Journal of Engineering and Applied Sciences 11 (4): 810-815, 2016
ISSN: 1816-949X

A New Method for Detecting Network Intrusion by Using a Combination of Genetic Algorithm and Support Vector Machine Classifier

Behrooz Mabadi Jahromy, Ali Reza Honarvar, Mojtaba Saif and Mohammad Ali Mabadi Jahromy
Department of Computer Engineering and Information Technology, Islamic Azad University,
Safashahr Branch, Safashahr, Iran

Abstract: The purpose of intrusion detection is to identify an unauthorized use, misuse and damage to computer systems and networks by either of two internal users and external attackers. In this study, we have presented a new approach based on machine learning techniques to identify malicious attacks and provide security at an accessible level for users. The introduced method uses a genetic algorithm with a statistical target function based on the data distribution to select the features and the support vector machine for classification. The results of the simulation proposed good quality of the indicative method.

Key words: Computer networks, intrusion detection, machine learning, genetic algorithms, support vector machine

INTRODUCTION

With the advancement of information technology, the need has been created to perform computing tasks everywhere and all the time. It also requires that people are able to perform their heavy computing tasks without expensive hardware and software through services. Modern networks provide the productivity and conserving in IT resources and increasing computing power, so that the processing power becomes a tool with a perpetual access (Mukhopadhyay *et al*., 2011). Although, the network has many benefits but security in the network is very important. Providing computer network security is of great value today. With the advent of advanced computer and technologies and internet services, computer network security is more important. There are various tools to provide a secure computer network.

One of the tools of network security, along with other tools has the task of intrusion detection is intrusion detection systems. The task of these systems is to automate the process of intrusion detection (Mar *et al*., 2014).

Intrusion refers to unlawful actions that risk the integrity and confidentiality or access to a resource. The intrusion can be divided into two internal and external. External intrusions are those intrusions that are done by authorized or unauthorized individuals from outside the network to the internal network and authorized personnel do internal intrusions in systems and the internal network from within the network. Intruders, generally, use the software defects, breaking passwords, eavesdropping on network traffic and weaknesses of network design, network services and computers, to intrude the systems and computer networks (Mukhopadhyay *et al*., 2011).

In order to counter the intruders to computer systems and networks, many ways as methods have been created entitled the intrusion detection methods that have the practice of monitoring events occurring in a computer system or network. The use of intrusion detection systems is very common for securing computer networks. However, in recent years, intrusion detection systems are more developed and are widely used in securing computer networks, still far from the ideal and still there is a very important subject, in respect of it (Rupali and Bhupendra, 2010). Intrusion detection systems are based on network against high traffic over networks today and may become the bottleneck. This limitation makes it, in addition to loss of network security; network performance under cover is also lowered, so that in high traffic, intrusion detection systems have also a large limitation computational capacity. Due to these problems, in recent years, machine learning as one of intelligent techniques are widely used in intrusion detection. Machine learning is a relatively new research field of artificial intelligence that at present, it spends its growth and development era and is a very active field of computer science. Limitations and problems in the infrastructure lead the researchers to the use and application of machine learning techniques to solve problems and create the opportunity that the machine

Corresponding Author: Behrooz Mabadi Jahromy, Department of Computer Engineering and Information Technology, Islamic Azad University, Safashahr Branch, Safashahr, Iran

