Ciência e Natura, v. 37 Part 2 2015, p. 168–174

ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

ciência e natura

Categories Malware using Neural Networks based on Feature Selection by Genetic algorithm

Fatemeh Farahmand [1], Seyed Javad Mirabed [2]

[1] Young Researcher club, Department of Computer, Bushehr branch, Islamic Azad University, Bushehr branch, Iran.

[2] Department of Computer, Islamic Azad University, Central Tehran Branch,Tehran,Iran.

ABSTRACT

Concurrent with the ever-increasing growth of information and communication technology (ICT) and the dramatic expansion of computer networks, we observe different forms of attacks and intrusions to networks; thus intrusion detection systems (IDSs) are considered as a vital part of each network connected to internet in the modern world. Neural networks are considered as a popular method used in IDS. Two major problems in these networks, i.e. long training time and inattention to features' domain, have made necessary development and/or improvement of the model. Feature selection techniques are used in the neural networks in order to develop a new model to speed up the attack detection, to reduce error notification rate and finally to enhance system's efficiency. In this study, for enhancing efficiency of the neural network in detecting intrusions, a genetic algorithm was used for selecting features. The suggested model was examined and assessed on NSL-KDD dataset which is the modified version of the KDD-CUP99. The experimental results indicate that the suggested model is very efficient in enhancing precision and recall of attack detection and reducing the error notification rate and also is able to offer more accurate detections in contrast to the basic models

Keywords: *neural networks, genetic algorithm, feature selection, IDS*

Recebido: dia/mês/ano Aceito: dia/mês/ano

1 Introduction

As a useful task in various sciences, data classification has always been considered by researchers; with many methods have been suggested for this task and each method has been designed to be used in different tasks. The main objective of data classification is to detect various classes in a single dataset by which the position of the new sample in this set can be determined. Classification is one of the most important objectives of model detection, which due to its many applications, is emphasized today. Many methods of classification have been invented so far including: 1) decision tree induction, 2) Bays Theory, 3) support vector machines (SVM), and 4) Neural Networks.

Artificial Neural Network method was introduced as a hopeful method for detecting intrusion which sometimes has showed a high potential capability to detect detrimental attacks. The major problem here is the long training time which along with inattention to features domain makes necessary development and/or improvement of the model. Irrelevant data and extra features will result in weak classification, numerous calculations and a low efficiency. In general, selecting proper features for machine learning algorithms reduces data size and algorithm's consuming space and as a result speeds it up. It sometimes improves the precision of classification. Feature selection techniques can be used to minimize features size and to delete features that are useless in making decision for reaching a certain class. Minimizing the dataset size and weighted learning algorithm in analysis is suggested for finding an optimal feature set for the intrusion detection system; it also can be useful in speeding up system and reducing the time needed to learn features of each attack.

Section 3 and section 4 of this paper deal with genetic algorithm-based neural network and the importance of feature selection methods, respectively. Section 5 introduces details of the suggested technique including how to use the genetic algorithm along with the neural network for detecting intrusion. And finally summary and conclusions are provided in the section 6 of the paper.

2 Literature Review

Intrusion detection system (IDS) detects attacks through selecting and analyzing the network traffic and uses various effective methods such as machine learning techniques in this way [1]. Recently, artificial neural networks have been introduced as an alternative and helpful means to model complicated systems and are used widely in predictions. In general, learning by neural networks is divided into two classes of supervised learning and unsupervised learning [2, 3]. [4] provides an important study about how to use neural network to detect and classify intrusions. This study aims at determining those neural networks that classify attacks very well and records a higher detection rate for attacks. This study focuses on classifying two types of records: a single class (normal or attack) and multi-class, where classification of attacks is determined by neural networks.

