Journal of AI and Data Mining
Vol 6, No 1, 2018, 157-162

Intrusion Detection based on a Novel Hybrid Learning Approach

L. Khalvati*, M. Keshtgary and N. Rikhtegar

Department of Computer & Information Technology, Shiraz University of Technology, Shiraz, Iran.

Received 27 August 2016; Revised 01 February 2017; Accepted 03 June 2017
**Corresponding author: l.khalvati@sutech.ac.ir (L. Khalvati).*

Abstract
The information security and the Intrusion Detection System (IDS) play a critical role in the internet. IDS is an essential tool for detecting different kinds of attacks in a network and maintaining data integrity, confidentiality, and system availability against possible threats. In this paper, a hybrid approach is proposed towards achieving a high performance. In fact, the important goal of this paper is to generate an efficient training dataset. In order to exploit the strength of clustering and feature selection, an intensive focus on intrusion detection combines the two, so the proposed method is using these techniques as well. At first, a new training dataset is created by K-Medoids clustering and Selecting Feature using the SVM method. Then Naïve Bayes classifier is used for evaluation. The proposed method is compared with another mentioned hybrid algorithm and also 10-fold cross validation. The experimental results based on the KDD CUP'99 dataset show that the proposed method has a better accuracy and detection rate and also false alarm rate than the others.

Keywords*: Intrusion Detection System, K-Medoids, Feature Selection, Naïve Bayes, Hybrid Learning Approach.*

1. Introduction

Today, internet access has become an important part of our daily life but the huge worldwide connections have caused security issues [1]. A secure network must have three features: confidentiality, integrity, and availability. Confidentiality means that accessing the network's data should be allowed only for the authorized people; integrity means that data should not be distorted during its transmission through the network; and availability means that whenever the information is required, it should be available to the authorized people.

Intrusion detection system (IDS) is a defensive system whose main goal is to detect actions that attempt to deny the network security features. Generally, there are two main types of intrusion detection systems: Signature-based Intrusion Detection System (SIDS) and Anomaly-Based Intrusion Detection System (AIDS) [2]. SIDS is the process of detecting harmful activities based upon known patterns of previous attacks, whereas AIDS is the process of detecting detrimental activities whenever the behavior of the system deviates from the normal behavior. AIDS can be executed by different techniques such as Naïve Bayes classifier, which is used in this paper, to improve the accuracy of IDS.

In the present work, we propose a multi-level approach through a combination of K-Medoids clustering, Selecting Feature using SVM algorithm and also Naïve Bayes classifier to improve the performance of IDS. First of all, K-Medoids clustering and Selecting Feature using the SVM algorithms are used to construct a new training dataset. Then the new training dataset is utilized to train the Naïve Bayes classifier. The results obtained demonstrate that the proposed method performs better in terms of accuracy, detection rate, and also false alarm rate.

The remainder of this paper is organized as what follows. Related work is discussed in Section 2. Section 3 represents the materials and methods that are used in this work. Section 4 describes the evaluation metrics. Our experiments are represented in Section 5. Finally, the paper is concluded in Section 6.

2. Related work

In the recent years, various hybrid IDS systems have been developed to achieve the best possible performance. In this section, we will review some of these methods that did not pay attention to building an efficient training dataset and normalization or made it by the K-Means algorithm.

Aslahi-Shahri et al. [3] have proposed a hybrid method that integrates SVM and genetic algorithm (GA). The experimental results on the KDDCUP'99 dataset have shown that this method is capable of achieving the good true-positive and also false-positive values.

Ravale et al. [4] have presented a hybrid approach based upon combining K-Means clustering algorithm and RBF kernel function of SVM method for IDS. The evaluation results show that their method performs better in terms of detection rate and accuracy when applied to the KDDCUP'99 dataset.

Esmaily et al. [5] have introduced a method based upon the integration of Decision Tree (DT) algorithm and Multi-Layer Perception (MLP) Artificial Neural Network (ANN). The results obtained reveal that the hybrid method is able to identify the attacks with high accuracy and reliability.

