WALIA journal 30(S1): 233-237, 2014
Available online at www.Waliaj.com
ISSN 1026-3861

Intrusion detection in computer networks through a hybrid approach of data mining and decision trees

Tayebeh Rouhani Nejad [1], Mohammadebrahim Shiri Ahmad Abadi [2,*]

[1]Department of Computer Engineering, Boroujerd Branch, Islamic Azad University, Boroujerd, Iran
[2]Department of Computer Engineering, Amirkabir University of Technology, Tehran, Iran

Abstract: Increasing development of information and a trend towards using digital resources has posed a challenge to information security. The most important issue is to protect computer systems against infiltrators and misusers. To protect computer systems and networks against infiltrators, several approaches have been designed called intrusion detection approaches. The purpose of intrusion detection approach is to detect any unauthorized activities, misuses, and damage to computer systems and networks by internal users or external attackers. Intrusion detection systems are one of the major factors of security substructures for several organizations. These systems are models, and hardware and software patterns that automatize processes. They notify user as an alarm. Sometimes these alarms are correct and sometimes incorrect. To avoid these alarms, data mining-based intrusion detection systems are used. In this paper, we suggest a hybrid approach of data mining with feature reduction techniques containing 41 features and decision tree algorithms to improve performance (97.19%).

Key words: *Hybrid algorithms; Data mining; intrusion; Intrusion detection; Feature reduction*

1. Introduction

According to increasing development and importance of computer networks in business, technical and engineering knowledge, and up to date technologies, the amount of attacks and intrusive activities into computer networks has been increased significantly. Studies suggest that the rate of computer attacks is increasing annually. Thereby, protection of data and information against these attacks especially in commercial and military environments is an important issue. Thereby, in a distributed network-based computer system, security of system is very important. Obviously not considering security precautions may result in disorder, deficiency, or temporary stop of system. Thus, the role of intrusion detection systems as special-purpose devices to detect abnormalities and attacks on computer networks has become a crucial issue. For a long time, research on intrusion detection has been mostly focused on abnormality detection and misuse-based techniques. Although in commercial goods, due to its high predictability and accuracy, misuse-based detection is generally preferred, in academic research, due to its theoretical potential for dealing with novel attacks, abnormality detection is normally accepted as a more powerful approach.

2. Intrusion detection system (IDS)

Intrusion detection systems are responsible to identify and detect any unauthorized use, misuse, and damage to systems by both internal and external users. The systems are created as software and hardware systems and have their own advantages and disadvantages. High accuracy and speed, and being free from security failure are some advantages of hardware systems. But simple use of software, compatibility in software condition, and the difference of various operating systems make software systems more common and suitable compared to hardware systems. In general, 3 main functions of intrusion detection systems are (Mahmood, 2011):

- Monitoring and Evaluation
- Detection
- Response

3. Classification of Attacks

Simulated attacks are divided into 4 groups:

- Denial-of-Service (DoS) attack: The purpose of DOS attacks is to disrupt resources or services that users are going to access and use (disrupting services).
- User-to-Root (U2R) attack: These attacks are launched successfully on victim´s machine and are available at the root.
- Remote-to-Local (R2L) attack: In these attacks, attacker intrudes into the user´s machine through remote unauthorized intrusion, starts to abuse user´s legitimate account, and sends the package on the network.

* Corresponding Author.

- Probing attack: In these attacks, computers scan to collect information or discover known vulnerabilities scanning capabilities.

4. KDD-99 Dataset

To test our IDS system we used the DARPA KDD99 Intrusion Detection Evaluation dataset (KDD99, 1999). This dataset was created by Lincoln Laboratory at MIT and was used in The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 the Fifth International Conference on Knowledge Discovery and Data Mining (KDD99, 1999). This dataset is one of the most realistic publicly available sets that include actual attacks (Aickelin et al., 2007). Therefore, researchers have been using this dataset to design and evaluate their intrusion detection systems. Additionally a common dataset allows researchers to compare experimental results. The KDD data set was acquired from raw tcpdump data for a length of nine weeks. It is made up of a large number of network traffic activities that include both normal and malicious connections. The main purpose of this dataset is to determine cooperation of 41 features at KDD-99 dataset to detect attacks.

5. Methods

Following steps should be taken respectively.