Tong et al. [5] used a hybrid neural network to determine anomalies and misbehaviors. Radial basis function (RBF) networks, as a Real-time pattern classification, and ELMAN network have been used for reconstructing the memory from previous records. RBF networks use a local recurrence exponential function for estimating both nonlinear local input and output. In contrast to multilayer perceptron networks, RBF network needs a shorter time to learn. Each ELMAN network has a set of content nodes. Content nodes receive their inputs from a node in the hidden layer and send its output to a node in the same layer. Since content nodes only depend on reactors in the hidden layer of the previous input, content nodes keep the input data. A recurrent neural network based on inputs grouping whose size has been reduced was used in [6]. Reducing size speeds the network up. For this purpose, input features have been divided into four classes: content features, basic features, time-based traffic features and host-based traffic features. In this network, the input layer is connected to the hidden layer based on the feature's related class. Moradi et al. [7] have tried to detect the Denial of Service Attack (DoS) in network branches using feed forward neural networks. For solving two problems of precision and sustainability of weak detection related to neural network based IDS, Wang et al. [8] used the combination of neural network and fuzzy clustering with the approach of anomalies detection. For them the imbalanced state of various attacks can be a main reason for

Recebido: dia/mês/ano Aceito: dia/mês/ano

these two problems and analyzed two types of attacks, R2L and U2R, for this purpose. They called their data provision method Fuzzy Clustering- Artificial Neural Networks FC-ANN in which the whole set of learning initially is broken into subsets with fewer numbers and less complication using fuzzy clustering method and a proper neural network is applied on each subset. Each neural network is able to learn each subset faster, stronger and more accurate. And finally the main output is induced from all neural networks using fuzzy aggregation method. In the third step, for refining errors of various neural networks, an ultra-learner, fuzzy aggregation module for relearning and relearning and combining different neural networks whose database needs continuous updating. An IDS with the fixed basis used different statistical methods; however this methods needs to collect many and sufficient data for developing a complicated mathematical model, which is impossible regarding the network's heavy traffic. As we observed, for bypassing the mentioned constraints there are different methods in which neural networks have worked successfully as a fixed basis for the mentioned methods.

Fundamentally, evolutionary methods including genetic algorithm have broad applications in various stages of designing neural networks, as they have unique capabilities in finding optimal amounts and capability search in unpredicted spaces. Hence, this study used genetic algorithm for designing both form and structure of the neural network.

3 Genetic Algorithm Based Neural Network

Neural networks have been used in different fields; however, they reached ideal results only when a rich source of data along with numerous observations are available; thus, this technique is useless when there are few data for the network training and this deficiency limits their applications considerably [9]. A brief review of previous studies makes it clear that nonlinear models and especially neural networks have had the most applications in such studies. However, using these methods is accompanied with certain constraints such as training data, learning algorithm stop at the local optimal point and need to determine a certain functional form. Various smart techniques have been suggested to tackle these constraints: generate artificial training samples, parameter fine-tuning for the inductive model and automatic measurement of network's weights using genetic algorithm were among the suggested techniques in this area. Different studies including [10-12] faced data constraints in solving problems. However, the suggested techniques in this area have several weak points: lack of potential to develop generate samples and lack of such samples in reality, choosing improper models, constraints in combining input variables, using linear techniques in adjusting parameters of the nonlinear model and author's prejudice about network's structure were some of constraints of such techniques. In this study, we decide to use the genetic algorithm as the method to select features.

4 Importance of feature selection for IDS

Since intrusion detection systems need a great deal of information to analyze a network, both analysis and classification are very difficult tasks. Therefore, an intrusion detection system needs to minimize the data size for processing purposes which is necessary to remove or decrease the complicity of relations among some features [13]. Since increasing the number of features increases the computing costs of a system, both designing and launching system with the lowest number of features seems necessary. On the other hand, paying attention to this issue is very important that we need to select a useful set of features in order to be able to develop an acceptable efficiency for the system. The main goal of selecting a feature is declining the feature vector’s dimension in classification, as an acceptable classification rate is reached too. Under this condition, features with less differentiation power are removed and features containing proper information for differentiating model classes are kept. Feature selection in IDSs simplifies the problem and makes detection faster and more accurate by deleting duplicate and irrelevant data. One of the most optimum techniques for data mining is genetic algorithm which is used to extract and select features. In the suggested model in this study, the genetic algorithm is used for selecting feature.