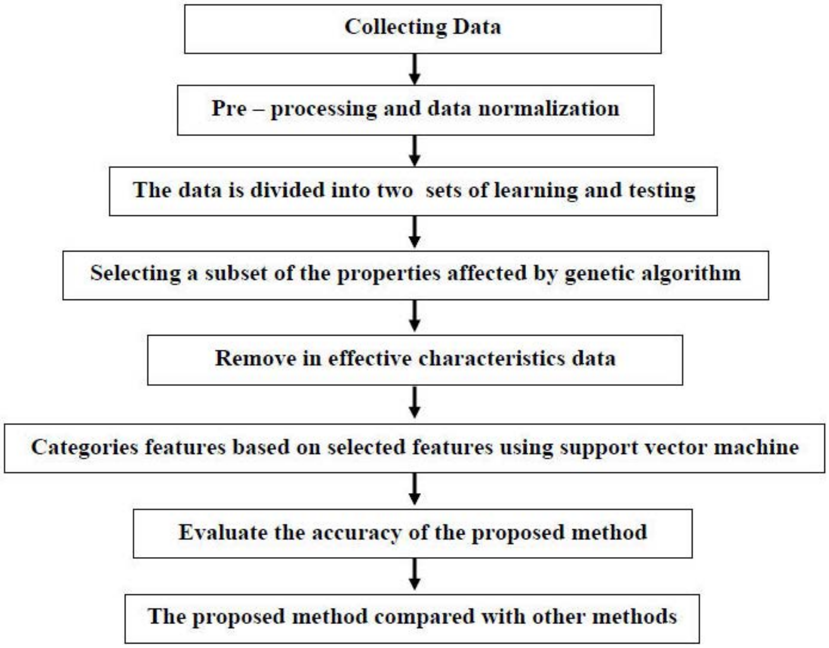
Collecting Data
Pre – processing and data normalization
The data is divided into two sets of learning and testing
Selecting a subset of the properties affected by genetic algorithm
Remove in effective characteristics data
Categories features based on selected features using support vector machine
Evaluate the accuracy of the proposed method
The proposed method compared with other methods

Fig. 1: Flowchart of the proposed method

learning has an important help and contribution in the field of intrusion detection systems (Buyya *et al*., 2013). Machine learning methods to develop an intrusion detection system in the network environment and enhance the security level, should consider the features of the environment that affect the development process. Good performance, scalability, user and service diversity, dynamism use of services are including aspects that should be considered. Since, the network environment has a distributed architecture, it is vulnerable and prone to attack and intrude. Given the complexity of the network intrusion detection process and on the other hand, the efficiency of machine learning techniques, this research plans to integrate the techniques of genetic algorithms to select effective features in the detection and identification of attacks as well as the support vector machine for classification the attacks and to introduce an efficient way to detect attacks in the network. The proposed approach uses the support vector machines and genetic to detect attacks. For test and evaluation, complex data of NSL_KDD have been used.

The framework of the proposed method: In this study, a new method is introduced for automatic detection of security threats based on machine learning algorithms that is able to explore data and then to determine a model for intrusion detection. To do this, the data NSL_KDD are used. After obtaining the required data, the preprocessing of data will be done to standardize the data and remove outliers. After pre-processing operations to increase the accuracy of the classifier and reduce the data dimension, we must use some feature selection algorithms to calculate the importance, influence of each attribute and non-important properties should be removed from the data set. In this study, a combination of genetic algorithms, to select the most important features and then the support vector machines to classify attacks based on optional features are used. Expressed flowchart in Fig. 1 shows the proposed method well.

As stated flowchart is well understood, after data collection and normalization and pre-processing operations, the data are divided into two parts of 70 and 30%, respectively to learn and then test the proposed method. The proposed method steps begin of data reduction by genetic algorithm. After this step, the number of basic features reduces. This increases the accuracy and speed of next steps, meaning the classification dramatically. In fact, after feature selection, classification attacks carried out by SVM. After learning classifier, the test data are given to the machine is to examine the quality classification based on test data.

Database: To detect attacks, data NSL_KDD are used. In this database, for each connection of TCP/IP, 41 quantitative and qualitative features have been extracted and all types of attacks are categorized into four main groups as follows (Che and Ji, 2012).

Denial of (DOS) service is a type of attack in which an attacker makes some computational resources or memory to handle authorization requests, so busy or fill up that accordingly, deprives authorized users from accessing the machine.

Unauthorized access by a remote machine (R2L) is a class of attacks in which an attacker has started its work with accessing to a normal user account on the system and is able to take advantage of vulnerable and in this way, root accesses to the system.

Unauthorized access to the main local user privileges (U2R) is a class of attacks in which an attacker sends small packets through the network to a machine. However, this attacker has not an account on that machine and thereby he takes advantage of some vulnerability so that he could access the machine as a user.

Searching and exploring (probe) is a class of attacks in which the attacker scans a network of computers to collect information or to find known vulnerabilities. An attacker can use this information to search for exploits with a map of the machines and services that are available over a network.

Selecting features based on genetic algorithms: For feature selection, a genetic algorithm is used to select the most important features. The goal is to split the existing properties to two important features which are effective on reconnaissance attacks and are ineffective properties in identifying attacks. Genetic algorithms are searching heuristics and optimization algorithms that run in parallel and are inspired by Darwin's principle of natural selection and genetic amplification (Guo *et al*., 2010). In other words, these algorithms are optimization techniques, based on the selection and recombination of promising solutions. In the traditional genetic algorithms, solutions are displayed as binary strings of 0 and 1.

