Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 7, July 2017, pg.178 – 188

Providing an Improved Intrusion Detection System using Graph Clustering and Genetic Algorithm

AmirReza Rezvani Habib Abadi
Dr. Mahdi Mollamotalebi
Msc, Department of Computer , College of Engineering, buin Zahra Branch, Islamic Azad University, Qazvin, Iran
Assistant Professor, Islamic Azad University, Qazvin, Iran
motalebi.academic@gmail.com

Abstract
One of the methods for network security is the use of intrusion detection systems. These systems send the relevant reports to the management sector by monitoring the network in order to detect malicious activities or violate security policies. They warn the system administrator, if they encounter abnormal data in the network, to take the necessary steps to prevent the attack. One of the problems with the implementation of intrusion detection systems is the high level of information and the high number of features of each attack. As a result, the performance of the intrusion detection system dramatically decreases. High-dimensional data sets from two directions reduce the performance of the intrusion detection system. On one hand, the size of the data increases with increasing the dimensions of the data, and on the other hand, a model based on high-dimensional data has a lower generalization capability. A way to cope with this problem is to reduce size of problem through selection of feature. By reduced size of problem, the computational complexity of the intrusion detection system is reduced and the performance of the intrusion detection algorithm improves in terms of diagnostic accuracy. In this paper, to select appropriate

intrusion detection features, a clustering-based graph approach is proposed. In this method, the primary features are divided into a number of clusters, and then appropriate features are chosen of each cluster using evolutionary search techniques. The proposed method was tested with the KDDCUP99 dataset and under the MATLAB software environment. Results from simulation and comparing them with the same and recent methods indicate suitable performance of the proposed method in intrusion detection in terms of intrusion detection accuracy and reduction of computational complexity.

Keywords- intrusion detection system, security of computer network, genetic algorithm, graph clustering, detection accuracy, computational complexity

Introduction

Intrusion detection systems are responsible for identifying and detecting any unauthorized use of the system, misuse or harm by both domestic and foreign users. Detection and prevention of infiltration today is considered as one of the main mechanisms for security of networks and computer systems. Intrusion detection systems are commonly used alongside firewalls and are complementary to them. To make security in a computer system, intrusion detection systems are required in addition to firewalls so as to detect it if the attacker crosses the firewall, anti-virus, and other security measures and enters into system. Intrusion or harassment refers to an activity that is carried out from the outside or inside the network and aims to violate one of the security aspects of the network, that is, confidentiality, integrity, or availability. An accepted definition of intrusion is a set of related activities that aim at unauthorized access to information, change in information, or damage to the network in such a way as to disrupt the network's operation. In this definition, the activity involves successful and unsuccessful efforts [1]. Intrusions are divided into internal and external divisions. External intrusions are those intrusions that are committed by authorized or unauthorized persons from outside the network to inside the network, and internal intrusions are transmitted through the members of the organization's network [2]. Infiltrators generally benefit from software vulnerabilities, password breaks, eavesdropping, network traffic, and design weaknesses in the network or services, for Intrusion into organizations and computer networks. An Intrusion Detection System is a set of tools and methods that are responsible for monitoring, detecting, and reporting unauthorized activities to the network administrator.

Accordingly, intrusion detection systems are one of the most important security tools used in networks and complement other security technologies such as firewalls, antivirus and honeypots. Two important applications of Intrusion detection systems are: i) collect information about interference and provide it to the network administrator so that attacks are foreseen and prevented, (ii) The recording of interruptions in the network, so that this information provides the organizations to detect and track down the attackers. Since technically, the creation of computer systems (hardware and software) is virtually impossible without security weaknesses, the detection of intrusion in the research of computer systems is followed by a special sensitivity. The purpose of an intrusion detection system is not to prevent attacks, but to detect and, if possible, identify types of attacks, as well as to detect security problems in a system or computer network and notify it to the system administrator. Therefore, intrusion detection systems are commonly used alongside fire walls and as

complementary to them [3]. The steps to create an intrusion detection system include collecting data, preprocessing data, intrusion detection, and reporting to the network administrator. In the meantime, detection of intrusion is the most crucial stage. After Dening introduced a model for intrusion detection in 1987, many studies focused on increasing the efficiency and accuracy in intrusion detection systems. In the early 1990s, the use of statistical approaches and knowledge of security experts was common. Since the late 1990s, the diagnosis of normal and abnormal behavior was changed from manual to automatic, and artificial intelligence and machine learning techniques were used to discover models from the contents of educational data [4].

