Journal of Information Security and Applications 44 (2019) 80–88

Contents lists available at ScienceDirect

Journal of Information Security and Applications

journal homepage: www.elsevier.com/locate/jisa

Cyber intrusion detection by combined feature selection algorithm

Sara Mohammadi [a], Hamid Mirvaziri [a], Mostafa Ghazizadeh-Ahsaee [a], Hadis Karimipour [b,*]

[a] Department of Computer Engineering, ShahidBahonar University, Kerman, Iran
[b] School of Engineering, University of Guelph, Guelph, Ontario, Canada, N1G 2W1

ARTICLE INFO

Article history:
Available online 1 December 2018

Index Terms:
Feature selection
Intrusion detection systems
Feature grouping
Linear correlation coefficient
Cuttlefish

ABSTRACT

Due to the widespread diffusion of network connectivity, the demand for network security and protection against cyber-attacks is ever increasing. Intrusion detection systems (IDS) perform an essential role in today's network security. This paper proposes an IDS based on feature selection and clustering algorithm using filter and wrapper methods. Filter and wrapper methods are named feature grouping based on linear correlation coefficient (FGLCC) algorithm and cuttlefish algorithm (CFA), respectively. Decision tree is used as the classifier in the proposed method. For performance verification, the proposed method was applied on KDD Cup 99 large data sets. The results verified a high accuracy (95.03%) and detection rate (95.23%) with a low false positive rate (1.65%) compared to the existing methods in the literature.

1. Introduction

IN spite of the significant development in network security, the existing solutions are unable to completely defend computer networks against the malicious threats. The traditional security techniques such as firewalls, user authentication and data encryption are not capable enough to fully safeguard the network security due to fast development of intrusion techniques [1,2]. Therefore, new defense mechanisms such as intrusion detection system (IDS) are suggested to facilitate system's security.

An IDS is a system that conducts the process of identifying attack behavior on a network. It plays an important role in monitoring and evaluating daily activities in computer systems to detect intrusions and security threats. Generally, IDSs are classified into two groups: signature-based or misuse-based detection systems and anomaly-based detection systems. Signature-based techniques detect anomalies by matching predefined attack's signatures [3,4]. The main advantages of these methods are their simplicity and low false positive rates, however, they are not able to detect new mimicry attacks.

Anomaly-based detection techniques rely on the assumption that the intruder's behavior is different from normal network behavior [3–6]. These techniques study the normal traffic of the network and identify each deviant behavior as malicious behavior. This system provides the possibility of detecting both unknown and known attacks. The main disadvantage of this system is its high false positive rate. Generally, an IDS is overwhelmed with huge amount of data with irrelevant and redundant features which cause a long-term problem in network traffic classification. One major limitation of current IDS technologies is the requirement to filter false alarms let the system be overwhelmed with data. This is due to irrelevant and additional features in dataset, which decrease the speed of detection.

KDD Cup 99 is a well-known evaluation dataset for intrusion detection field. It consists of more than five million of training samples and two million of test samples. Such a large-scale dataset slows down the classification process, or even degrade classifier's performance due to insufficient memory. Besides that, big data usually contain redundant and noisy features that present critical challenges to knowledge discovery and data modeling.

This work is motivated by above mentioned drawback. Utilizing a preprocessing step, performance of the IDS can be improved. Dimensionality reduction, as a preprocessing step, is used in this work to eliminate non-essential features from datasets. It reduces the size of the problem thereby accelerating the whole process. Dimensionality reduction includes feature extraction and feature selection which are fundamental steps in IDS construction. This article is focused on the feature selection step in IDS.

Feature selection is the preprocessing method, which can effectively solve the IDS problems by selecting relevant features and eliminating redundant and irrelevant features. Relevant features have useful information about the classes, which are essential for proper operation of the classifier. Unlike the relevant features, irrelevant features result in performance degradation in the operation of classifier [7]. These features carry no extra information

* Corresponding author.
E-mail addresses: sara.mohammadi.68@eng.uk.ac.ir (S. Mohammadi), hmirvaziri@uk.ac.ir (H. Mirvaziri), mghazizadeh@uk.ac.ir (M. Ghazizadeh-Ahsaee), hkarimi@uoguelph.ca (H. Karimipour).

https://doi.org/10.1016/j.jisa.2018.11.007
2214-2126/

about the classes since the selected features contain the same information. The advantages of feature selection include data understanding, data reduction, limitation of required storage space and reduction of processing cost [8].

The combinatory method proposed in this work takes advantage of both wrapper and filter methods. A decision tree named ID3 is used for system modeling. For classification, least square support vector machine (LSSVM) is selected. Support vector machine (SVM) is a supervised learning method that is used for classification and regression problems. In general, SVM may lead to computational complexity when a dataset is very large. LSSVM is usually employed for large datasets [15,16,25,27]. In this work, LSSVM is used as a classifier of the filter and ID3 is used as a classifier of our combinational feature selection method. A preprocessing feature selection step is proposed to improve the performance of IDS. A combination of filter and wrapper method as a feature selection is suggested in this work. Feature grouping based on linear correlation coefficient (FGLCC) can reduce the computational complexity of wrapper method named cuttlefish algorithm (CFA) by eliminating irrelevant and redundant features from original dataset. CFA is used to search for the best subset to improve the accuracy of classifier. The aim is to achieve high accuracy and a low false positive rate by using both accurate and optimal searching of the wrapper method and efficiency of the filter method.

Efficiency of the proposed FGLCC-CFA method is validated using several case studies on the benchmark KDD Cup 99 intrusion detection dataset. Performance of the FGLCC-CFA is compared with FGLCC and CFA methods individually. The experimental results show that the proposed method has higher accuracy and a lower false positive rate while reducing the processing time compared to the FGLCC and CFA [18], ID3-BA [31], KMSVM [29] and N-KPCA-GA-SVM [21] methods.

The key contributions of this paper are listed as follows:

- A new filter-based feature grouping method based on linear correlation coefficient (FGLCC) is proposed.
- A novel combinatory method is proposed to improve the performance of the FGLCC combined with cuttlefish algorithm (CFA).
- False positive rate is significantly reduced.
- Attack detection rate and accuracy rate are increased for different type of attacks.