Anita et al. [6] have applied a hybrid approach based upon the K-Nearest Neighbor, K-Means, and Decision Table Majority rule based on the KDDCUP'99 dataset. The important achievement of this paper was the reduction of false alarm rate in the intrusion detection system and improving its efficiency.

Guo et al. [7] have proposed a new and easy-to-implement hybrid learning method named distance sum-based support vector machine (DSSVM). By applying DSSVM to the KDDCUP'99 dataset, the results obtained show that the proposed method performs well in both the detection rate and the computational costs.

Moussaid et al. [8], firstly, did a pre-processing phase for normalizing each TCP connection, and then the SVM technique was applied to the KDD KDDCUP'99 dataset to reduce the number of features. Finally, the K-Means algorithm was used to test the performance of the chosen attributes. The results obtained showed that choosing 10 features by SVM had a better performance.

Aziz et al. [9] have developed a multi-layer hybrid machine-learning method. This method consists of three layers: at first, the principal component analysis (PCA) is used for feature selection; and then the genetic algorithm (GA) is used for generating the anomaly detectors; and finally, several different classifiers including Naïve Bayes, multi-layer perceptron neural network, and decision trees are used. The results obtained demonstrated that the Naïve Bayes classifier had a better accuracy in the case of the U2R and R2L attacks, while the j48 decision tree classifier had a better accuracy in detecting the DOS and Probe attacks.

Ihsan et al. [10] have discussed different normalization techniques and their effect on different classifiers such as the Naïve Bayes classifier. The results obtained illustrate that the hybrid normalization performs better than the conventional normalization techniques.

Xia et al. [11], at first, created an efficient train dataset using the K-Means and Ant Colony algorithms, and then the effectiveness of four different feature selection methods including Feature removal method, Sole feature method, hybrid method for feature selection, and Gradually Feature Removal method (GFR) by the SVM classifier was evaluated. The results obtained showed that the GFR method performed better than the others.

Mukherjee et al. [12] have investigated the performance of four different feature selection methods using Correlation-based Feature Selection, Information Gain, Gain Ratio, and Feature Vitality-Based Reduction Method by performing the Naïve Bayes classifier on the reduced dataset. The results of this research work show that the selected attributes by Feature Vitality Based Reduction Method gives a better intrusion detection performance.

3. Materials and methods

In this section, we describe the dataset and algorithms used in this research work.

3.1. Dataset and data pre-processing

Since KDD CUP'99 is the most commonly used dataset for simulating intrusion detection [1], we will use 10% of it in our experiments. Each record in this dataset includes 41 features and a class label. The features are listed in table 1, and the class labels can be categorized into 5 classes: normal, Denial of Service (DOS), unauthorized access from a remote machine (R2L), User to Root (U2R), and probe. Data pre-processing is the first step in the data analyzing procedure. This phase includes different methods like removing repeated data, normalization, and discretization. Here, we will describe the pre-processing methods that are used in this paper, as what follow.

What one notes is that there are a lot of duplicate records in the KDD cup99 dataset that may cause

biased results of classifiers towards more frequent records, and so their elimination is a necessity for achieving more accurate results. By removing duplicate records, the size of dataset is reduced from 494,021 to 145,586 records. Furthermore, each dataset consists of different attributes describing records. These features are qualitative or quantitative with different ranges of values and influence on the data analysis process. However, normalization can eliminate this effect by scaling data into a specific range.

Table 1. Network data features.