One of the problems with the implementation of intrusion detection systems is the high number of parameters associated with the data set, as well as the number of features of each attack.

The existence of a large number of unrelated and redundant features in the data set negatively affects the performance of the machine learning algorithm and also increases the complexity of time [5]. Therefore, reducing the size of the data set is one of the basic tasks in data mining and machine learning applications. Reducing the dimensions of the data set reduces the complexity of the time, and, on the other hand, reduces the parameters of the intrusion detection algorithm. As a result, performance of intrusion detection algorithm improves. In other words, a model which is made based on reduced features has extended generalizations compared to the original model. In recent years, two general strategies (feature selection and feature extraction [6,7]) have been proposed to reduce the dimension.

A feature selection tool that is also known as variable selection and subset selection, by searching among available sub-sets, selects a subset of the primary features, while in the feature extraction strategy, the original features are transmitted to a new space with less dimension. In the proposed method in this paper, the initial features are first represented as a weighted graph. The nodes of this graph, features, and edges show the similarity between the features. Then, the graph features are divided into a number of clusters, and using evolutionary search capabilities, the final features are selected for intrusion detection.

Related works

In recent decades, many intrusion detection systems have been developed to detect abnormalities. In general, intrusion detection systems are based on two different approaches. I) signature-based diagnosis; ii) anomaly-based diagnosis.

Detection of signature-based intrusion compares the patterns of behavior in the network with previously observed samples. This method is very effective against the prevalence of attacks, in order to determine the possibility of attacks. However, due to various types of attacks and behavioral patterns that can be exploited by the attackers, its efficiency is limited. This approach is the simplest method of detecting attacks to computer network because only the current activities that are in progress are examined. The last activity report is compared with a list of existing patterns, and this comparison is done using the comparison methods of the spellings [8]. The anomaly-based detection approach compares the observed conditions with the normal conditions of the system in order to detect the serious differences that usually occur in the event of an attack. Systems that operate on this approach have documented history that describes the state of the various components of the system in normal conditions.

The main advantage of abnormal-based diagnostic methods is that they can, with the smallest cost, detect different and unknown types of attacks whose patterns have not been previously detected. The history of the system used in a training phase, which may last for days or weeks, is recorded and reviewed. The problem with this approach is that due to the complexity and variety of different behaviors that may occur on a network, the creation of a history requires a high degree of accuracy. In addition, the exact diagnosis of the cause of anomalies is not possible. Due to the large dimension of network data, many intrusion detection systems have used feature selection. This is to find important features that have a greater impact on the classing, and designing the system is based on them, so that other irrelevant or overriding features can be ignored.

Research has shown that the use of the feature selection phase increases the accuracy of the intrusion detection system [8]. Cho and his colleagues in 2003 presented a model that uses fuzzy logic and Markov's model for detection. In this method, the Markov model was used to reduce the size. In 2004, Fleuret [10] made a feature selection based on conditional common information, which resulted in a simple and fast algorithm.

Chebrolu et al. (2005) determined the important features that influenced the development of the intrusion detection system, using the Markov algorithm and the decision tree; they also used the Bayesian network and the regression tree to develop Intrusion Detection System[11]. Further, Amiri et al. [12] proposed a new method for intrusion detection using the cross-sectional feature selection method. In the proposed method, in this paper, the value of the features is measured using the cross-metric criterion and the inappropriate features are eliminated. Further, in 2015, Eesa et al. [8], using the fish optimization algorithm, selected features that were appropriate for intrusion detection, and the intrusion detection system was designed based on these features.

The proposed method

In this section, the proposed method is described to select feature in intrusion detection. Since a graph-based algorithm has been used for clustering features in the proposed method, the problem is represented graphically at the first stage of the proposed method. Then, on the basis of graphic representation of problem, the primary features are divided into several clusters. Further, at the third stage of this proposed method, search the subset of the final feature is described based on the data scattering criterion. Figure 1 displays the overview of the proposed method. As shown in this figure, the proposed method consists of three main stages of the graphic representation of the problem, the clustering of genes and the selection of appropriate genes from each cluster.