The rest of the paper is organization as follows. Section 2 provides a literature review about the related works. Intrusion detection and feature grouping are explained in Section 3. Section 4 describe the proposed FGLCC-CFA method. Section 5 presents experimental results followed by the conclusion in Section 6.

2. Related works

Generally, feature selection includes three main methods: 1) filter methods [8,9], 2) wrapper methods [8,10,11], and 3) combination methods [10,12]. A filter method is a ranking method that ranks the features by their importance and then selects the best feature. This method is fast with low computational complexity and is scalable for high dimensional datasets [13]. A new method of feature selection including discretization, filtering and classification to improve the classification is proposed by Bolon-Canedo et al. [17]. The proposed method is examined on the KDD Cup 99 using both binary and multiple class classification. It results in a low false positive rate and fast performance however; the detection rate is below the average detection rate in similar existing methods. Amiri et al. [15] proposed a forward feature selection algorithm using the mutual information method to measure the relation among features. Then the LSSVM as a classifier was used to select the optimal feature set and to build the IDS. Horng et al. [35] proposed an SVM-based IDS, which combines a hierarchical clustering and the SVM. The hierarchical clustering algorithm was used to provide the classifier with fewer and higher quality training data to reduce the average training and test time and to improve the classification performance of the classifier. Mukkamala and Sung [33] proposed a novel feature selection algorithm to reduce the feature space of KDD Cup 99 dataset from 41 dimensions to 6 dimensions and evaluated the 6 selected features using an IDS based on SVM. The results show that the classification accuracy increases by 1 percent when using the selected features.

The second feature selection technique, wrapper method which is a searching method, selects a set of features that maximizes the objective function. The wrapper method is more accurate than the filter method, however it is computationally burdensome for high dimensional datasets [14]. Whereas, the datasets used in IDS have high dimensions like KDD Cup 99, wrapper methods are time-consuming. In addition, the complexities in time and computations may cause low accuracy. Recently, a wrapper method as an extension of the bee's algorithm was proposed in [31]. The results on KDD Cup 99 show that the proposed method has high false positive rate equals to 3.15% and low detection rate of 80%. In [18] CFA was employed as a search strategy to select optimal set of features and decision tree (DT) to evaluate selected subset of features. Aside from the fact that the proposed method is not compared with other existing techniques in terms of accuracy and speed, it has a high false positive rate about 3.91%. Ant Colony Optimization (ACO) technique for subset features selection as a wrapper method was presented in [23], and SVM was employed for classification technique. ACO has ability to select strong features that leads to quick and efficient detection rate. The main disadvantage of this method is its high false positive rate about 2%. A local search algorithm, which uses *K*-means clustering to group the training data set into two clusters: normal operation condition and denial of service (DoS) attacks is suggested in [22]. The proposed wrapper method in [22] reduces computational burden; however, it is only tested for DoS attacks.

Overall, majority of the techniques in the literature are either tested on one or two types of attacks or results in a high false positive rate and high processing time. To address the previously mentioned problems, this paper propose a combination of feature selection method (FGLCC-CFA). The upper phase (filter method (FGLCC)) tries to eliminate irrelevant and redundant features from the original data. This phase increases the pace of processing for the lower phase (wrapper method (CFA)) by decreasing the search range from the entire original feature space to the pre-selected features (the out-put of the filter method). This work has led to high detection and accuracy rates and low false positive rate and processing time compared to the literature works. In addition, to verify the performance of the proposed method, it is tested for all major types of intrusion attacks.

3. Intrusion detection system

An intrusion is the act of causing a disturbance in the system operation, compromising its safety, integrity and accessibility. Intrusion detection has two main elements; model and resource. Models explain the legal behavior of the resources and the techniques which compare the current system activities and the constructed models with each other in order to identify the intrusive actions. Resources, such as user accounts and file systems, are to be protected by the intrusion detection.

In recent years, cyber-attacks led to irreparable damage and financial losses in large-scale networks. These type of attacks can access vital information about the network and prevent legal users from being served by servers by Denial of Service (DoS). The duty of the IDS is to recognize and detect any suspicious activities in

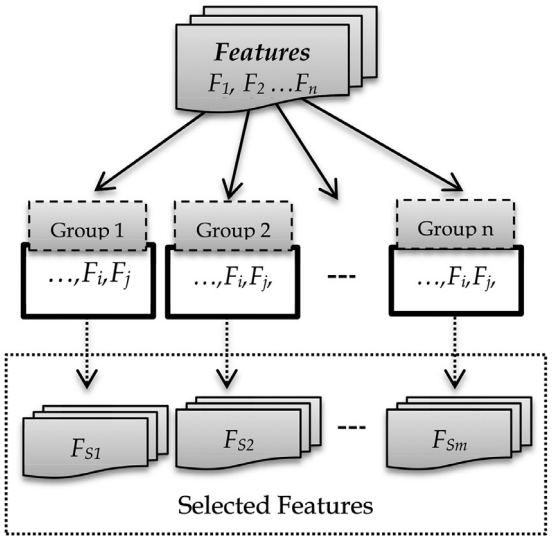

Fig. 1. Example for strategy of selection in feature grouping.

the system by gathering information and monitoring the system behavior [20]. The focus of this paper is on the network intrusion detection systems (NIDS) which are designed to detect attacks that target computer networks.

3.1. Feature grouping selection methodology

Feature grouping uses the relationship of features in a dataset to construct groups and design selection strategy to extract specific features from a group. It reduces the variance in the estimation and improves the stability of feature selection. Feature grouping is one of the best methods for large data set analysis. It leads to a better understanding and explanation of data [24].

The number of groups represents the number of features that should be selected. Moreover, different number of features can be selected from different groups. The goal is to construct groups utilizing one of the clustering methods to extract useful features based on specific criteria from each group. Different clustering methods and metrics results in different cluster constructions. The number of clusters affects the amount of features selected. For example, different strategies could be adopted if we expect to select 8 features from a dataset. We could create 8 groups by a clustering method and select 1 feature in each group or we could construct 4 groups and select 2 features per group instead. Fig. 1 shows one example of feature grouping.