5.1. Feature Selection

This is an important step to improve performance of data mining algorithms. Selection of a subset of main features is done according to specific criteria and is a main and often useful approach in data mining to reduce dimensions. Selection of features results in reduction of features. Thereby, it removes irrelevant, redundant or noisy features and has a significant effect on certain applications including speeding up a data mining algorithm, improving learning accuracy, and better understanding of model (Liu et al., 2010). Weka software is a set of the most up to date machine learning algorithms and devices for data pre-processing. Due to the fact that all facilities of Weka are available to users as suitable medium users, users can implement different approaches on their data and choose the best algorithm for their purpose. In this paper, we used Weka 3.7 a machine learning tool along with 2 techniques (Information Gain (IG) and Gain Raito (GR)) to choose a subset of features. The descriptions of these 2 techniques are as follows:

5.1.1. Information gain (IG)

The IG evaluates attributes by measuring their information gain with respect to the class. It discretizes numeric attributes first using MDL based discretization method (Witten et al., 2011). Let C be set consisting of c data samples with m distinct classes. The training dataset ci contains sample of class I. Expected information needed to classify a given sample is calculated by:

$$I = (C_1, C_2, \ldots, C_m) = -\sum_{i=1}^{m} \frac{c_i}{c} \log_2 \left(\frac{c_i}{c}\right) \quad (1)$$

Where is the probability that an arbitrary sample belongs to class Ci. Let feature F has v distinct values {f1, f2, ..., fv }which can divide the training set into v subsets {C1,C2, ..., Cv }where Ci is the subset which has the value fi for feature F. Let Cj contain Cij samples of class i. The entropy of the feature F is given by:

$$E\ (F) = \sum_{i=1}^{v} \frac{c_{ij+\ldots+c_{mj}}}{c} \times I\ (c_{ij+\ldots+c_{mj}}) \quad (2)$$

Information gain for F can be calculated as:

$$Gain\ (F) = I\ (C_1, C_2, \ldots, C_m) - E(F) \quad (3)$$

5.1.2. Gain ratio (GR)

The information gain measures prefer to select attributes having a large number of values. The gain ratio an extension of info gain, attempts to overcome this bias. Gain ratio applies normalization to info gain using a value defined as (Nguyen et al., 2010).

$$SplitInfo\ (C) = -\sum_{i=1}^{v} \left(\frac{|c_i|}{|c|}\right) \log_2 \left(\frac{|c_i|}{|c|}\right) \quad (4)$$

The above value represents the information generated splitting the training data set C into v partitions corresponding to v outcomes of a test on the feature F (Zaman and Karray, 2009).

The gain ratio is defined as:

$$Gain\ Ratio\ (F) = Gain\ (F)/\ SplitInfo\ (S) \quad (5)$$

5.2. Decision Tree Algorithms

In this, the target concept is represented in the form a tree, where the tree is built by using the principle of recursive partitioning. In this, attributes are selected as a partitioning attribute or as a node based on the information gain criteria and then the process continues repeatedly for every child node until all attributes are considered and a decision tree is constructed. Some pruning techniques may further be considered so that the size of the tree is reduced and the over fitting is thereby avoided (Mitchell, 1997). Some of the reasons of importance of decision trees in data mining environment are: their 1- results can be interpreted and do not require all input parameters. 2- process of their structures is relatively rapid and they are flexible. In this section

some effective algorithms in intrusion detection are described. These algorithms are more effective and have higher speed compared with other algorithms.

5.2.1. C4.5 Algorithm

C4.5 algorithm is a process of building preliminary decision tree from training dataset in which training data is presented as a decision tree. Indecision tree internal nodes indicate features, their edges represent feature selection criteria, and leaves are classes. Construction of decision tree has 2 stages: growing and trimming. In the growing stage decision tree is constructed from training data and in the trimming stage a part of tree is trimmed and its branches are not tested. Decision tree can be used for classification of a new sample. In fact, we move across the tree to reach a leave. In each decision node not having leaves, the result of features of node test is calculated, and then the movement continues towards the root of the chosen tree.

5.2.2. Random Forest Algorithm

Random forest as suggested by Breiman (Breiman, 2001) is an ensemble of unpruned classification or regression trees, induced from bootstrap samples of the training data, using random feature selection in the tree induction process. Prediction is made by aggregating the predictions of the ensemble by majority voting for classification. It yields generalization error rate and is more robust to noise. However, similar to most classifiers, RF can also suffer from the curse of learning from an extremely imbalanced training data set. As it is constructed to minimize the overall error rate, it will tend to focus more on the prediction accuracy of the majority class, which often results in poor accuracy for the minority class.

5.2.3. Random Tree Algorithm

Random tree generates several decision trees randomly. On building a tree, algorithm chooses a "remained" feature randomly. On each expansion, the node should be investigated without any impurity.