The initial population of the genetic algorithm includes a series of binary chromosome which are 41 bits; each bit representing each of the available properties. If a bit of a feature in network access data has the value 1, it means that feature plays a vital role in the network attack detection and 0 means that it is unimportant and is ineffective in identifying patterns of attack detection. After making the initial population, by using the parameters available in the chromosomes, each chromosome performance is calculated by an evaluation function in the genetic algorithm. Our main goal is to find optimal or near-optimal parameters that produce the most accurate solution. The proposed evaluation function in this paper is a statistical method which is expressed in the following section. This means that the proposed feature selection method is a filter-based approach and does not use the classifier at every stage to evaluate the selected features that it increases the feature selection speed (Chandola *et al*., 2012). In the final step, by using genetic operators, a new generation is produced. According to the evaluation function for each chromosome, chromosomes with a higher value will be selected and produce new population, the crossover and mutation operators are used.

The objective function of the proposed feature selection method: The method of feature selection by using genetic algorithm by using input data and by a measure of independence and a new separability has focused on the reduction of network intrusion data and acquiring accuracy of classification. An efficient method of selecting a subset based on genetic algorithms for data network is proposed. Since, the distribution of the between-class scattering matrix to within category scattering of subsets can offer their participation in the categories. A sub-optimal characteristic is selected in the categorization process in terms of the independence of features. A measure called the separability Score is introduced that the selected features evaluation is calculated based on these criteria.

The separability score: The purpose of the separability score is selecting the optimum features for classification. Imagine that $(x, y)\in(R_d\times Y)$ is a sample where R_d is a d-dimensional feature space and $Y = \{1, 2, ..., c\}$ is set of category labels. n_i represents the number of samples that belong to the ith class and N present the total number of samples. Think that x_{ij} implies on the j sample in i class:

$$u = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{n_i} x_{ij} \qquad (1)$$

$$u_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \qquad (2)$$

Therefore, the between-class scattering matrix (S_b) (between-class scatter matrix) and among-class scattering matrix (S_w) (within-class scatter matrix) are defined as follows:

$$S_b = \sum_{i=1}^{c} n_i (u_i - u)(u_i - u)^T \qquad (3)$$

$$S_w = \sum_{i=1}^{c}\sum_{j=1}^{n_i} (u_i - x_{ij})(u_i - x_{ij})^T \qquad (4)$$

Here, S_b measures the intervals between the average vector of each category and the overall average while S_w

measures the categories mean distribution around their average of the vectors. For a subset of desired features, separability of category is Scatter-Matrix-based class separability, the ratio of the trace or determinant of the between-class scattering matrix with the in-class scattering matrix. The separability of S is as follows:

$$S = \frac{S_b}{S_w} \tag{5}$$

A subset with a large S is considered as a well-regarded subset and means small within-class scattering and big between-class scattering. Therefore, a large S ensures that the categories are scattered well by means of distribution. This is a simple, powerful and integrated criterion to categorize. In order to select an optimal subset, class separability idea is used because the contributions of these subsets reflect this classification.

Classification by using support vector machine: After specifying the features, it turns to categorizing the data by using Support Vector Machine (SVM). It is safe to say, SVM algorithms are the finest and most powerful machine learning algorithms (Bankovic *et al*., 2007; Muda *et al*., 2011). This new method can be used for linear and nonlinear data classification (Sangkatsanee *et al*., 2011). In recent years, due to good results, this algorithm is turned to a common technique used to classify. With a suitable nonlinear mapping, SVM algorithm will convert the training data space to a higher dimension and then in this new dimension, it looks for a hyper-plane that separate the samples of a class from other classes. With a suitable nonlinear mapping, data collection of two classes can be separated by a hyper-plane. In this way, each class is taught against other classes. Data related to the class will be labeled +1 and data from other classes will be labeled -1. If the number of classes is N, N support vector machines are trained that each one corresponds to one of the classes. After teaching classes in the test phase, each of test samples applies to the N support vector machine, the winner class is the class that it SVM has the highest output (Mukhopadhyay *et al*., 2011; Salah *et al*., 2013).