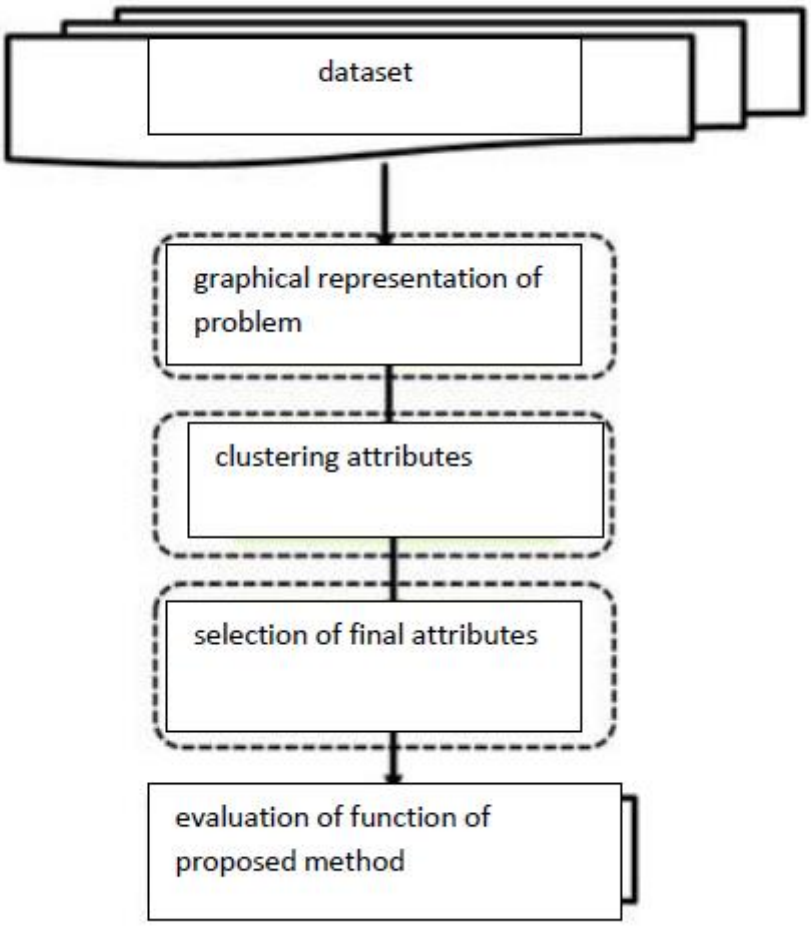

Fig 1. Flowchart of the proposed method

Analysis of performance of the proposed method

One of the main goals in the issue of selecting features is the removal of unrelated features with redundancy. It has been shown that any method of feature selection that cannot effectively remove these features is not a good method, and the set of features selected by that method is generally not the best subset of the feature [13]. The reason for this is mainly due to the unrelated features as well as the redundancy among the features, which increases the parameters of the classification algorithm and thus reduces the performance of the classification algorithm. Therefore, the researchers conclude that the selected features must have the most relation with the target class and have the least redundancy with each other. As a result, one of the goals of the first proposed method is to select a feature sub-set with maximum predictability of target class and minimum redundancy between its features. To achieve this, in search process of genetic algorithm, higher fit function is attributed to the chromosomes whose feature sub-set has higher classification accuracy and lesser redundancy. In other words, in the calculation of the fitness function, two criteria of communication and redundancy have been simultaneously influenced, caused The set of features with the highest predictability and the least redundancy has a greater chance of survival. In addition, the

clustering of features, in both sides, improves the performance of the genetic algorithm. On one hand, all clusters should be selected with a specific number of features, which will allow the subset of the selected feature represents all the initial features. On the other hand, the repair operator used in this proposed method, which modifies the produced chromosomes using features clustering and reduces the search space and direction to the genetic algorithm. Therefore, the use of this repair operator increases the ability of the algorithm to find the right answer and increases the speed of convergence of the genetic algorithm.

Practical results

To test the performance of the proposed method on various data sets, experiments have been carried out using the classifiers introduced in the previous section. In experiments conducted on the proposed method, data sets are randomly divided into educational data and experimental data. In doing so, 66% of dataset are considered as educational data and the rest are considered as experimental data. In each experiment, after the identification of the educational data and experimental data, each method of feature selection has been executed ten times, and averages of ten different implementations are used to compare different methods. In most of the previous articles and methods, in order to evaluate the intrusion detection performance, 10 times the implementation of the proposed method and the mean of accuracy have been used for performance evaluation. For this, 10 times implementation of the proposed method has been used in this research. The methods compared in this section are compared based on two criteria of number of selected features and classification accuracy.