5.3 Adaboost Algorithm

Adaboost is a popular boosting algorithm for producing a powerful classifier asliner combination of simple weak classifiers. This algorithm was formulated by Yoav Freund and Robert Schapire (Freund and Schapire, 1995). In boosting approach, classification algorithm called weak learning (Polikar, 2006) is implemented repeatedly using different training data chosen according to previous implementation; and finally the most repeated answer is chosen. Although this is a time consuming approach, the answers are reliable. In this approach, every sample of training dataset is assigned a weight. At the beginning the weight of all samples are equal and then according to performance of classifier, the weight of each sample is modified. In fact, the weight of samples classified incorrectly is increased and the weight of samples classified correctly is decreased. Adaboost is the most popular boosting approach.

6. Evaluation of Intrusion Detection Systems

To investigate classification results, standard criteria of confusion matrix are used. Confusion matrix may be used to predict the performance of a classifier for experimental data. Confusion matrix is usually represented as 2 classes, but it can be constructed for any number of classes. On using a classifier, 4 following results are predicted. Using following expressions, data were analyzed:

True negative (TN): the proportion of valid records correctly classified.

True positive (TP): the proportion of attack records correctly classified.

False positive (FP): the proportion of records falsely classified as attacks while in fact they are valid activities.

False negative (FN): the proportion of records classified falsely as valid activity while in fact they are attacks.

Table 1: Confusion matrix for a classification problem of 2 classes

		Predicted Records	
		Classification –	Classification +
Actual Records	Classification –	TN	FP
	Classification +	FN	TP

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$ER = 1 - Accuracy$$

$$F - measure = \frac{2 * Precision_i * Recall_i}{Precision_i + Recall_i}$$

$$Precision = \frac{TP}{TP + FP}$$

7. Results

In this paper, an approach is proposed using GR and IG feature reduction algorithms with selection of 41 features. These algorithms can improve performance of intrusion detection system. In the following section they are compared individually. To evaluate the proposed approach, C4.5 decision tree, random forest, and random tree have been employed. Table 2 illustrates evaluation results of individual algorithms using all dataset and features and hybrid algorithm using GR and IG feature reduction algorithms with selection of 41 features using Adaboost hybrid classifier.

Fig. 1 illustrates the proposed approach of IG +Adaboost +Random tree algorithm which has superior performance compared with hybrid approaches along with feature reduction and single approaches.

In Fig. 2, the proposed approach has lower error rate compared with hybrid approaches along with feature reduction and single approaches.

Table 2: Comparison of result

Feature Reduction Methods	Attribute	TP	FP	Accuracy	precision	Recall	F-Measure
C4.5	-	0.895	0.001	94.690	99.900	89.480	93.287
Random Forest	-	0.902	0.001	95.060	99.940	90.180	93.797
Random Tree	-	0.917	0.001	95.800	99.900	91.700	95.060
IG + Adaboost+ C4.5	33	0.936	0.000	96.780	99.980	93.580	96.291
IG+ Adaboost+Random Forest	37	0.913	0.001	95.630	99.940	91.320	94.702
IG+ Adaboost+Random Tree	**40**	**0.944**	**0.001**	**97.190**	**99.940**	**94.440**	**96.880**
GR+ Adaboost+C4.5	38 & 39	0.932	0.001	96.530	99.879	93.180	95.965
GR+ Adaboost+Random Forest	31	0.920	0.001	95.950	99.918	91.980	95.213
GR+ Adaboost+Random Tree	30	0.927	0.000	96.330	99.960	92.700	95.756

