumptions arise.
- Remote to user attacks or R2L: In this type of attack, the attacker sends packets of data to a machine through the network. Then it uses the vulnerability to illegally access to find a local user.

The Following Categories of Attacks

We have 21 types of attack in the training data set, each of them belong to one of five categories. In table 1, we classify all the attacks in their respective classes.

Table 1: The following series of attacks in different categories in the dataset KDD CUP '99

Category	Sub category
Normal	Normal
PRB	ipsweep, nmap, portsweep, satan
DOS	Back, land, Neptune, pod, Smurf, teardrop
U2R	buffer_overflow, loadmodule, multihop, perl, rootkit

R2L	ftp_write, guess_passwd, imap, phf, spy, warezclient, warezmaster

Feature Selection

Feature Selection refers to process of selection subset of features to increase performance. One of the most efficient feature selection methods is Information Gain (IG). This method calculates information obtained from a feature for predicting a class with identifying the presence or absence of a given feature in data (Lee & Xiang, 2001).Thus the entropy of the random variable X is defined as (1).

Number of classes is i:

$$H(x) = -\sum_{i} P(X_i)\log 2.P(X_i)$$

The entropy of the random variable after observing the amount of Y as follows (2):

$$H(x|Y) = -\sum_{j} P(Y_j)\sum_{i} P(x_i|y_j)\log 2.P(X_i|Y_j)$$

and calculated by (3):

$$IG\ (X|Y) = H(X) - H\ (X|Y)$$

The Results and Evaluation

To evaluate the results of our classification, confusion matrix was used. Confusion matrix may be used briefly to predict the performance of a classification in data. Usually two classes are mentioned but can be used for any number of classes. The result of a confusion matrix with four categories is shown below. We carried out analysis using the following terms:

TN (True negative): the percentage of valid records that are classified correctly.

TP (True positive): the percentage of attacked records that have been correctly classified.

FP (False positive): the percentage of records that were mistaken as attack while they were authentic activity.

FN (False negative): the percentage of records that wrongly are known as right activities while they are attacking.

Table 2: Clutter Matrix to a Classification Problem with Two Hands

		Predicted Records	
		Category -	Category +
Actual Records	Category -	TN	FP
	Category +	FN	TP

$$Accuracy=(TP+TN)/(TP+FP+TN+FN) \quad (4)$$

$$Precision=TP/(TP+FP) \quad (5)$$

$$Recall=TP/(TP+FN) \quad (6)$$

$$F\text{-}Measure=2\times (Precision \times Recall)/(Precision + Recall) \quad (7)$$

4 algorithms are widely used, compare and contrast analysis are done in this section. In this section we show that the features can significantly influence the performance of an algorithm.

Evaluate on the IG + SVM classifier

Table 3: The Results of the Proposed Method Using IG + SVM Classifier

Category	Recall	Precision	F-measure
Normal	99.20	78.29	87.45
Dos	58.8	99	73.77
PRB	84.1	80.1	82.05
U2R	12.5	57.14	20.55
R2L	5.8	64.19	10.63
Accuracy: 93.90			
Average precision: 75.57			
Average recall: 52.08			
Average F-measure: 54.88			
Classification error: 6.1			

At this stage, SVM classifier is used to evaluate the proposed learning approach. Table 3 shows the results of using SVM classifier. The results show the high efficiency of the proposed method using a support vector machine classifier in the average accuracy of 75.72%, average call times of 52.08%, the average F-measure equal to 54.88%, gross basis equal to 93.90% and the classification error rate is 6.1%.

Evaluate on the IG + Naive Bayes classifier

Table 4: Results of the Proposed Method Using IG + Naive Bayes Classifier

Category	Recall	Precision	F-measure
Normal	68	96	80
Dos	99	94	96
PRB	95	31	47
U2R	32	48	38
R2L	67	1	3
Accuracy: 83.80			
Average recall: 72.20			
Average precision: 54			
Average F-measure: 52.80			
Classification error: 16.20			

At this stage, Naive Bayes classifier is used to evaluate the proposed learning approach. Table 4 shows the results of using Naive Bayes classifier and 22 items. The results show the efficiency in the average accuracy of 54%, average call times of 72.20%, the average F-measure equal to 52.80%, gross basis equal to 83.80% and the classification error rate is 6.20%.

Evaluate the proposed approach on the classification of the IG + J48

Table 5: Evaluation of the Proposed Method on the Classification of the IG + J48

Category	Recall	Precision	F-measure
Normal	99	99	99
Dos	100	100	100
PRB	99	99	99
U2R	96	98	97
R2L	59	70	64
Accuracy: 95			
Average recall: 90.6			
Average precision: 93.2			
Average F-measure: 91.8			
Classification error: 5			

At this stage, J48 classifier is used to evaluate the proposed learning approach. Table 5 shows the results of using J48 classifier and 22 items. The results show high efficiency of J48 in the average accuracy of 90.6%, average call times of 93.2%, the average F-measure equal to 91.8%, gross basis equal to 95% and the classification error rate is 5%.