Size of selected feature subset

In this section, different methods of intrusion detection based on feature selection are compared for the accuracy of intrusion detection on the data sets.

Four types of SVM, DT, NB, and KNN classifiers have been used to examine the capability of generalizing different feature selection methods. Table 1 displays the classification results for SVM, DT, NB, KNN classifiers.it should be noted that all the results of these tables have been obtained from average classification accuracy in 10 implementations for different methods.

Table 1. Comparison of intrusion detection accuracy of the proposed method with the method based on mutual information

	Support vector machine	Decision tree	**Simple Bayesian**	Closest neighbor
Mutual information[12]	96.57	94.13	95.63	96.36
Feature selection [8]	96.92	95.82	96.04	96.83
proposed method	97.14	96.19	96.78	97.03

As seen in this table, in all classifiers, the proposed method is superior to the other method and the accuracy of detection is higher. In figure 2, detection accuracy in the proposed method compared to article [8] and [12] has been displayed.

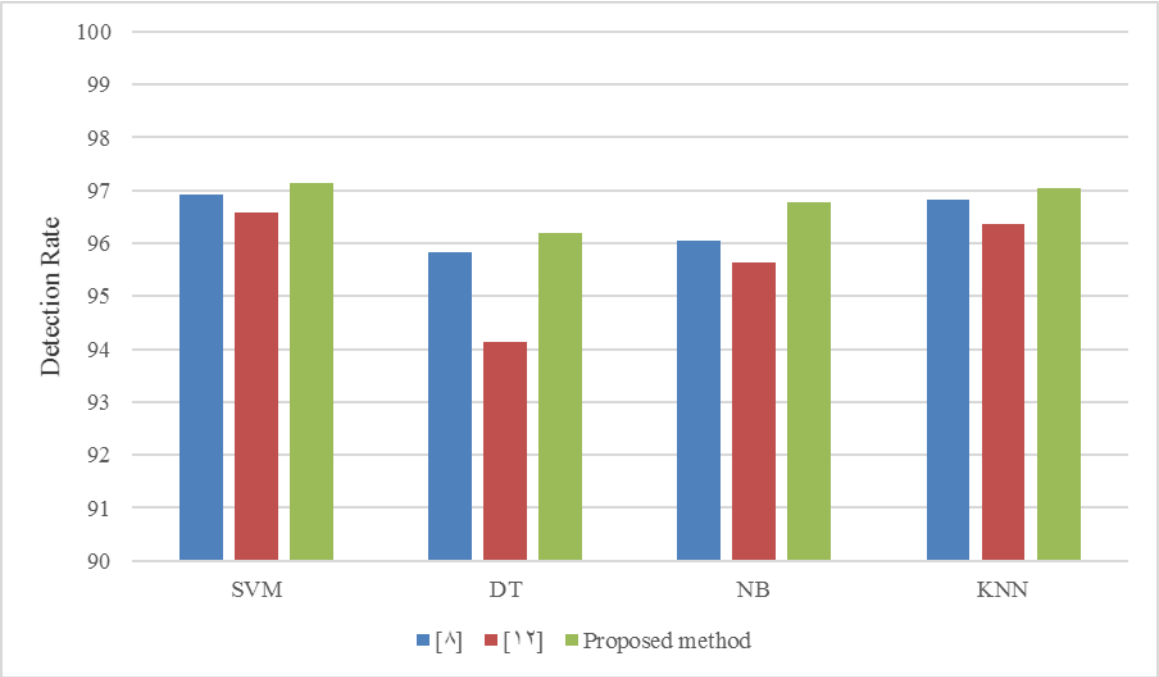

Fig 1. Evaluation of intrusion detection accuracy of proposed method on different classifiers

As shown in figure 1, in four different classifiers, the proposed method compared to two other methods has higher detection accuracy. For instance, accuracy of intrusion detection in the proposed method and on classifier SVM equals to 97.14, while detection accuracy for other two methods and on this classifier equal to 96.57 and 96.92. Rest three classifiers same as this classifier of the proposed method have superiority. On the other hand, study on data of this table indicates that the highest difference has been raised between the proposed method and other methods in classifier NB, which the reason is that classifier NB has associated to selection features, so that the least relationship between selection features will lead to higher detection accuracy. The proposed method in this paper will enable to select more suitable features and as a result intrusion detection accuracy on this classifier compared to other methods will be higher.