Evaluation criteria of classifier accuracy: To compare the accuracy of the classifier, based on the proposed method, the two criteria of precision and recall are used:

$$\text{Precision} = \frac{\text{(Number of True Positive (TP))}}{\text{(True Positive (TP)+False Positive (FP))}}$$

$$\text{Recall} = \frac{\text{True Positive (TP)}}{\text{(True Positive (TP)+False Negative (FN))}}$$

RESULTS AND DISCUSSION

The results of testing and evaluation: For implementing the proposed method, a computer device with corei7 processor with 4 GB of main memory has been run on the Windows 7 operating system. An introduced genetic algorithm for feature selection has been stated in the previous section. To simulate this method, parameters must be set in the algorithm that these parameters are expressed in Table 1.

To obtain accuracy of the proposed classifier in this study, the hold-out method of validation is used. Hold-out validation is a method of assessment that specifies how much the results of a statistical analysis is applicable and independent of the training data on a data set (Song *et al*., 2013). This technique in particular will be used in forecasting application to identify the extent of the model will be useful in practice (Altwaijry, 2013). In general, a round of cross validation includes data segmentation into two subsets of the supplement, analysis on a subset (training or learning data) and validation analysis by using data from another set (validation data or test). In this study, 70% of data are used for training and the remaining 30% are used for the test.

Genetic algorithm extraction feature: In this study, for feature selection, genetic algorithm with a statistical function based on a standard separability is used to assess the competencies of sub-features. This method does not use the feedback from classification algorithm and selects features based on the data distribution among the members of a class and data distribution among different categories. Bit number 0 in the final chromosome of genetic algorithm means the lack of choice and the number of bits 1 means choosing a specific property corresponding to that bit. Not selected columns have little effect on data classes' determination and will be deleted from the data collection. Genetic algorithm used in this study has been implemented in the NSL_KDD data. The

Table 1: Genetic algorithm parameter settings and stop conditions

Parameters	Size
Population	100
Maximum population	100
Genes per chromosome	(the total number of properties available) 41
Elite	2
Probability of crossover	0.6
The possibility of mutation	0.2
Criteria for stop	Reach the number of generation 200 or lack of progress for 25 consecutive repetitions

Table 2: The classifier accuracy for different categories of attacks

	Percentage				
Features type/category	Normal	Probe	Dos	U2R	R2l
Features selected by genetic algorithm	98.4	98.8	97.8	97.4	99.1
Features unselected by genetic algorithm	88.5	82.4	87.6	90.6	93.4

Table 3: Evaluation results of support vector machine accuracy

	Percentage				
Features type/category	Normal	Probe	Dos	U2R	R2l
Precision	98.1	98.2	97.4	98.2	98.9
Recall	98.6	97.9	97.9	97.1	98.8

total number of initial columns in the database had 41 features that 28 features had bit 1 in the final chromosome and the remaining of 13 features had bit 0. This means that 28 features were diagnosed as important the features and the remainder of features were omitted and classification operation will be done with these 28 features.

Assessment of the proposed feature selection method: For an accurate evaluation of the proposed feature selection method by using genetic algorithms, we should calculate and evaluate the classifier performance based on the features selected, individually for all classes. These assessments that are introduced for four categories of attack in the previous section and for two sets of feature of the GA (two sets of important and not important features) are examined and are stated in Table 2.

As it is clear from the results presented in Table 2, the proposed method has excellent accuracy for all different categories. When unselected features are used for the classifiers, a little precision will be achieved that this fact indicates the high quality of genetic algorithms for feature selection presented in the previous chapter.

Evaluation of SVM accuracy to classify attacks: To evaluate the accuracy of the SVM classifier based on features selected by genetic algorithm, the criteria of percision and recall are used. These measures are calculated separately for all different categories of attacks that Table 3 shows the results.

Compare the proposed method with other methods: There are methods that will make their operations in the field of intrusion detection systems in the network. The proposed methods by Memon and Chandel (2014) and Li *et al.* (2012) are the best methods available in the field. The accuracy of the proposed method with a combination of genetic algorithm and support vector machine for each category with these methods have been compared that Table 4 illustrates this well.