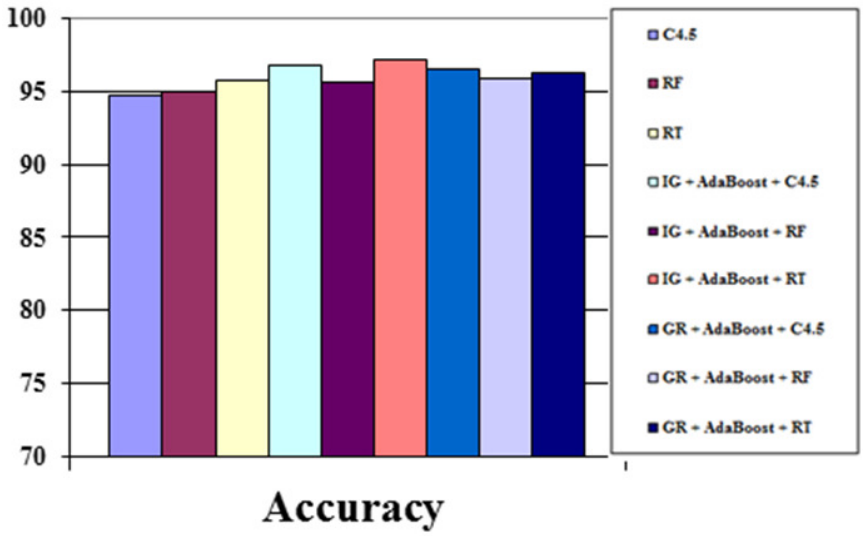

Fig. 1: Comparison of performance of all algorithms

8. Conclusions

Nowadays, intrusion detection systems are one of the major factors of security in networks. Proper use of these systems is useful for networks. Selection of proper algorithms to investigate intrusion detection can improve performance. Also, feature reduction algorithms have effects on improvement of intrusion detection process. In this paper, GR and IG feature reduction approaches and decision tree algorithms were used to improve the rate of intrusion detection. The results revealed that random tree algorithm along with Adaboost hybrid algorithm, IG feature reduction algorithm, feature reduction of41, accuracy of 97.19%, and classification error of 2.81% result in an intrusion detection system of high performance.

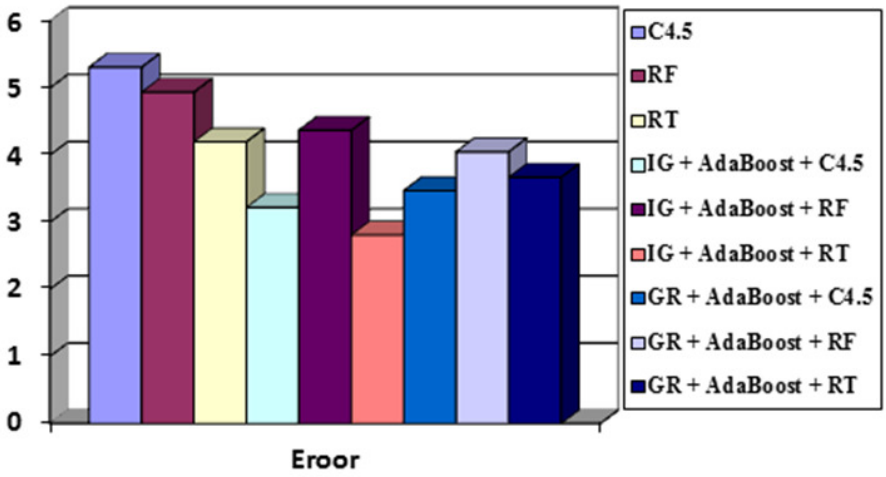

Fig. 2: Comparison of error rate of all algorithms of this paper

References

Aickelin U, Twycross J and Hesketh-Roberts T (2007). Rule generalization in intrusion detection systems using SNORT. International Journal of Electronic Security and Digital Forensics , 1: 101–116.

Breiman L (2001). Random Forests. Machine Learning, 45: 5-32.

Freund Y and Schapire RE (1995). A decision theoretic generalization of online learning and an application to boosting. in Proc. of the European Conference on Computational learning theory, 23–37.

KDD'99 archive: The Fifth International Conference on Knowledge Discovery and Data Mining. <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>.

Liu H, et al. (2010). Feature Selection: An Ever Evolving Frontier in Data Mining, in Proc. of JMLR: Workshop and Conference Proceedings 10: 4- 13 The Fourth Workshop on Feature Selection in Data Mining.

Mahmood SM (2011). Using ant and self-organization maps algorithms to detect and classify intrusion in computer networks, MSc. Thesis University of Mosul.

Mitchell TM (1997). Machine Learning, McGraw Hill, ISBN 0-07-042807-7.-

Nguyen H, Franke K and Petrovic S (2010). Improving Effectiveness of Intrusion Detection by Correlation Feature Selection. in Proc. of Reliability and Security ARES 10 th International Conference on Availability, Krakow, 17-24.

Polikar R (2006). Ensemble Based Systems in Decision Making. IEEE Circuits and Systems Magazine, Third Quarter, 21-45.

Witten IH, Frank E and Hall MA (2011). Data Mining Practical Machine Leanrning Tools & Techniques Third edition, Morgan kaufmann, ISBN. 978-0-12-374856-0.

Zaman S , Karray F (2009). Features selection for intrusion detection systems based on support vector machines, in Proc. of the 6th IEEE Conference on Consumer Communications and Networking Conference , Las Vegas, NV.