Selection features

Other experiments were designed to examine what method will have higher classification accuracy in case there are equal number of selected features. In doing so, in both intrusion detection methods, number of selected features has remained fixed and detection accuracy is calculated. Figures 2-5 display average classification accuracy for intrusion detection methods on classifiers SVM, DT, NB and KNN. In figures 2-5, horizontal axis displays number of selected features and vertical axis displays classification accuracy. Since the dataset used in this research has 41 primary features, number of features under study has been selected of 10 to 40 so as to examine increased or decreased trend of classification accuracy in change of number of features.

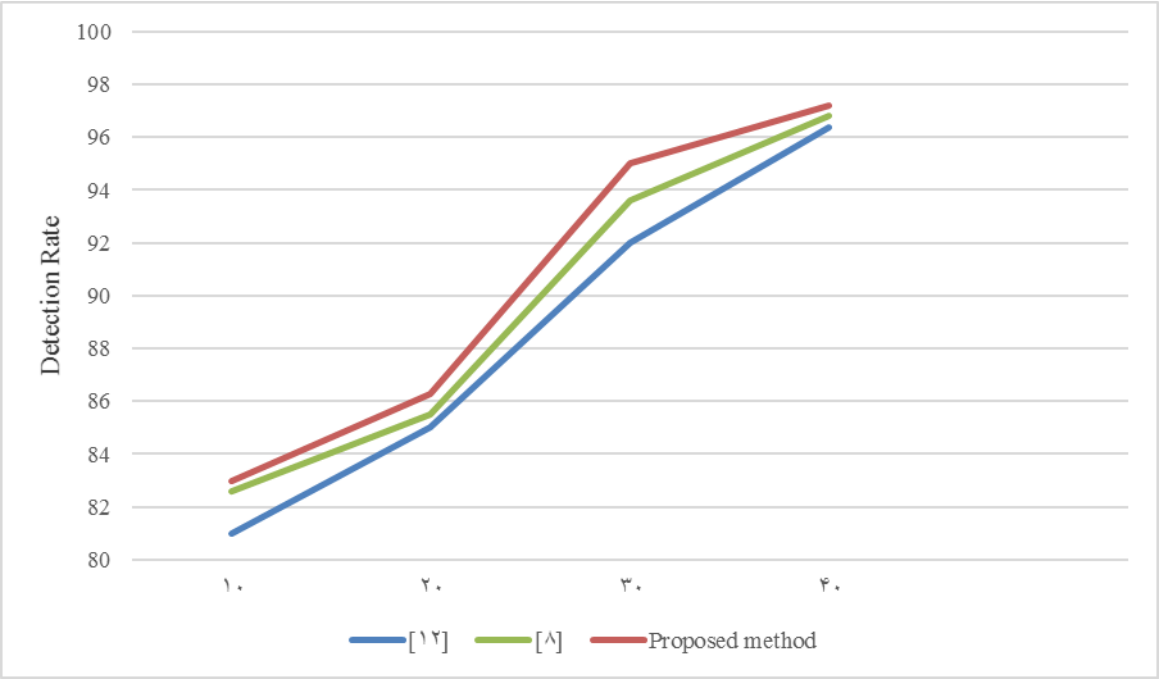

Fig 2. Detection accuracy for different methods on classifier SVM

As seen in this figure, the proposed method has higher accuracy than two other methods. For instance, when 20 features have been selected, intrusion detection accuracy has been 86.3% for the proposed method, while it is 85.5 and 85 for two methods in article [8] and [12]. Further, data of this figure display that when number of features increases from 10 to 30, change and rise in intrusion detection accuracy has been tangible, but it does not rises with more increase. In following, detection accuracy for classifiers DT, NB and KNN has been displayed.