3.2. Connection between feature selection and intrusion detection

Researchers have combined feature selection with intrusion detection to classify intrusions from normal connections. Although selecting important and relevant features from input data builds a fast and accurate IDS and conversely, selecting irrelevant features cause enormous problems in network traffic classification. The successful combination of feature selection and intrusion detection improves the detection and accuracy rates, decrease the false positive rate and the complexity of computation in the system [21,25]. Therefore, the intrusion detection classifier requires an accurate method to extract important features from data sets. The framework of intrusion detection based on the feature selection step and the classification step is shown in Fig. 2.

4. Proposed FGLCC-CFA feature selection method

This section describes the proposed combination feature selection method for intrusion detection systems which is called FGLCC-CFA. The proposed method combines filter and wrapper methods to select an optimum subset of features from a large dataset (KDD Cup 99) and improve IDS's performance.

4.1. Feature grouping based on linear correlation coefficient (FGLCC)

Correlation coefficient measure is the basis of linear correlations method which is used to determine the linear relations between two random variables [25]. It is well known because of its simplicity and low computational cost. Correlation coefficient between two variables (features) shows the degree of relations between them which is equal to the rate of their covariance divided by the product of their standard deviations. Consider two data sets $X=\{x_1,x_2,\ldots, x_n\}$ and $Y=\{y_1,y_2,\ldots, y_n\}$ where n is the number of samples. The linear correlation coefficient corr(X, Y) for two variables X and Y can be defined as follows:

$$p(X,Y)=\frac{cov(X,Y)}{\sigma_X\sigma_Y}=\frac{\sum_{i=1}^{n}\left(x_i-\bar{X}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{X}\right)^2\sum_{i=2}^{n}\left(y_i-\bar{Y}\right)^2}} \quad (1)$$

Where cov(X, Y) is the covariance between X and Y; σ_X and σ_Y are standard deviation which are defined as follows. $\bar{X}$ and $\bar{Y}$ are the average value of X and Y, respectively.

The first step in the proposed FGLCC algorithm is to compute correlation coefficient between features and classes. The feature with the maximum correlation will be selected and inserted in vector S as the first feature. Afterselecting the first feature, the number of selected features is proportional to the number of groups. Then this selected feature will be removed from feature set *F* and added to the *S* set. In the next step, the algorithm computes correlation between every individual feature and all other features in *F*. The results will be collected in a matrix ($Matrix_{corr}$) which is defined as follows:

$$Matrix_{corr}=\begin{bmatrix} a_{11} & \cdots & a_{1j} \\ \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{ij} \end{bmatrix} \quad (2)$$

$a_{ij}=corr(f_i;f_j),(f\in F);$
$vector(i)=[a_{i1},a_{i2},\ldots,a_{ij}],$
$i=\{1,2,\ldots,n\},\ j=\{1,2,\ldots,n\}$

where *n* is equal to the existed features in *F* set. The vectors in $Matrix_{corr}$ will be ranked by *K*-means algorithm. *K*-means is a method of clustering which divides data into the clusters. Finally, FGLCC computes the evaluation function E_{FGLCC} for all of the features in each group and selects the feature from each group which has maximum E_{FGLCC}. In other words, the feature which has maximum correlation with the class ($corr(C; f_i)$) and minimum relation with other selected features ($corr(f_i; f_s)$) is inserted to the S set. Details of each step for the proposed FGLCC algorithm is shown in Fig 3.

4.2. Strategy of selection in cuttlefish algorithm

Cuttlefish algorithm (CFA) imitates the mechanisms of the cuttlefish which changes its color. The CFA has two key processes; the first one is reflection, which is used to imagery the light reflection mechanism, and the second one is visibility, which is used to simulate the visibility of matching patterns.

CFA is a heuristic algorithm which extracts an optimal subset of features from initial features. It has employed a decision tree as a classifier. Decision tree (DT) is one of the well-known machine

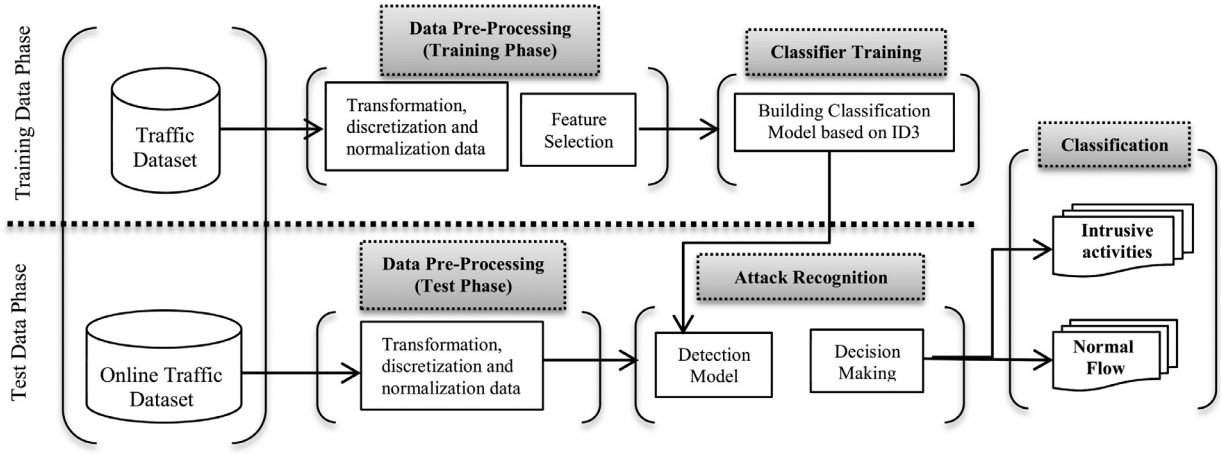

Fig. 2. The framework of intrusion detection system based on ID3.

Algorithm 1: FGLCC Algorithm

Input: Feature set $F = \{f_i, i = 1,..., n\}$, G

Output: S: the selected feature subset

Step 1. Initialization:

Set $S = \emptyset$ and C = class label, G = number of groups

Step 2. Compute $corr(C; f_i)$ for each feature, i = 1, … , n

Step 3. Select feature f_i:

Select feature f_i:

$$\arg\max_{f_i}(corr(C; f_i)), i = 1,..., n;$$

Then, set:

$$F = F - \{f_i\}, S = S \cup \{f_i\}, n = n - 1$$

Step 4. Compute $corr(f_i; f_j)$ between feature

Select feature f_i ($f_i \in F$) and other features f_j ($f_j \in F$)

Insert the results in $Matrix_{corr}$.