#	Network data feature	#	Network data feature	#	Network data feature
1	Duration	15	su_attempted	29	same srv rate
2	protocol type	16	num_root	30	diff srv rate
3	Service	17	num_file creations	31	srv diff host rate
4	Flag	18	num shells	32	dst host count
5	src_byte	19	num_access_files	33	dst_host_srv_count
6	dst_byte	20	num_outbound_cmds	34	dst_host_same_srv_rate
7	Land	21	is_host_login	35	dst_host_diff_srv_rate
8	wrong_fragment	22	is_guest_login	36	dst_host_same_src_port_rate
9	Urgent	23	Count	37	dst_host_srv_diff_host_rate
10	Hot	24	srv_count	38	dst_host_serror_rate
11	num_failed_login	25	serror_rate	39	dst_host_srv_serror_rate
12	logged_in	26	srv_serror_rate	40	dst_host_rerror_rate
13	num_compromised	27	rerror_rate	41	dst_host_srv_rerror_rate
14	root_shell	28	srv_rerror_rate	42	Class label

In this paper, a hybrid normalization technique combining a probability function for qualitative attributes and Mean Range Normalization for quantitative attributes is used to transform their values in the range of [0-1]. (For more details, see [10].) In order to illustrate this technique, suppose that X, which is a qualitative attribute, takes on the {a, b, a, a, b, a, b} values, where N = 7. The probability function for the values of X is known as follows [10]:

$$f_x(x) = \Pr(X = x) = \Pr(\{s \in S : X(s) = x\}) \quad (1)$$

Thus for instance, f_x (a) = 4/7 and f_x (b) = 3/7. Moreover, Mean Range Normalization is used for the quantitative attributes [10]. It is defined as (2):

$$x_i = \frac{v_i - \min(v_i)}{\max(v_i) - \min(v_i)} \quad (2)$$

- $\mathbf{v_i}$: current value of an attribute
- $\mathbf{Min(v_i)}$: minimum value of that attribute
- $\mathbf{Max(v_i)}$: maximum value of that attribute

Therefore, all the qualitative and quantitative attributes values would be in the range of [0-1].

3.2. Construction of small training dataset

This paper aims to make an efficient train dataset using clustering and feature selection algorithms, as discussed in the following sub-sections.

3.2.1. K-Medoids clustering

Due to the fact that the K-Medoids algorithm is robust and not sensitive to noise and outlier values [13], we employed it to create a new train dataset. K-Medoids is a famous clustering algorithm, which is used to break the dataset up into the groups based on what follows [13]:

- Select k of the n instances randomly as the medoids for the initial clusters.
- Assign each data instances to the closest medoid to generate the initial clusters.
- Repeat the following steps until the cluster membership stabilizes.
- Find the most central point of each cluster.
- Re-assign each data to the closest medoid selected in the earlier step.

In this work, since the U2R and R2L attack patterns are so similar to normal instances, we elected k = 3 to cluster the dataset into three groups. Then we selected the most similar data in each cluster.

3.2.2 Feature reduction strategy

Feature reduction strategy is the process of finding and choosing a useful subset of features. Finding an optimal feature selection method is so important [14]. In this paper, to make an efficient dataset, Selecting Feature using SVM [8] algorithm performs on the new dataset created by the above steps. Table 2 also shows the selected features by this algorithm.

Table 2. Selected features by Feature selection using SVM method.

Method	Features
Feature selection using SVM	2, 3, 4, 5, 6, 8, 13, 22, 23, 24.

3.4. Naïve Bayes Classifier

Naïve Bayes classifier, known as a conditional probability model, is one of the most useful and efficient learning algorithms. This method works based on the Baye's theorem and also a strong assumption that is defined as Conditional Independence and supposes that the probability of one feature does not have any effect on the probability of the other ones [15].

4. Performance evaluation Metrics

There are three performance metrics that were utilized for measuring the efficiency of algorithms in this work.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (3)$$

$$Detection\ Rate = \frac{TP}{(TP + FP)} \quad (4)$$

$$False\ Alarm\ Rate = \frac{FP}{(FP + TN)} \quad (5)$$

- True positive (**TP**): Number of samples that are correctly classified as attacks.
- True negative (**TN**): Number of normal samples that are correctly classified as normal.
- False positive (**FP**): Number of normal samples that are incorrectly classified as attacks.
- False negative (**FN**): Number of attack samples that are incorrectly classified as normal.