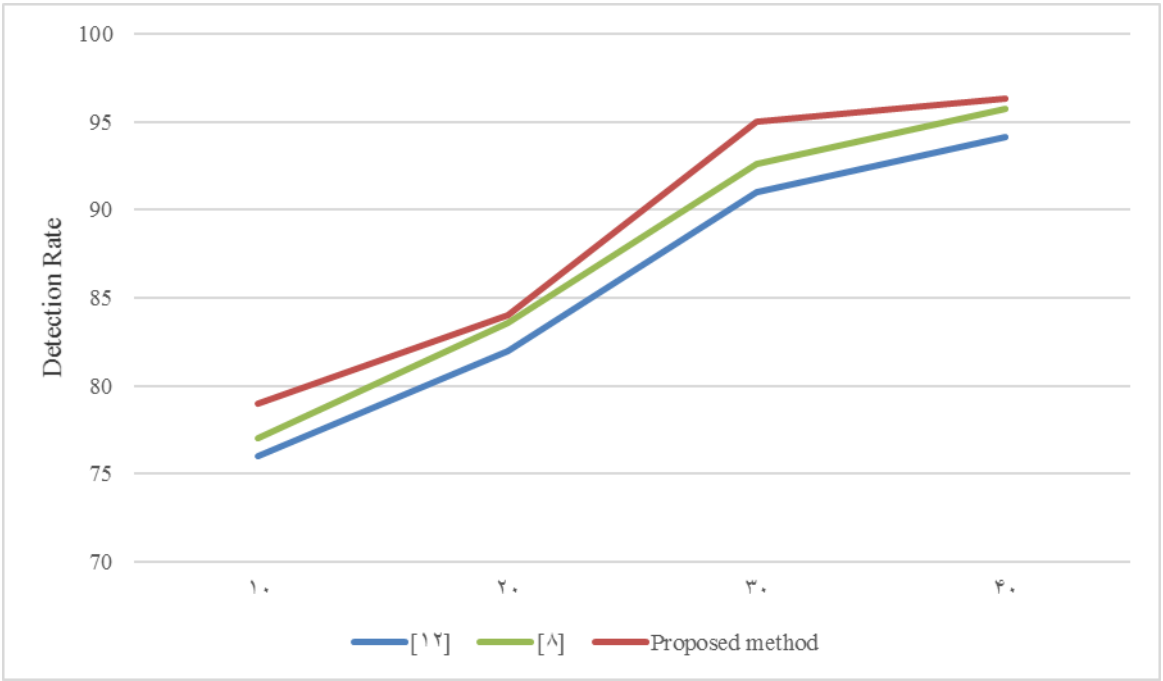

Fig 3. Detection accuracy for different methods on classifier DT

As shown in figure 3, when number of selected features has changed from 10 to 20, slope of diagram has not changed. The reason is that these features have not been suitable and sufficient for intrusion detection regarding number of primary features. Further, here it is observed that when number of selected features increases from 20 to 30, detection accuracy increases with sharper slope. The reason is that number of 30 features have been sufficient for intrusion detection and intrusion detection accuracy reaches to maximum with this number of features. On the other hand, number of features from 30 to above has no effect on increased accuracy.