Step 5. Rank vectors in $Matrix_{corr}$ by K-means clustering and assign them in G groups.

Step 6. Do the following for each group in G :

Select feature f_i:

$$\arg\max_{f_i}(E_{FGLCC} = corr(C; f_i) - \beta \times \sum_{f_s \in S} corr(f_i; f_s)))$$

(corr: correlation coefficient function)

Then, set:

$$S = S \cup f_i,\ n = n - 1$$

Step 7. Return S includes selected features.

Fig. 3. Feature grouping based on linear correlation coefficient algorithm.

learning techniques which contain three components: nodes, arcs and leaves. The tree is constructed during a training phase. During the test step, according to the result of the test data along the way, each test data is classified by the navigation from the root of the tree down to the leaf. There are two popular DT algorithms which are ID3 and C4.5. ID3 is used in our proposed method. Additional information about CFA algorithm is available in [18].

To increase the accuracy and speed of the FGLCC algorithm, it is combined with CFA algorithm. At the first step, all of the features are ranked through the FGLCC algorithm. After that, the best features which have the highest ranks are selected. These features are used as an input of wrapper method (CFA) to improve the performance of the CFA. CFA searches between the input features to select an optimum subset for higher accuracy and lower false positive rates compared to the previous values. Fig. 4 explains the proposed FGLCC-CFA algorithm in detail.

5. Experiment and results

Classification has four main steps: data gathering, data preprocessing, classifier training, and attack detection. These steps are described as follows:

5.1. Data collection

KDD Cup 99 dataset has been the most widely used dataset for the evaluation of anomaly detection methods. This dataset is derived from DARPA 1998. KDD Cup 99 includes training and testing datasets named “10%KDD Cup 99’’ and “Corrected label KDD”, respectively. Each of the single connection vectors in both KDD training and testing datasets contains 41 features and is labeled as either normal or an attack. The attacks fall in one of the following four categories:

1. Denial of Service Attack (DoS): The aim of the attacker is to make the network busy by sending many requests at once, therefore the network cannot handle legitimate requests. The examples are back, land, Neptune, pod, smurf and teardrop attacks.
2. User to Root Attack (U2R): The attacker in U2R attacks starts out with access to normal user account on the network system and exploit vulnerability to gain root access to the system. The examples are rootkit, per, load module and buffer-overflow attacks.
3. Remote to Local Attack (R2L): The attacker in R2U attacks send packets to a target machine over a network, then exploits vulnerability to gain local access as a user of that machine. The examples are ftp, write, spy, phf, guess-passwd, imap, warezclient, wrezmaster and multihop attacks.
4. Probing Attack: The attacker in Probe attacks scans a network to gather important information about target computers. The examples are nmap, ipsweep, portsweep and satan attacks.

It should be mentioned that the test dataset includes new attacks which are different from the attacks used in the training dataset. Therefore, the algorithm can be tested against both unseen and known attack, which makes the attack detection analysis more realistic. The datasets contain a total number of 24 training

Algorithm 2: FGLCC-CFA Algorithm

Input: candidate subset of selected features S.
Output: bestSubset:the best selected features from S

1. Initialize population P[n] with random solutions; initialize t
2. Evaluate fitness of the population using DT, keep the best solution in $AV_{Bestsubset}$ and in bestSubset
3. Remove one feature from bestSubset
4. While (stopping criterion is not met)

Case 1,2:
Sort the population P in descending order
K=random(N/2)
for(i=0; i<k; i++)
{ R=random($AV_{Bestsubset}$.selectedFeatures.Size)
$AV_{Bestsubset}$
V= p_i. selectedFeatures.Size-R;
$Visibility_i$[V]
newSubset = []
Chose R features randomly from p_i.selectedFeatures and add them to $Reflection_i$
Chose V features randomly from p_i.uselectedFeatures and add them to $Visibility_i$
newSubset=[]=$Reflection_i \oplus Visibility_i$
Evaluate newSubset using DT
If (newSubset is better than $AV_{Bestsubset}$)
$AV_{Bestsubset}$=newSubset
}

1. **Case 3,4:**
for(i=0; i<t; i++)
{ Use bestSubset and exchange one feature randomly between selectedFeatures and unselectedFeatures
If(newSubset is better than bestSubset)
bestSubset=newSubset}

2. **Case 5:**
for(i=0; i< $AV_{Bestsubset}$. SelectedFeatures.Size; i++)
{ Create newSubset by removing feature i from $AV_{Bestsubset}$.SelectedFeatures
If(newSubset is better than bestSubset)
bestSubset=newSubset}

3. **Case 6:**
for(i=0; i<k; i++)
{ Generate newSubset randomly
evaluate newSubset using DT
if(newSubset is better than p_i)
Pi=newSubset;
if($p_i$better than $AV_{Bestsubset}$)
$AV_{Bestsubset}$= p_i}

5. End while

Return bestSubset

Fig. 4. Pseudo code of proposed FGLCC-CFA method.

attacks, with an additional 14 unseen attack only in the test data set. The name and detail description of the training attack types are listed in [34].

As mentioned before, each datum in this dataset is labeled as attack or normal data and includes 41different features. These features are classified into three main groups. First group includes basic features (1 to 9 features) which extracted from TCP/IP connections. Second group are content-based features, which includes features 10 to 22. This group applies information of network packet payloads. Third group belongs to the features of traffic, which includes features 23 to 41. A list of KDD'99 features with detailed descriptions are listed in Table 1.

The number of records in the training dataset "10%KDD Cup 99" and testing dataset "KDD Corrected label" are equal to 494,020 and 3,110,288, respectively. To avoid computational burden due to the large size of data sets, a random number of records are selected from training and testing datasets. The number of selected data in KDD Cup 99 dataset is shown in Table 2.