Recebido: dia/mês/ano Aceito: dia/mês/ano

5 The suggested combined method

In this section, the suggested combined method is analyzed. The suggested method is based on the evolutionary theory. The main dataset includes N features which is divided into different subsets, n_1, n_2, ….. to be used in the genetic algorithm. Then, these subsets are moved for training to the feed backward neural network with a fixed number of neurons in the hidden layer. Number of input layer's neurons depends on number of features of the subset in question. Any subset of data is divided into training and test parts. After training, network is tested with a set of new data (test dataset). Number of wrongly classified samples is considered as the error of the subset in question. Thus, any subset of features has an estimation error which will be helpful in detecting the best subset. Thus, the neural network helps the genetic algorithm population members in finding the best solution. The last subset that is achieved with GA (which is the most optimal) is trained once more with a neural network for a more population.

The algorithm is applied as follows:

Stage 1: Initialization, formation of the initial population regarding the considered number, adjusting remix parameters and mutation on population, initialization of duplications of genetic algorithm and number remained populations in each duplication;

Stage 2: Any member in a population shows a different solution. Competence rate (converse error) of each member is measured by neural network (after passing sufficient courses of education) for the whole set and out of which μ members of the best are selected as the parents of the next generation;

Stage 3: The following stages are carried out M times;

Stage 4: Formation of the new population (offspring) from the current population of the present generation using combined reactors and mutation;

Stage 5: Calculating competence rate of offspring using neural network;

Stage 6: Selecting the best ones out of offspring and parents as remnants of this generation, in fact it helps us to move towards optimal answer. Then we will move to the next generation with the new population. The key point is that before completing this stage, the error rate of the remaining people needs to be calculated and if the result is less than a standard rate, the algorithm is ended, otherwise it will send to the next generation. In other words, for completing algorithm, we used two mechanisms of iteration and achieving a certain level of error;

Stage 7: the algorithm is ended.

After finding subset of the final reduced features, all features that are not available in the reduced set are deleted from the set of features and are entered into the neural network for training and testing. In this study, genetic algorithm was used to assimilate similar features and reducing them. Using genetic algorithm speeds up processing and increases attacks detection rate.

5.1 Dataset

NSL-KDD [14] dataset was used to assess the suggested method. This dataset includes selected records of KDD-CUP99 [15] whose problems have been raised like the duplicate records. Since 1999 KDD-CUP99 has been the most popular dataset used to assess intrusion detection methods. It was developed by Stolfo et al. [16] based on IDS assessment program, DARPA'98 [17].

DARPA'98 is a set consisting about five million records of various links each which has about 100 bytes capacity. The KDD training collection includes about 4900000 vectors of various bonds and each vector includes 41 feature which are classified as normal and attack. Table 1 shows the relative distribution of various classes of data in the training and testing sets.

Table 1. Approximate distribution of training data and test data sets NSL-KDD

Category	Training	Test
NORMAL	48%	19%
PRB	20%	1%
DOS	26%	73%
U2R	0.2%	0.07%
R2L	5%	5%

5.2 Assessment Criteria

In this task, the main goal is detecting intrusion in the first place and then classifying the intrusion in the second place. In fact, if attacks and intrusions are classified correctly, selecting

Recebido: dia/mês/ano Aceito: dia/mês/ano

the proper decision for any sort of attack is made regarding the case and decisions of the network administrator. Precision and recall are two variables or parameters which are analyzed in this part; these two parameters are known and important in assessing data-mining and/or machine learning based algorithms. Precision's formal definition is as follows:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

And recall's definition is as follows:

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

Where TP (true positive) represents data which are truly dedicated to the positive class, FP (false positive) represents data which are falsely dedicated to the positive class and FN (false negative) represents data that falsely dedicated to negative class.