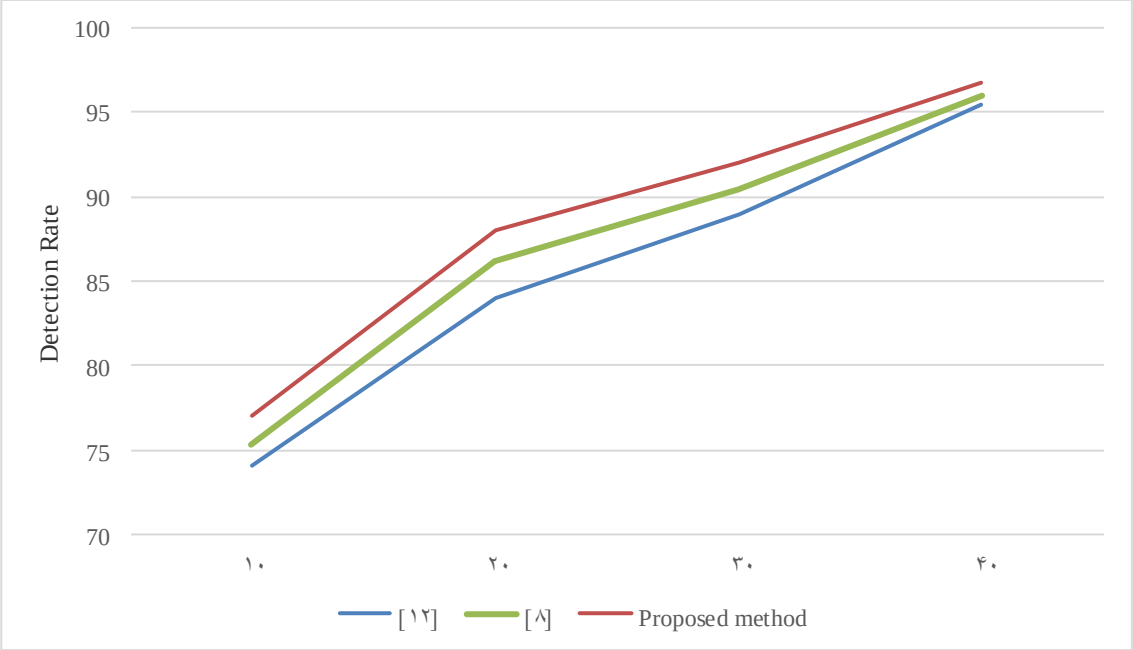

Fig 4. Detection accuracy for different methods on classifier NB

In figure 4, when number of selected features has increased from 10 to 20, intrusion detection accuracy has changed. This has occurred since the number of primary features have been 41 and 10 features have not been sufficient for intrusion detection. Further, when number of selected features has increased from 10 to 20 and 30, increased accuracy is seen, which the reason is that 30 features have been sufficient for intrusion detection; with this number of features, intrusion detection accuracy has increased to maximum. On the other hand, increased features from 30 to 40 have no effect on increased accuracy and 30 primary features are agent for total features; detection with the same features has been accurate to the same size. According to chart above, it can conclude that when classifier NB is used to detect intrusion, 30 features will be suitable; increased features from 30 to above will raise no change in intrusion detection.