5.2. Data preprocessing

Data processing includes four main phases:

1. Data transfer: detection model needs all of the input records in format of vectors of real numbers. Therefore, symbolic feature in the dataset should be transformed into the numeric values.
2. Data discretization: the aim of discretization is to limit the continuous values to the limit sets. Discretized data causes a better classification. Most of the features in KDD Cup 99 dataset are continuous. The method of discretization in [26] is used in this paper.

$$Normalized\_X = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{3}$$

3. Data normalization: each feature in KDD Cup 99 dataset has different data limitation; so they are normalized into the specific range of [0,1] using the following equation:
4. Feature selection: each feature in dataset has 41 features, but not all of these features are necessary to build an IDS. Therefore, the most important features should be selected to achieve the highest efficiency.

5.3. Classifier training

Once the optimal feature subset is selected by the propsoed FGLCC, it will be sent to the classifier training stage where LS-SVM is employed. The classifier distinguishes attacks from Normal data. At the next phase, when selected features are applied to the CFA algorithm, each subset of selected features in each step of CFA are given to the ID3 classifier for evaluating, and then the best subset of features are selected.

5.4. Attack detection

In this paper, decision making is based on two classes: attacked data and normal data. After feature selection processing, ID3 will be trained to detect intrusions from normal traffics.

5.5. . Performance evaluation

Four performance measures are employed to evaluate the performance of ID3 system. These measures include detection rate (DR), accuracy rate (AR), false positive rate (FPR) and fitness function (F) which are defined in Eqs. (4)–(6).

$$AR = \frac{TP + TN}{TP + TN + FN + FP} \tag{4}$$

$$DR = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

Table 1
Different Groups of Features in KDD Cup 99 Dataset.

Group	Frature name	Description	Type
G1	1. Duration	Length (number of seconds) of the connection	Continuous
	2. Protocol-Type	Type of the protocol, e.g. tcp, udp, etc.	Discrete
	3. Service	Network service on the destination, e.g., http, telnet, etc.	Discrete
	4. Src-bytes	Number of data bytes from source to destination	Continuous
	5. Dst-bytes	Number of data bytes from destination to source	Continuous
	6. Flag	Normal or error status of the connection	Discrete
	7. Land	1 if connection is from/to the same host/port; 0 otherwise	Discrete
	8. wrong-fragment	Number of "wrong" fragments	Continuous
	9. Urgent	Number of urgent packets	Continuous
G2	10. Hot	Number of "hot" indicators	Continuous
	11. Num-faild-logins	Number of failed login attempts	Continuous
	12. Logged-in	1 if successfully logged in; 0 otherwise	Discrete
	13. Num-compromised	Number of "compromised" conditions	Continuous
	14.Root-shell	1 if root shell is obtained; 0 otherwise	Discrete
	15. Su-attempted	1 if "su root" command attempted; 0 otherwise	Discrete
	16. Num-root	Number of "root" accesses	Continuous
	17. Num-file- creations	Number of file creation operations	Continuous
	18. Num-shells	Number of shell prompts	Continuous
	19. Num-access- files	Number of operations on access control files	Continuous
	20. Num- outbound-cmds	Number of outbound commands in an ftp session	Continuous
	21. Is-hot-login	1 if the login belongs to the "hot" list; 0 otherwise	Discrete
	22. Is-guest-login	1 if the login is a "guest"login; 0 otherwise	Discrete
G3	23. Count	Number of connections to the same host as the current connection in the past two seconds	Continuous
	24. Srv-count	Number of connections to the same service as the current connection in the past two second	Continuous
	25. Serror-rate	% of connections that have "SYN__errors	Continuous
	26. Rerror-rate	% of connections that have "REJ" errors	Continuous
	27. Same-srv-rate	% of connections to the same service	Continuous
	28. Diff-srv-rate	% of connections to different services	Continuous
	29. Srv-serror-rate	% of connections that have "SYN" errors of connections that have "REJ" errors%	Continuous
	30. Srv-rerror-rate	% of connections to different hosts	Continuous
	31. Srv-diff-host-rate		Continuous
G4	32. Dst-host-count	Count for destination host	Continuous
	33. Dst-host-srv-count	Srv-count for destination host	Continuous
	34. Dst-host-same-srv-rate	Same-srv-rate for destination host	Continuous
	35. Dst-host-diff-srv-rate	Diff-srv-rate for destination host	Continuous
	36. Dst-host-same-src-port-rate	Same-src-port-rate for destination host	Continuous
	37. Dst-host-srv-diff-host-rate	Diff-host-rate for destination host	Continuous
	38. Dst-host-serror-rate	Serror-rate for destination host	Continuous
	39. Dst-host-srv-serror-rate	Srv-serror-rate for destination host	Continuous
	40. Dst-host-rerror-rate	Rerror-rate for destination host	Continuous
	41. Dst-host-srv-rerror-rate	Srv-serror-rate for destination host	Continuous

Table 2
Selected data in 10%KDDcup and corrected.

Data set	Train	Test
Normal	973	606
DOS	3915	2299
Probe	41	42
R2l	13	160
U2r	5	10

where true positive (TP) is the number of attacks that have been classified properly, true negative (TN) is the number of normal records that have been classified properly, false positive (FP) is the number of normal records that have been classified as attacks, and false negative (FN) is the number of attacks that have been classified as normal.

Fitness is a statistical technique used to evaluate the accuracy of the system. Classifier system will determine which subset of features are the best according to the fitness defined in Eq. (11):

$$Fitness\,(F) = (\alpha \times DR) + (\beta \times (100 - FPR))$$

Where $\alpha \varepsilon [0,\ 1]$ and $\beta = 1 - \alpha$. α and β are two parameters which indicate the significance of the DR and FPR qualities, respectively. In this experiment, $\alpha = 0.7$ and $\beta = 0.3$ are considered as DR is more important than FPR.

5.6. Results

In this paper, the parameter β of the filter method (FGLCC) is adjusted between 0.3 and 1. This range is suggested by [15,28,32]. β is a penalty parameter and those features which have more correlation with selected features are punished by receiving a different value of parameter β. $\beta = 0$ only considers the correlation between feature and class and $\beta = 1$ assigns all weight to the correlations between features; none of these values are acceptable. β should be adjusted with a suitable value to keep the balance among the correlations between features and the correlations of features and class.

Fig. 5 shows the accuracy and false positive rates of filter method (FGLCC) for different values of β. As seen in Fig. 4, FGLCC has a high AR and low FPR when $\beta = 0.3$. Therefore, FGLCC is applied on KDD Cup 99 when $\beta = 0.3$.

