Cumhuriyet Üniversitesi Fen Fakültesi
Fen Bilimleri Dergisi (CFD), Cilt:36, No: 3 Özel Sayı (2015)
ISSN: 1300-1949

Cumhuriyet University Faculty of Science
Science Journal (CSJ), Vol. 36, No: 3 Special Issue (2015)
ISSN: 1300-1949

A Hybrid Intrusion Detection System Based on Multilayer Artificial Neural Network and Intelligent Feature Selection

Mehdi MANSOURI[1,*], Mohadese TORABI GOLSEFID[2], Naser NEMATBAKHSH[3]

[1] *Young Researchers and Elite Club, Najafabad Branch, Islamic Azad University, Najafabad, Iran*

[2] *Computer Department, Faculty of Engineering, Shahin Shahr branch, Islamic Azad University, Shahin Shahr, Iran*

[3] *Faculty of Computer Engineering, Najafabad branch, Islamic Azad University, Najafabad, Iran*

Received: 01.02.2015; Accepted: 05.05.2015

Abstract. Increased intrusions into computer networks and cyber-attacks have rendered the immunization of cyberspace one of the most important issues of managers and experts in the recent years. Since cyber-attacks have become more sophisticated and hackers have become more professional, mere use techniques such as firewall, cryptography, biometrics, and antiviruses is not sufficient anymore. Therefore, it is necessary to employ efficient intrusion detection systems. Considering 5 classes of cyber-attacks, a detection intrusion system, of abuse detection type, based on the combination of a multilayer artificial neural network and an intelligent feature selection method was introduced in this research. The research results indicated that the feature selection phase using the proposed method yielded more favorable outcomes than the compared method in terms of the evaluation criteria.

Keywords: Intrusion, Intrusion Detection System, Effective Feature Selection, Multilayer Artificial Neural Network.

1. INTRODUCTION

With the spread of attacks and intrusions which occur in different forms in networks nowadays, techniques such as user identification, data protection, avoidance of programming errors, firewall and so on which are considered as the most primary techniques in computer security will no longer be sufficient. It is obvious that if a password is not valid, then the user in question will not be granted access. Data protection is difficult due to the great volume of data which are transferred through the network. Moreover, it cannot be claimed that systems in which applications are rapidly changing are free from any programming errors. Each error in the system structure can let the intruder take advantage of the system. It is even possible that a mistake by the user results in information accessibility.

Firewalls are vulnerable, too. They cannot provide security for the network on their own. In the presence of firewalls, external attacks will be stopped; however, it is not possible to confront internal attacks. Considering the weaknesses of traditional methods in establishing security and protecting the information completely, presenting more efficient approaches is needed. Many methods which monitor the events occurred in a system or a computer network have been proposed as intrusion detection methods. Along with firewalls, intrusion detection systems are used as a security supplementary. Undoubtedly, it is essential to conduct research on the design of an intrusion detection system which identifies attacks at an ideal rate and announces fewer errors. Intrusion detection system is a program that monitors the events occurred to a computer networks for the sake of security problems. Such systems fall into two main groups including abuse detection and anomaly detection in terms of their intrusion detection method. In the abuse detection technique, these systems use known signatures pertaining to intrusions or

*Corresponding author. *Email address: Mansori.mehdi@gmail.com*

Special Issue: The Second National Conference on Applied Research in Science and Technology

http://dergi.cumhuriyet.edu.tr/cumuscij

vulnerabilities and search for the activities which are similar to such signatures. In the anomaly detection technique, intrusion detection systems are be able to detect the intrusion through search for abnormalities in the network traffic [1].

This paper aimed at designing an intrusion detection system of the abuse type with the help of a hybrid method for selecting features. Given the large size of data in intrusion detection and the importance of knowledge discovery in data-mining, the data needs to be preprocessed. The intelligent feature selection which was used in this research decreased the dimensions of problem to some extent and improved the performance of the system. KDD CUP99 dataset is used in this research for learning and evaluation purposes. It is a standard dataset which has been used in many research projects [2, 5]. In KDD CUP99 dataset, the attacks fell into four groups including Probe, U2R, R2L, and DOS. Considering the normal case as another group, the inputs would belong to one of the five above-mentioned categories [6].