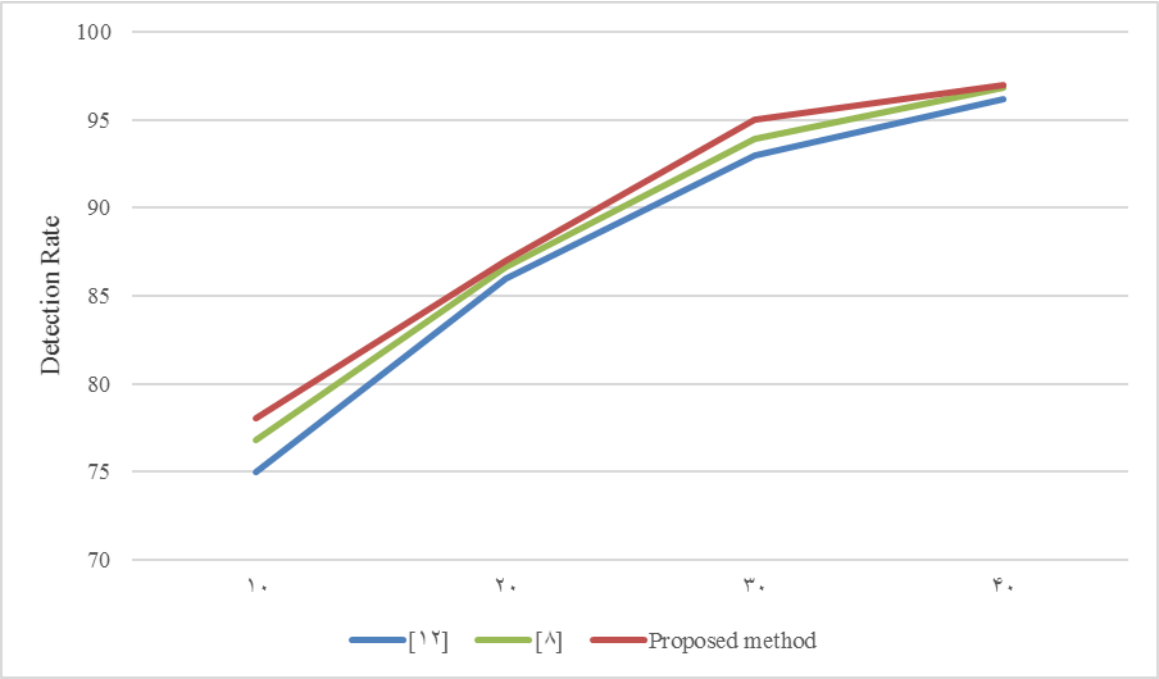

Fig 5. Detection accuracy for different methods on classifier KNN

As shown in figure 5, when number of features increases from 10 to 40, intrusion detection accuracy increases. When number of selected features changed from 10 to 20, slope of diagram has more increased than other states. The reason is that 10 features have not been sufficient regarding number of primary features, and when number of features has increased to 20, intrusion detection accuracy has increased. Further, it is observed that the detection accuracy increases when number of selected features increases from 20 to 30. The reason is that number of selected features has not reached to suitable size and these features do not represent primary features. On the other hand, increased features from 30 to 40 has no effect on increased accuracy, so that accuracy has not much increased. As a result, it can say that the best value for these features in intrusion detection using classifier KNN is 30 features. Further, according to figure 5, it is observed that there is no difference on accuracy in three methods of feature selection with 40 selected features. The reason is that since there are 41 primary features, no difference is observed between compared methods and three methods have the same functions. In figures 3-5, intrusion detection accuracy in proposed method is compared with other methods in terms of number of different selection features. As seen in these figures, the increased number of selection features results in increased accuracy of proposed method. For instance, in classifier KNN, when 10 features have been selected, the proposed method has had accuracy of 78%, while accuracy of intrusion detection is about 95% when there are 30 selected features. Further, comparison of 4 classifiers SVM, KNN, DT and NB indicates that classifier SVM with accuracy of 97.2% has highest intrusion detection accuracy. As a result, it can say that classifier SVM is more suitable for intrusion detection than rest classifiers.

Conclusion

Information systems and networks develop the most important parts of modern living which the current living is impossible without them. Disturbance in these networks leads to huge costs. Despite constant research, progress and security of computer networks have remained as an open problem, and the reason is complexity and interconnected nature of computer networks. One of problems in implementation of intrusion detection systems is high information and features at each attack. Large number of these irrelevant features in dataset puts a negative effect on performance of machine learning algorithm and increases computational complexity, thus reduced dataset size is a major task in data mining and machine learning applications. Reduced dataset size reduces computational complexity on one hand and reduces intrusion detection algorithm parameters on the other hand. As a result, function of intrusion detection algorithm increases. In other words, a model which is made based on reduced features has higher generalizability than the primary model. In recent, years, two general methods have been presented to reduce the size: feature selection and feature extraction. Feature select which is recognized with sub-set selection selects a sub-set of primary features by searching among existing sub-sets, while primary features are transmitted to a new space with less size in feature extraction.

References

[1]. Al-mamory, S.O. and F.S. Jassim.2015. On the designing of two grains levels network intrusion detection system. Karbala International Journal of Modern Science.p:1-11

[2]. Qin, T., et al.2015. Robust application identification methods for P2P and VoIP traffic classification in backbone networks. Knowledge-Based Systems. 82: p. 152-162.

[3]. Liu, Y. and Y.F. Zheng, FS_SFS.2006. A novel feature selection method for support vector machines. Pattern Recognition.39(7): p. 1333-1345.

[4]. Seyed Mehdi Hazrati Fard, Ali Hamzeh , and S. Hashemi.2013. Using reinforcement learning to find an optimal set of features. Computers & Mathematics with Applications. 66(10): p.1892–1904.

[5]. Xin Sun, et al. 2012. Feature evaluation and selection with cooperative game theory. Pattern Recognition. 45(8): p.2992–3002.

[6]. Ahmed K. Farahat, Ali Ghodsi, and M.S. Kamel. 2013. Efficient greedy feature selection for unsupervised learning. Knowledge and Information Systems .35(2):p. 285-310.

[7]. Aghdam, M.H., N. Ghasem-Aghaee, and M.E. Basiri. 2009. Text feature selection using ant colony optimization. Expert Systems with Applications. 36(3): p. 6843-6853.

[8]. Eesa, A.S., Z. Orman, and A.M.A. Brifcani. 2015. A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Expert Systems with Applications. 42(5): p. 2670-2679.

[9]. Cho S-B, Park H-J. 2003. Efficient anomaly detection by modeling privilege flows with hidden Markov model. Computers and Security .22(1):p.45-55.

[10]. Fleuret, F.2004. Fast binary feature selection with conditional mutual information. The Journal of Machine Learning Research. 5, 1531–1555.

[11]. S. Chebrolu, A. Abraham, and J.P. Thomas. 2005. "Feature deduction and ensemble design of intrusion detection system", Computers & Security. 24(4): pp. 295-307.

[12]. Amiri, F., et al. 2011. Mutual information-based feature selection for intrusion detection systems. Journal of Network and Computer Applications. 34(4): p. 1184-1199.