Table 4: Compare the proposed method accuracy with other methods

	Percentage				
Method/category	Normal	Probe	Dos	U2R	R2l
Introduced method	97.1	96.2	95.9	95.6	97.8
The method presented by Mukhopadhyay *et al.* (2011)	96.4	93.2	95.6	94.1	96.7
The method presented by Li *et al.* (2012)	96.8	91.2	94.4	91.0	94.9

As well, it is clear from the results in Table 4, the proposed method, compared with the previous two methods has greater or equal accuracy that this shows the high quality of classifiers presented in this thesis.

CONCLUSION

In this study, a new approach to create security and identify malicious attacks on the network based on a combination of genetic algorithm for feature selection and support vector machine for data classification have expressed. The introduced method uses genetic algorithms and data distribution criteria for selecting features and support vector machine for classification that the results of each were evaluated. Evaluation results indicate high precision and good quality of the data classification method proposed in each category and also in comparison with other methods are available. The results show that the method introduced seems to be very efficient in comparison with other existing methods in the field of intrusion detection.

REFERENCES

Altwaijry, H., 2013. Bayesian Based Intrusion Detection System. In: IAENG Transactions on Engineering Technologies, Kim, H.K., S.I. Ao and B.B. Rieger (Eds.). Springer, New York, USA., ISBN: 9789-400747869, pp: 29-44.

Bankovic, Z., D. Stepanovic, S. Bojanic and O. Nieto-Taladriz, 2007. Improving network security using genetic algorithm approach. Comput. Electr. Eng., 33: 438-451.

Buyya, R., J. Giddy and D. Abramson, 2013. An evaluation of economy-based resource trading and scheduling on computational power grids for parameter sweep applications. Proceedings of the 2nd International Annual Workshop on Active Middleware Services, October 4, 2013, Edinburgh.

Chandola, V., A. Banerjee and V. Kumar, 2012. Anomaly detection for discrete sequences: A survey. IEEE Trans. Knowledge Data Eng., 24: 823-839.

Che, Z. and X. Ji, 2012. An efficient intrusion detection approach based on hidden markov model and rough set. Proceedings of the International Conference on Machine Vision and Human-Machine Interface, April 24-25, 2010, Kaifeng, China, pp: 476-479.

Guo, Y., B. Wang, X. Zhao, X. Xie, L. Lin and Q. Zhou, 2010. Feature selection based on rough set and modified genetic algorithm for intrusion detection. Proceedings of the 5th International Conference on Computer Science and Education (ICCSE), August 24-27, 2010, Hefei, Anhui, P.R. China, Pp: 1441-1446.

Li, Y., J. Xia, S. Zhang, J. Yan, X. Ai and K. Dai, 2012. An efficient intrusion detection system based on support vector machines and gradually feature removal method. Expert Syst. Appl., 39: 424-430.

Mar, J., Y.C. Yeh and I.F. Hsiao, 2014. An ANFIS-IDS against deauthentication DOS attacks for a WLAN. Proceedings of the International Symposium on Information Theory and its Applications, October 17-20, 2010, Taichung, pp: 548-553.

Memon, V.I. and G.S. Chandel, 2014. A design and implementation of new hybrid system for anomaly intrusion detection system to improve efficiency. Int. J. Eng. Res. Applic., 4: 1-7.

Muda, Z., W. Yassin, M.N. Sulaiman and N.I. Udzir, 2011. Intrusion detection based on K-means clustering and Naive bayes classification. Proceedings of the 7th International Conference on Information Technology in Asia, July 12-13, 2011, Kuching, Sarawak, pp: 1-6.

Mukhopadhyay, I., M. Chakraborty, S. Chakrabarti and T. Chatterjee, 2011. Back propagation neural network approach to intrusion detection system. Proceedings of the 2011 International Conference on Recent Trends in Information Systems, December 21-23, 2011, Kolkata, pp: 303-308.

Rupali, D. and V. Bhupendra, 2010. Feature reduction for intrusion detection using linear discriminant analysis. Int. J. Comput. Sci. Eng., 2: 1072-1078.

Salah, S., G. Macia-Fernandez and J.E. Diaz-Verdejo, 2013. A model-based survey of alert correlation techniques. Comput. Networks, 57: 1289-1317.

Sangkatsanee, P., N. Wattanapongsakorn and C. Charnsripinyo, 2011. A practical network-based intrusion detection and prevention system. Comput. Commun., 34: 2227-2235.

Song, J., H. Takakura, Y. Okabe and K. Nakao, 2013. Toward a more practical unsupervised anomaly detection system. Inf. Sci., 231: 4-14.