2. RESEARCH BACKGROUND

Recent studies indicate that there are different types of attacks which are more difficult to detect than the previously-known attacks. This matter clarifies the importance of applying methods which are more generable. Some approaches were taken into account in recent years. Some of the data-mining approaches used for abuse detection and anomaly detection include statistical methods, artificial intelligence network, Markov's hidden model, decision tree, and data-mining association regulations which are able to discover adaptable and useful patterns of features in designing the intrusion detection systems [7, 9]. In most data-mining events pertaining to the problem of intrusion detection, classification and clustering techniques were selected to detect the type of attack. Statistical methods are among the most comprehensive methods used in the intrusion detection system. The statistical techniques are used in cases in which there is a general idea on the expected relationships.

In this regard, in a paper named Classifying the Traffic with a Statistical Method, Zuev et al. used the statistical methods to classify the network traffic in 2005. In this research, training and evaluation data were taken from the Cambridge dataset. They were based on the probable data classification technique for categorizing the data. In this method, each datum belongs to a class with a certain degree of probability. Their approach was able to cope with the unstable nature of the Internet traffic. They used the Simple Bayes' Theorem as the classifier which is a simple method of supervised learning. Results indicated that their method was more accurate than other data-mining methods due to benefiting from statistical methods in data-mining [10].

Statistical methods are more difficult to interpret because of their mathematical structure; however, the results and outcome interpretations are more accurate in this method than others. After wholly statistical methods, researchers have used machine-learning and data-mining tools to create intelligent intrusion detection systems. Many research works have been conducted on intrusion detection systems based on artificial neural networks. For instance, Wang et al. conducted a research named A New Approach to Detecting Intrusion Using the Artificial Neural Network and Fuzzy Clustering in 2010. In this research, they presented a new approach named FC-ANN based on artificial neural networks and fuzzy clustering. Their proposed method resulted in an intrusion detection system of high rate and a low FP rate for more stability. The general process of proposed method includes the following steps:

First, the fuzzy clustering technique is used to generate different training sets. After that, the artificial neural network models are trained in accordance with different training sets.

Finally, a meta-learner model and a cumulative fuzzy model are used to integrate the results. In the proposed method, the attacks fell into four groups including R2L, U2R, Probe, and Dos. The selection is conducted at random in order to decrease the number of datasets. The criteria which are often used to evaluate the intrusion detection include NP, TP, and FP. However, given the fact that the number of samples of U2R, Probe and R2L attacks is very small in testing and training sets, using such criteria to test the efficiency of the system would be futile. Therefore, they used other criteria such as F-value, Recall and Precision which do not depend on the size of training samples. The results of KDD CUP99 indicated that the proposed method i.e. the FC-ANN method was better than BPNN and other well-known methods such as decision tree and Simple Bayes' Theorem [11].

3. SAMPLING THE TRAINING DATA

Given the fact that the number of training samples was very large in the KDD CUP99 dataset, the training data should be sampled in the initial step. In this paper, 5 classes of Probe, R2L, U2R, Dos and Normal were considered for detection. These classes are indicated in Table 1.

Table 1. Computed parameters at two considered.

Class Number	Class Name	Instance count	
		Train	Test
1	Normal	391458	229853
2	Probe	4107	4166
3	Dos	1126	16189
4	R2L	52	288
5	U2R	97277	60593

4. NORMALIZATION

In dataset KDD CUP99, the training data had features of symbolic, discrete and continuous values. In some cases, the domains of some variables are different from those of others, and data-mining algorithms are not able to explore such data efficiently.

Symbolic features such as protocol_type (3 different signs), service (70 different signs), and flag (11 different signs) were turned into integers in [0, S-1]. S refers to the number of samples here.

The domains of integers pertaining to each feature were different. Continuous features which include smalls domains of integers such as wrong_fragment: [0, 3], urgent: [0, 14], hot: [0, 101], num_failed_logins: [0, 5], num_compromised: [0, 9], su_attempted: [0, 2], num_root: [0, 7468], num_file_creations: [0, 100], num_shells: [0, 5], num_access_files: [0, 9], count: [0, 511], srv_count: [0, 511], dst_host_count: [0, 255], and dst_host_srv_count: [0, 255] were scaled in [0.0, 1.0] linearly.

Logarithm scaling (on the basis of 10) would be used for continuous features which include very large domains of integers such as duration: [0, 58329], src-bytes: [0, 1.3 billion], and dst_bytes: [0, 1.3 billion] in order to decrease the domains of such features.