Performance of the FGLCC is compared with two well-known feature selection algorithms in the literature: feature grouping based on mutual information (FGMI) [30] and linear correlation coefficient feature selection (LCFS) [15]. The proposed FGLCC method is compared with FGMI [30] and LCFS [30] because both of these algorithms are well-known feature grouping method (which use a basic algorithm like our FGLCC) in the literature. As seen from Fig. 6, proposed method has a better performance compared to other feature grouping techniques, which are used in FGMI, and LCFS. As shown in Fig. 6, the maximum different between DR and

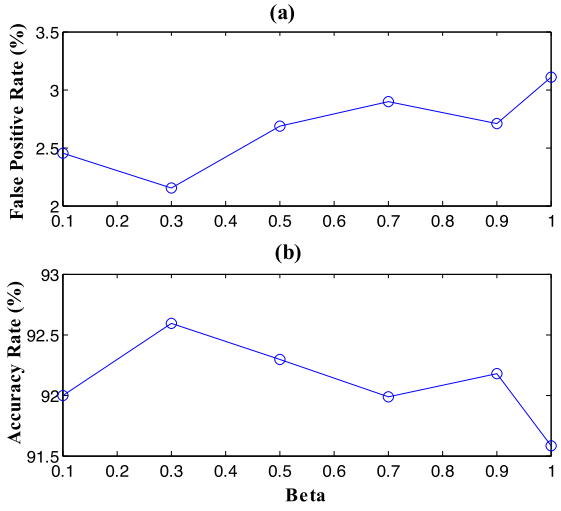

Fig. 5. (a) FPR of proposed algorithm for different β, (b) AR of proposed algorithm for different β and in KDD Cup 99 dataset.

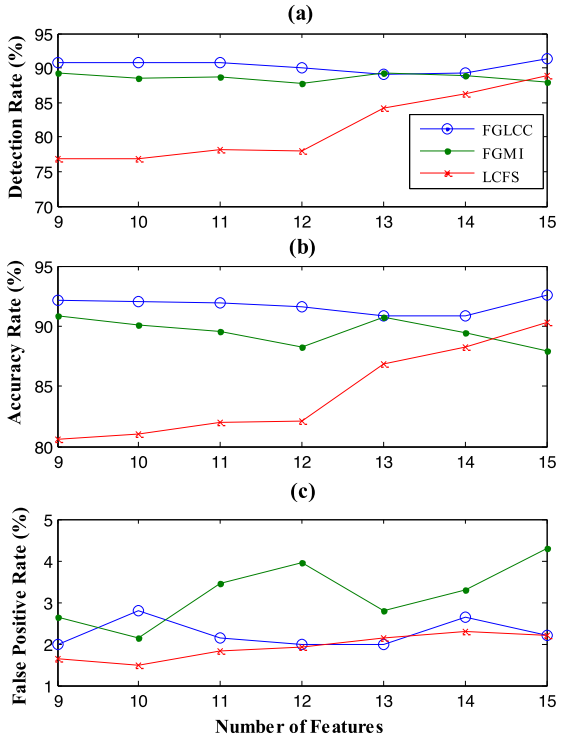

Fig. 6. (a) DR, (b) AR and (c) FPR of FGLCC compared with FGMI and LCFS on KDD Cup 99.

FPR are in 15, 9 and 15 features for the algorithms FGLCC, FGMI [30], and LCFS [15], respectively. Therefore, these number of features are considered as the number of selected features for those algorithms. Proposed FGLCC algorithm results in a better detection rate and a lower false positive rate compare to FGMI and LCFS. Overall, FGLCC accurately detects the attack with an average of 91% accuracy while the detection rate in FGMI and LCFS has an average of 88% and 80%, respectively.

To improve the performance of the FGLCC, it is combined with CFA algorithm. Table 3 shows the selected features for our proposed filter method (FGLCC) between original 41 features of KDD Cup 99, and our combinational method between 15 selected features by FGLCC. Detailed steps for the algorithm are provided in the previous section. Fig. 7 shows the results of feature selection

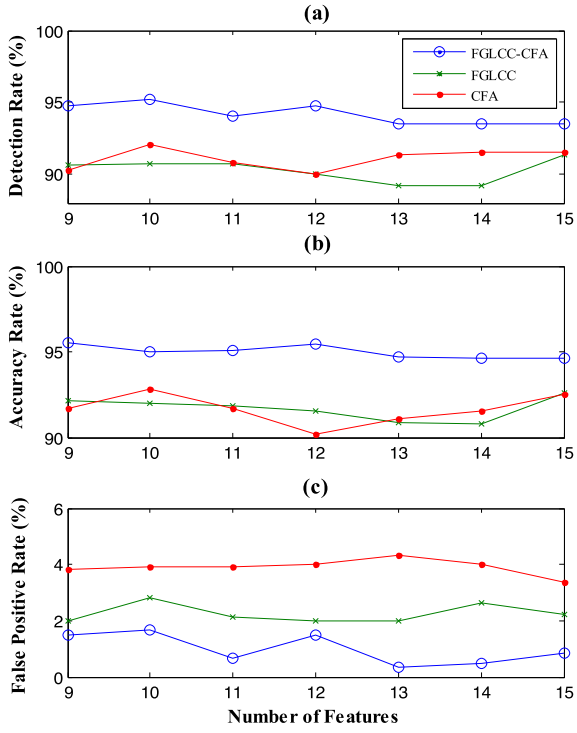

Fig. 7. (a) DR, (b) AR and (c) FPR of FGLCC-CFA compared with FGLCC, CFA on KDD Cup 99.

using proposed FGLCC-CFA algorithm compared to FGLCC and CFA on Corrected Labels of KDDcup99 dataset. As can be seen in Fig. 7, the proposed combined FGLCC-CFA outperforms FGLCC and CFA. The best result is achieved for 10 selected features as the highest difference between DR which was equal to 95.23% and FPR equal to 1.65%. This highest difference between DR and FPR for FGLCC and CFA algorithms is in 15 and 10 selected features, respectively.