5.3 Results

Regarding the nature of the casual research in the genetic algorithm and production of a different initial population in any application, in this study we applied the suggested combined algorithm,, regarding the predefined parameters, ten times and then considered the average of precision and recall rates of IDS as the average value for these criteria from the predicted results. Likewise, it was observed that in any application, after selecting important features, input data were decreased about 70 percent.

There are several advantages in using feature selection in the suggested model including reducing the training and testing time in the neural network and hence decreasing computation costs and as a result decreasing computer resources such as memory and processing time which are necessary for detecting attacks. For this purpose, training and testing time was measure in two modes, without feature selection (BP) and with feature selection (GA-BP). Table 2 shows the results. Times are shown in millisecond (ms).

Comparing value of precision, recall split by attack class for both mentioned modes are shown in table 3 and 4, respectively, what is clear cut here is that there is a direct relationship between feature selection and precision and recall values of IDS.

Table 2. Average time training and testing neural networks, without / with feature selection (After 10 times the implementation of the proposed method)

	Without feature selection(BP)	Feature Selection (GA-BP)	Reduce Time
Time Training	249473ms	72546ms	70.92%
Time Test	25247ms	4165ms	83.25%

Table 3. Precision of the neural network for each class of attack

	Normal	DoS	PRB	U2R	R2L
BP	76.69%	97.46%	99.73%	52.96%	64.13%
The proposed method (GA-BP)	96.63%	97.47%	99.78%	56.14%	98.74%

Table 4. recall of the neural network for each class of attack

Recebido: dia/mês/ano Aceito: dia/mês/ano

	Normal	DoS	PRB	U2R	R2L
BP	98.21%	85.78%	97.49%	41.14%	52.76%
The proposed method (GA-BP)	98.54%	84.07%	98.33%	46.37%	64.39%

The mentioned criteria roughly in both modes of BP and GA-BP in neural network are higher than the logistic regression (refer to Figure 1 and 2).

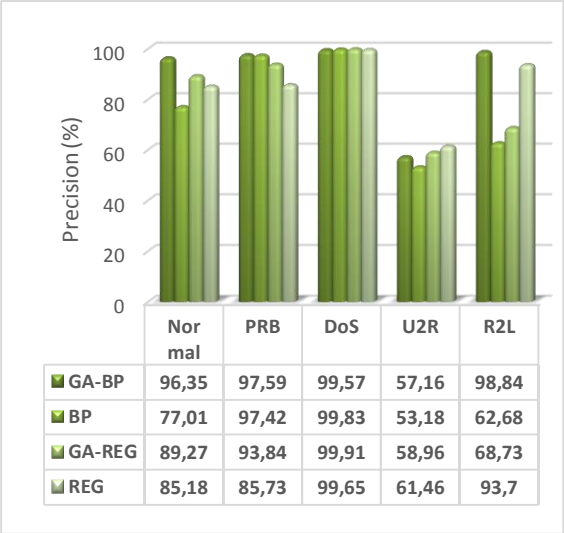

Figure 1 :Compare measure of precision for each class of attack for different methods

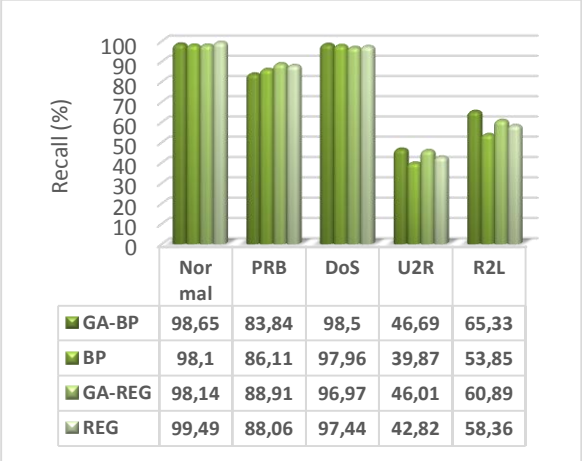

Figure 2 :Compare measure of recall for each class of attack for different methods

6 Conclusions

In this study, genetic algorithm was used as a technique of feature selection in combination with the neural network for improving its efficient for the intrusion detection problem. Relying on the natural nature of revolutionary theory and discovering dependencies between different data can result in reducing duplicate data in the dataset checked by the intrusion detection system. Likewise, NSK-KDD dataset was used for training and testing the suggested model.