Other features were either 0 or 1 (such as logged_in), or they were continuous features which were already in [0.0, 1.0] (such as diff_srt_rate). Therefore, these features were not supposed to be scaled.

In this paper, the following equation was used to scale the features:

$$N(m,n) = (C(m,n) - m(n))/(M(n) - m(n)) \quad (1)$$

In which M(n) is the maximum value pertaining to column n; m(n) is the minimum value pertaining to column n; C(m, n) is the integer pertaining to row m and column n; and N(m, n) is the normal value of row m of feature n.

5. EFFECTIVE FEATURE SELECTION

In this step, the objective is to find a subset of features with the appropriate efficiency. It should be comparable to a complete set of features, too. Additional and irrelevant features are deleted from datasets in a way that this process will decrease the dimensions of problem [13]. Finding a subset of features in a large set is a problem which can occur in many fields. Considering the fact that increasing the number of features would increase the computational cost of a system, it appears essential to design and implement the systems with the least number of features. Therefore, an effective subset of features should be selected in this step so as to maintain an acceptable efficiency for the system.

Since this research aimed to investigate the impact of effective feature selection step on the improvement of intrusion factors, two methods of Best First and Rep Tree were used to select the features and do the comparison consequently. After applying Best First, 12 features were selected out of 41 available features, and other ones were omitted. Table 2 indicated the selected features.

Table 2. Names of Features after Applying Feature Selection Method Best First

Feature	Abbreviated
Service	A1
protocol_type	A2
Flag	A3
src_bytes	A4
dst_bytes	A5
logged_in	A6
Count	A7
srv_count	A8
serror_rate	A9
rerror_rate	A10
diff_srv_rate	A11
dst_host_same_src_port_rate	A12

Moreover, after applying Rep Tree, 9 features were selected out of 41 features, and other ones were omitted. In Table 3, the selected features are indicated in abbreviations.

Table 3. Names of Features after Applying Rep Tree

Feature	Abbreviated
Service	A1
src_bytes	A2
Count	A3
srv_count	A4
serror_rate	A5
rerror_rate	A6
dst_host_count	A7
dst_host_srv_count	A8
dst_host_srv_diff_host_rate	A9

The Gridystepwise method was used to select the features in a way that 6 features were selected, and other ones were omitted. In Table 4, the selected features are indicated in abbreviations.

Table 4. Names of Features after Applying Gridystepwise.

Feature	Abbreviated
protocol_type	A1
src_bytes	A2
dst_bytes	A3
rerror_rate	A4
diff_srv_rate	A7
dst_host_same_src_port_rate	A9

After applying three above-mentioned methods of feature selection, the features which were selected in common were considered as inputs to be used in the learning step. In Table 5, the final features were selected to be applied to the learning step.

Table 5. Finally-Selected Features.

Fature	Abbreviated
protocol_type	A1
src_bytes	A2
rerror_rate	A4

6. LEARNING THE TRAINING DATA

In this step, the multilayer artificial neural network was used to learn the training data. The multilayer artificial neural network is a type of neural network in which the number of neurons of input layer is equal to the number of features, and the number of neurons of output layer is equal to the number of classes.

Learning is a process by which the artificial neural network adapts itself to a stimulus in a way that it gives the desirable response after modifying the appropriate networks parameters. Moreover, neural networks are responsible for classifying each stimulus during training. Therefore, when a stimulus is applied to the network, it detects the stimulus and puts it in the available classes.

This step is actually named intrusion detection. The neural network which is used for learning in this step has 3 inputs and 5 neurons for the output. The number of inputs indicates the number of features, whereas the number of outputs indicates the number of classes. After learning and training, the intrusion system should be evaluated.

7. EVALUATION CRITERIA

After performing the above-mentioned steps, the feature selection methods should be evaluated. Usually, criteria such as false/positive, true/negative, true/positive and false/negative are used to evaluate the accuracy of an intrusion detection system. True/positive indicates that the system has truly detected a particular attack which has really occurred. True/negative indicates that the system did not detect a normal connection and non-attack as an attack. False/positive indicate that the system has wrongly detected a particular type of attack. Finally, false/negative indicates that the system was unable to detect an attack which has really occurred. Given the fact that the number of samples is not appropriate for some classes in the datasets, it is necessary to adopt criteria whose validity would not be negatively influenced by the number of testing and training samples.