Evaluate the proposed approach on adaboost + IG + J48 Hybrid Classification

Table 6: Evaluation of the Proposed Method on the Classification of the IG + J48 + Adaboost

Category	Recall	Precision	F-measure
Normal	99	99	99
Dos	100	100	100
PRB	99	99	99
U2R	95	98	97
R2L	76	61	68
Accuracy: 97			
Average recall: 93.8			
Average precision: 91.4			
Average F-measure: 92.6			
Classification error: 3			

At this stage, J48 classifier is used to evaluate the proposed learning approach. Table 6 shows the results of using J48 classifier. The results show high efficiency of J48 in the average accuracy of 91.8%, average call times of 93.8%, the average F-measure equal to 92.4%, gross basis equal to 97% and the classification error rate is 3%.

Table 7: Comparison of Model 2 with Single Models

	SVM	Naive Bayes	J48	The Proposed Method
The Final Accuracy	93.9	83.8	95	97
The Wrong Classification Rate	6.1	16.2	5	3

SVM Algorithm, Naive Bayes, J48, adaboost are ways used in this research project. The important thing is to study feature selection techniques that many studies have been able to extract 22 features valuable for prediction and classification with higher efficiency than other methods. The proposed method of decision tree adaboost with j48 based classifier combination and 22 characteristics in Table (4) has shown better performance.

Conclusion

Intrusion detection systems are the systems that must show security and alert or alarm sound. Security is a very important issue that should be designed in a higher probability of intrusion detection system. Sometimes hardware is not for computing the address of the person or organization and the required hardware can be made for the calculations. Therefore, methods and algorithms that have been introduced in the table above can be used. In order to achieve a high degree of confidence, a security policy is often considered. This policy controls the function of various parts of the system and defines the requirements for supervision. In each of the algorithms, strategies to optimize are presented. Reducing the rate of negative and positive alarm and intrusion detection should also be considered carefully. The positive and negative alarm rate means that sometimes the warnings are not real and there are just so many of them. This aspect should also be taken into account in any software designers and the accuracy and diagnosis should not reduce. The previous methods had lower accuracy. This paper presents a combination of hybrid approach to data mining algorithms and aggregated by category adaptive boost random tree decision tree algorithm, the algorithm is relatively strong and isolated in a single algorithm that combines classifier and the operative reduction and feature extraction method with 22 attribute of individual specification and categories of cumulative (combined) paid The results provide an accuracy method of classify the attacks and intrusion detection.

References

Altwaijry, H. (2013). Bayesian based intrusion detection system. In IAENG Transactions on Engineering Technologies (pp. 29-44). Springer Netherlands.

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine learning,36(1-2), 105-139.

Hanguang, L., & Yu, N. (2012). Intrusion detection technology research based on apriori algorithm. Physics Procedia, 24, 1615-1620.

Horng, S. J., Su, M. Y., Chen, Y. H., Kao, T. W., Chen, R. J., Lai, J. L., & Perkasa, C. D. (2011). A novel intrusion detection system based on hierarchical clustering and support vector machines. Expert systems with Applications, 38(1), 306-313.

Khan, L., Awad, M., & Thuraisingham, B. (2007). A new intrusion detection system using support vector machines and hierarchical clustering. The VLDB Journal—The International Journal on Very Large Data Bases, 16(4), 507-521.

Kim, D. S., & Park, J. S. (2003, January). Network-based intrusion detection with support vector machines. In Information Networking (pp. 747-756). Springer Berlin Heidelberg.

Lee, W., & Xiang, D. (2001). Information-theoretic measures for anomaly detection. In Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium on (pp. 130-143). IEEE.

Madbouly, A. I., Gody, A. M., & Barakat, T. M. (2014). Relevant Feature Selection Model Using Data Mining for Intrusion Detection System. arXiv preprint arXiv:1403.7726.

Moore, S. A., D'addario, D. M., Kurinskas, J., & Weiss, G. M. (2009). Are Decision Trees Always Greener on the Open (Source) Side of the Fence?.weather, 1, 40.

Muda, Z., Yassin, W., Sulaiman, M. N., & Udzir, N. I. (2011). A K-Means and Naive Bayes learning approach for better intrusion detection. Information technology journal, 10(3), 648-655.

Panda, M., & Patra, M. R. (2007). Network intrusion detection using naive bayes. International journal of computer science and network security, 7(12), 258-263.

Panda, M., Abraham, A., & Patra, M. R. (2012). A hybrid intelligent approach for network intrusion detection. Procedia Engineering, 30, 1-9.

Salama, M. A., Eid, H. F., Ramadan, R. A., Darwish, A., & Hassanien, A. E. (2011). Hybrid intelligent intrusion detection scheme. In soft computing in industrial applications (pp. 293-303). Springer Berlin Heidelberg.

Sangkatsanee, P., Wattanapongsakorn, N., & Charnsripinyo, C. (2011). Practical real-time intrusion detection using machine learning approaches.Computer Communications, 34(18), 2227-2235.

Srinivasu, P., & Avadhani, P. S. (2012). Genetic Algorithm based Weight Extraction Algorithm for Artificial Neural Network Classifier in Intrusion Detection. Procedia Engineering, 38, 144-153.

Yang, D., Chen, G., Wang, H., & Liao, X. (2007). Learning vector quantization neural network method for network intrusion detection. Wuhan University Journal of Natural Sciences, 12(1), 147-150.

Zhang, J., & Chen, X. (2012). Research on Intrusion Detection of Database based on Rough Set. Physics Procedia, 25, 1637-1641.