Decreasing computation costs, computer resources such as memory and processing time are among advantages of the model. Similarly, with feature selection in the suggested model, both precision and recall values of the IDS are improved. The results showed that the suggested method enjoys a considerable performance in intrusion detection and is able to perform more precise detections in contrast to the basic models introduced in this study.

References

Dimitrakakis, C. and Mitrokotsa, A.(2013).Intrusion detection in MANET using classification algorithms: The effects of cost and model selection. Ad Hoc Networks, 11, pp. 226–237.

Singh, J.,Dutta, P. , Pal, A.(2012).Delay prediction in mobile Ad Hoc network using artificial neural network. Procedia Technology, 4, pp. 201-206.

Wu, S. X. , Banzhaf, W. (2012).The use of computational intelligence in intrusion detection systems: A review. Applied Soft Computing, 10, pp. 1–35.

Bhutan, M.H., Bhattacharyya, D.K. , Kalita, J.K. (2014).Network Anomaly Detection: Methods, Systems and Tools. IEEE Communications Surveys & Tutorials, 1(16), pp.303-336.

X. Tong, X. , Wang, Z. , Yu, H. (2009).A research using hybrid RBF/Elman neural networks for intrusion detection system secure model.

Recebido: dia/mês/ano Aceito: dia/mês/ano

Computer Physics Communications, 180, pp. 1795–1801.

Sheikhan, M. , Jadidi ,Z. , Farrokhi, A. (2010) .Intrusion detection using reduced-size RNN based on feature grouping.Neural Comput & Applic.

Moradi , Z. , Teshnehlab, M. , Rahimi, A. (2011) .Implimantaion of Neural Networks for Intrusion Detection in MANET. International Conference on Emerging Trends in Electrical and Computer Technology (ICETECT), India.

Wang, G. et al. (2010) .A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering. Expert Systems with Applications, 37, pp. 6225–6232.

Andonie, R. (2010).Extreme Data Mining: Inference from Small Datasets. International Journal of Computers Comminications & Control, 5, pp. 280-291.

Huang, C. , Moraga, C. (2004).A Diffusion-Neural-Network for Learning from Small Samples. International Journal of Approximate Reasoning, 35, pp. 137-161.

Tsai , T. I. , Li, D. C. (2008).Approximzte Modeling for High Order Non-Linear Functions Using Small Sample Sets.Expert System with Applications, 34, pp. 564-569.

Li , D. C. , Liu, C. W. (2009).A Neural Network Weight Determination Model Designed Uniquely for Small Data Set Learning. Expert System with Applications, 36, pp. 9853-9858.

Chebrolu, S., Abraham, A. , Thomas, J. P. (2009).Feature deduction and ensemble design of intrusion detection systems. Computers & Security, 24, 2005, pp. 295-307.

Nsl-kdd data set for network-based intrusion detection systems, Available on: http://nsl.cs.unb.ca/NSL-KDD/, March 2009.

Tavallaee, M. et al.(2009).A Detailed Analysis of the KDD CUP 99 Data Set. in Proceeding of the 2009 IEEE symposium on computational Intelligence in security and defense application (CISDA).

Stolfo, S. J. et al.(2000).Cost-sensitive modeling for fraud and intrusion detection: Results from the JAM project. in Proceedings of the 2000 DARPA Information Survivability Conference and Exposition.

Lippmann, R. et al.(2000).The 1999 DARPA off-line intrusion detection evaluation. Computer Networks, 34, pp.579-595.

Recebido: dia/mês/ano Aceito: dia/mês/ano