Precision, Recall, and F-value are among the criteria which do not depend on the size of testing and training data. They were used to evaluate the system here. The values which are close to 1 will be appropriate for these criteria. The following equations indicate how Precision and Recall are calculated in a way that FP, TP, and FN refer to False Negative, True Positive, and False Negative, respectively. Furthermore, Equation 3 indicates how F-value is calculated, and β is equal to the relative importance of Precision in comparison with Recall. Its value is usually considered to be one.

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \quad (3)$$

$$F - value = \frac{\left(1 + \beta^2\right) * Recall * Precision}{\beta^2 * \left(Recall + Precision\right)} \quad (4)$$

8. Results and Findings

The results of evaluating the criteria Precision, Recall and F-value indicate that the proposed feature selection method had a positive impact on these criteria in comparison with other methods. Table 6 compares these results with those of another method.

Table 6. Comparison of Criteria Precision, Recall, and F-value.

Class Name	Evaluation criteria	Method	
		FC-ANN [2]	Proposed Method
Normal	Precision	0.913	0.99
	Recall	0.991	0.994
	F-value	0.95	0.992
Probe	Precision	0.481	0.494
	Recall	0.8	1
	F-value	0.601	1
Dos	Precision	0.999	0.995
	Recall	0.967	0.983
	F-value	0.983	0.986
R2L	Precision	0.932	0.987
	Recall	0.586	0.865
	F-value	0.719	0.791
U2R	Precision	0.833	1
	Recall	0.769	0.843
	F-value	0.8	0.834

9. CONCLUSIONS

In this research, an abuse type intrusion detection system was proposed using the artificial neural network and intelligent feature selection. The problem of intrusion detection includes 5 classes in intrusion detection. After sampling, normalization was done along with decreasing the dimensions of the problem from 41 to 3 through hybrid feature selection method. The results indicated that Precision, Recall, F-value performed better on the same dataset than the compared

method in terms of the evaluation criteria, and it improved the performance of the intrusion detection system.

REFERENCES

[1] D. Fisch, A. Hofmann and B. Sick, " On the versatility of radial basis function neural networks: A case study in the field of intrusion detection", Information Sciences, Volume 180, Issue 12, pp. 2421-2439, 2010.

[2] Wang, G., et al., "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering". Expert Systems with Applications, 2010. 37(9): p. 6225-6232.

[3] Li, Y., et al., "Network anomaly detection based on TCM-KNN algorithm, in Proceedings of the 2nd ACM symposium on Information", computer and communications security. 2007, ACM: Singapore. p. 13-19.

[4] Muna Mhammad T. Jawhar. Et al.," Design Network Intrusion Detection System using hybrid Fuzzy-Neural Network". Computer Science and Engineering, 2011. V (4): Issue (3).

[5] Horng, S.-J., et al., "A novel intrusion detection system based on hierarchical clustering and support vector machines". Expert Systems with Applications, 2011. 38(1): p. 306-313.

[6] M. Sheikhan and M. Sharifi Rad, "Misuse detection based on feature selection by fuzzy association rule mining", World Applied Sciences Journal, 10 (Special Issue of Computer & Electrical Engineering), pp. 32-40, 2010.

[7] Pietraszek, T. and A. Tanner, Data mining and machine learning-Towards reducing false positives in intrusion detection. Inf. Secur. Tech. Rep., 2005. 10(3): p. 169-183.

[8] Sangkatsanee, P., N. Wattanapongsakorn, and C. Charnsripinyo, "Practical real-time intrusion detection using machine learning approaches". Computer Communications, 2011. 34(18): p. 2227-2235.

[9] Yun Wang, "A multinomial logistic regression modeling approach for anomaly intrusion detection", Computers & Security, Volume 24, Issue 8, November 2005, P662-674.

[10] Zuev, D. and A. Moore . Traffic Classification Using a Statistical Approach.Passive and Active Network Measurement. C. Dovrolis, Springer Berlin Heidelberg.,2005 . 3431: p. 321-324.

[11] Wang, G., et al., "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering". Expert Systems with Applications, 2010. 37(9): p. 6225-6232.

[12] Witten, I. H., & Frank, E. (2005). Data mining: Practical machine learning tools and techniques. Boston: Morgan Kaufmann Publishers.

[13] A. Z. Al-Garni, A. Jamal, A. M. Ahmad.2006. "Neural network-based failure rate prediction for De Havilland Dashtires". Journal of Engineering Applications of Artificial Intelligence 19 681-691.