In addition to the basic methods (FGLCC and CFA), FGMI [30] and LCFS [15], the proposed combinatory method (FGLCC-CFA) is compared with three literature works; two novel hybrid methods include N-KPCA-GA-SVM [19] and KMSVM [22], and a reputable and new wrapper method, a combination of Bees algorithm with ID3, which is called ID3-BA [31]. All of these techniques are used for verification purposes in the literature. The results of DR, AR, FPR and fitness of well-known algorithms on corrected labels of KDDcup99 dataset are summarized in Table 4. Overall, FGLCC-CFA has the 4.28% and 3.45% improvement in DR, 2.64% and 2.36% in AR, 23.25% and 57.69% in FPR compared to the FGLCC and CFA respectively.

To evaluate the efficiency of the proposed FGLCC-CFA algorithm, its processing time is compared with the FGLCC and CFA algorithms. The building and testing time based on Corrected Labels of KDD Cup 99 dataset are available in Table 5. It can be observed that FGLCC-CFA significantly reduces the building time and testing time compared to the wrapper method (CFA). It is slightly slower than FGLCC which is negligible considering the improved performance. The reason is that CFA is computationally onerous algorithm so its combination with FGLCC increases the execution time. Here the goal is to compromise between speed-up and accuracy to get the best result possible.

To further verify the functionality of the proposed method, its performance is tested using 10-fold cross validation method on KDDcup99 data set. As shown in Table 6, FGLCC-CFA has the highest DR of 99.85% and AR of 99.84% and the lowest FPR of 0.19% among all.

Table 3
Selected features by FGLCC and FGLCC-CFA methods on KDD Cup 99.

Methods	# Feature	Selected features
FGLCC-CFA	10	$f_{23}, f_{36}, f_4, f_{22}, f_{29}, f_{10}, f_{24}, f_{41}, f_{35}, f_{13}$
FGLCC	15	$f_{23}, f_{36}, f_4, f_{22}, f_{29}, f_{10}, f_{32}, f_6, f_{40}, f_{39}, f_{35}, f_{27}, f_{13}, f_{24}, f_{41}, f_{30}$

Table 4
Summary of classification performance comparison.

Methods	DR%	AR%	FPR%	Fitness%
FGLCC-CFA	**95.23**	**95.03**	**1.65**	**95.46**
FGLCC	91.32	92.59	2.15	93.28
CFA	92.05	92.83	3.90	93.26
FGMI	89.32	90.83	2.64	91.73
LCFS	88.92	90.30	2.19	91.58
N-KPCA-GA-SVM	91.21	92.51	2.01	93.24
ID3-BA	91.57	92.59	3.16	93.15
KMSVM	88.71	87.01	-	-

Table 5
Summary of processing time comparison.

Methods	Building time (s)	Testing time (s)
FGLCC-CFA	83.28	43.50
FGLCC	78.37	38.21
CFA	350.60	110.45

Table 6
Classification performance for ALL ATTACKS through 10Fold-Cross validation on KDDcup99 train dataset.

Methods	DR%	AR%	FPR%	Fitness%
FGLCC-CFA	**99.85**	**99.84**	**0.19**	**99.84**
FGLCC	90.68	92.51	1.05	93.16
CFA	95.74	95.80	1.86	95.16
N-KPCA-GA-SVM	92.11	93.51	1.75	93.95

6. Conclusion

In this work a unique filter-based feature selection algorithm is proposed to increase the detection rate in feature selection while reducing the false positive rate. The proposed FGLCC led to a better result in detection, accuracy and false positive rates compared to LCFS and FGMI. To further improve the performance of the FGLCC, it is combined with CFA for intrusion detection system. FGLCC-CFA algorithm combines the filter and wrapper methods which utilize the benefits of both of them. In the CFA method, selecting an optimum subset of features is a time-consuming task. To solve this problem, the proposed FGLCC filter is utilized to rank the initial features and select the best one. The selected features are input features to CFA. The proposed FGLCC-CFA algorithm makes use of the high speed of FGLCC and the high accuracy of CFA to select a subset of features.

Performance of the FGLCC-CFA is verified in comparison to the original filter (FGLCC), wrapper (CFA) methods and three other methods such as ID3-BA, N-KPCA-GA-SVM and KMSVM. The results show higher AR and DR equal to 95.03% and 95.23%, respectively and a lower FPR of 1.65%.

References

[1] Tavallaee M, Stakhanova N, Ghorbani AA. Toward credible evaluation of anomaly-based intrusion-detection methods. *IEEE Trans Syst* Man Cybern Part C (Appl Rev) 2010;40(5):516–24.
[2] Tapiador JE, Orfila A, Ribagorda A, Ramos B. Key-recovery attacks on KIDS, a keyed anomaly detection system. IEEE Trans Dependable Secure Comput 2015;12(3):312–25.
[3] Hwang K, Cai M, Chen Y, Qin M. Hybrid intrusion detection with weighted signature generation over anomalous internet episodes. IEEE Trans Dependable Secure Comput 2007;4(1):41–55.
[4] Kabir E, Hu J, Wang H, Zhuo G. A novel statistical technique for intrusion detection systems. Future Gener Comput Syst 2018;79:303–18.
[5] Maggi F, Matteucci M, Zanero S. Detecting intrusions through system call sequence and argument analysis. IEEE Trans Dependable Secure Comput 2010;7(4):381–95.
[6] Karimipour H, Dinavahi V. Robust massively parallel dynamic state estimation of power systems against cyber-attack. IEEE Access Dec. 2017;6:2984–95.
[7] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27(8):1226–38.
[8] El-Khatib K. Impact of feature reduction on the efficiency of wireless intrusion detection systems. IEEE Trans Parallel Distrib Syst 2010;21(8):1143–9.
[9] Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Electr Eng 2014;40(1):16–28.
[10] El-Alfy ESM, Alshammari MA. Towards scalable rough set based attribute subset selection for intrusion detection using parallel genetic algorithm in MapReduce,. Simul Modell Pract Theory 2016;64:18–29.
[11] Bostani H, Sheikhan M. Hybrid of binary gravitational search algorithm and mutual information for feature selection in intrusion detection systems. Soft Comput 2017;21(9):2307–24.
[12] Ganapathy S, Kulothungan K, Muthurajkumar S, Vijayalakshmi M, Yogesh P, Kannan A. Intelligent feature selection and classification techniques for intrusion detection in networks: a survey. EURASIP J Wirel Commun Netw 2013;2013(1):271.
[13] Peng Y, Wu Z, Jiang J. A novel feature selection approach for biomedical data classification. J Biomed Inf 2010;43(1):15–23.
[14] Huang J, Cai Y, Xu X. A hybrid genetic algorithm for feature selection wrapper based on mutual information. Pattern Recognit Lett 2007;28(13):1825–44.
[15] Amiri F, Yousefi MR, Lucas C, Shakery A, Yazdani N. Mutual information-based feature selection for intrusion detection systems. J Netw Comput Appl 2011;34(4):1184–99.
[16] Huang G-B, Zhou H, Ding X, Zhang R. Extreme learning machine for regression and multiclass classification. IEEE Trans Syst Man Cybern Part B Apr. 2012;42(2):513–29.
[17] Bolon-Canedo M, Sanchez-Marono N, Alonso-Betanzos A. Feature selection and classification in multiple class datasets: an application to KDD Cup 99 dataset. Expert Syst Appl 2011;38(5):5947–57.
[18] Eesa AS, Orman Z, Brifcani AMA. A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Expert Syst Appl 2015;42(5):2670–9.
[19] Kuang F, Xu W, Zhuang S. A novel hybrid KPCA and SVM with GA model for intrusion detection. Appl Soft Comput 2014;18:178–84.
[20] R. Di Pietro and L. V. Mancini (Eds.) ,"Intrusion detection systems," vol. 38, Springer Science & Business Media, 2008.
[21] Wang J, Wen R, Li J, Yan F, Zhao B, Yu F. Detecting and mitigating target link-flooding attacks using SDN. IEEE Trans Dependable Secure Comput 2018;1(1).
[22] Kang SH, Kim KJ. A feature selection approach to find optimal feature subsets for the network intrusion detection system. Cluster Comput 2016;19(1):325–33.
[23] Mehmod T, Rais HBM. Ant Colony Optimization and Feature Selection for Intrusion Detection. In: Advances in machine learning and signal processing. Springer International Publishing; 2016. p. 305–12.
[24] Liu X, Lang B, Xu Y, Cheng B. Feature grouping and local soft match for mobile visual search. Pattern Recognit Lett 2012;33(3):239–46.
[25] Ambusaidi MA, He X, Nanda P, Tan Z. Building an intrusion detection system using a filter-based feature selection algorithm. IEEE Trans Comput 2016;65(10):2986–98.
[26] Mazumder S, Sharma T, Mitra R, Sengupta N, Sil J. Generation of sufficient cut points to discretize network traffic data sets. In: International conference on Swarm, Evolutionary, and memetic computing. Berlin, Heidelberg: Springer; 2012. p. 528–39.
[27] Ghorbani AA, Lu W, Tavallaee M. Network intrusion detection and prevention: concepts and techniques, vol. 47. Springer Science & Business Media; 2009.
[28] Kwak N, Choi CH. Input feature selection for classification problems. IEEE Trans Neural Netw 2002;13(1):143–59.
[29] Ravale U, Marathe N, Padiya P. Feature selection based hybrid anomaly intrusion detection system using K means and RBF kernel function. Procedia Comput Sci 2015;45:428–35.
[30] Song J, Zhu Z, Price C. Feature grouping for intrusion detection based on mutual. J Commun 2014;9(12):987–93.
[31] Eesa AS, Orman Z, Brifcani AMA. A new feature selection model based on ID3 and bees algorithm for intrusion detection system. Turk J Electr Eng Comput Sci 2015;23:615–22.
[32] Battiti R. Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 1999;5(4):537–50.

[33] Mukkamala S, Sung AH. Significant feature selection using computational intelligent techniques for intrusion detection. Proc Adv Methods Knowl Discovery Complex Data 2005;5(4):285–306.
[34] MIT Lincoln Labs. 1998 DARPA Intrusion Detection Evaluation February Available on: http://www.ll.mit.edu/mission/communications/ist/corpora/ideva/index.html.
[35] Horng S-J, Su M-Y, Chen Y-H, Kao T-W, Chen R-J, Lai J-L, Perkasa CD. A novel intrusion detection system based on hierarchical clustering and support vector machines. Expert Syst Appl 2011;38(1):306–13.

Sara Mohamadi received the B.Sc. degree in Information Technology from Fasa university, Fasa, Iran, in 2014, and the M.Sc. degree in Artificial Intelligence from Shahid Bahonar University of Kerman, Kerman, Iran, in 2017. She is currently research associate at Shahid Bahonar University of Kerman, Kerman, Iran. Her research interests include machine learning, artificial intelligent and cyber-security analysis.

Hamid Mirvaziri received the B.Sc. degree from Shahid Bahonar University, Kerman, Iran, in 1999, the M.Sc. degree from Guilan University, Guilan, Iran, and the Ph.D. degree from National University of Malaysia, Bangi, Malaysia, in 2010, all in Computer Engineering. His research area is in the field of signal processing and web and network security. Currently he is Assistant Professor in the Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran. His research is focused in signal processing and web and network security.

Mostafa Ghazizadeh-Ahsaee received the B.Sc. degree from Shahid Bahonar University, Kerman, Iran, in 2004. He received the M Sc. and Ph.D. degree in Computer Engineering both from the Department of Electrical and Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran, in 2007 and 2014, respectively. Currently he is an Assistant Professor of Computer Engineering Department of Shahid Bahonar University of Kerman, Kerman, Iran. His research fields of interest are Bayesian networks, machine learning and data mining algorithms, database management systems and also parallel and distributed systems.

Hadis Karimipour received the Ph.D. degree in Energy System from the Department of Electrical and Computer Engineering in the University of Alberta in Feb. 2016. Before joining the University of Guelph, she was a postdoctoral fellow in University of Calgary working on cyber security of the smart power grids. She is currently an Assistant Professor at the School of Engineering, Engineering Systems and Computing Group at the University of Guelph, Guelph, Ontario. Her research interests include large-scale power system state estimation, cyber-physical modeling, cyber-security of the smart grids, and parallel and distributed computing. She is member of IEEE and IEEE Computer Society. She serves as the Chair of the IEEE Women in Engineering (WIE) and chapter chair of IEEE Information Theory in Kitchener-Waterloo section.