5. Results and discussion

The total procedure of our work is illustrated in figure 1. All the experiments were produced WEKA 3.6 toolkit. We created a train dataset by K-Medoids clustering and Feature selection using the SVM method. Subsequently, its performance was measured by the Naïve Bayes classifier. In order to evaluate the proposed hybrid method, it was compared with three other methods based on K-Medoids and GFR feature selection method, K-Medoids without feature selection, and the most famous method namely 10-fold cross-validation. Tables 3 and 4 show confusion matrices associated with them, respectively.

As depicted in table 3, the proposed method obtains better results in detecting the DOS attack and also a normal behavior.

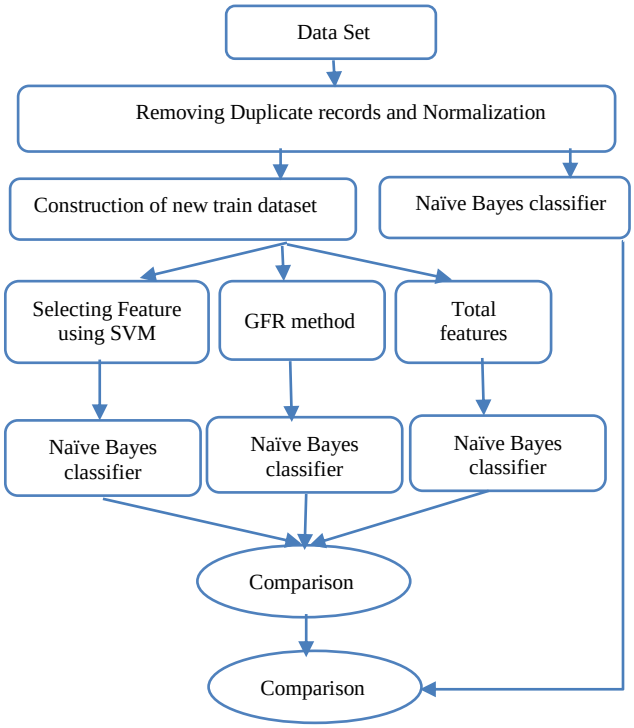

Figure 1. Proposed method procedure.

Table 3. Confusion matrix obtained by proposed method.

	DOS	Normal	Probe	U2R	R2L	Accuracy
DOS	49380	2232	2920	0	40	**90.5**
Normal	113	82242	4881	499	97	**93.6**
Probe	26	878	1224	0	3	57.4
U2R	0	561	56	378	4	37.8
R2L	0	27	0	0	25	48.1

Table 4 represents the confusion matrix obtained by K-Medoids, GFR, and the Naïve Bayes classifier. It can be observed that this method performs better in terms of detecting Probe U2R and also the R2L attacks.

Table 4. Confusion matrix obtained by utilizing K-Medoids, GFR and Naïve Bayes classifier.

	DOS	Normal	Probe	U2R	R2L	Accuracy
DOS	41968	1396	9909	873	426	76.9
Normal	2	80867	3029	1525	2409	92.1
Probe	0	253	1247	388	243	**58.5**
U2R	0	52	49	860	38	**86.1**
R2L	0	10	0	13	29	**55.8**

Various algorithms have different abilities in detection of normal and abnormal behaviours. Table 5 shows the performance of the mentioned methods regarding the accuracy and detection rate. As shown in table 5 and also figure 2, the proposed method outperforms the others in terms of accuracy, detection rate, and false alarm rate.

Table 5. Comparison between accuracy and detection rate.

	Accuracy (%)	Detection rate (%)	False alarm rate
Proposed method	**91.5**	**90.1**	**6.36**
K-Medoids+ GFR+Naïe Bayes	85.8	86.36	7.92
K-Medoids+Total features+Naïve Bayes	85.1	85.05	8.76

As shown in table 5, the proposed method is superior to the others.

Figure 2. Comparison between detection rate and accuracy among proposed method, K-Medoids+GFR+Naïve Bayes, and K-Medoids+Naïve Bayes.

Table 6 represents the results across accuracy, detection rate, and also false alarm rate, which are obtained from 10-fold cross-validation Naïve Bayes classifier and our proposed hybrid learning approach. It can be found that the proposed method performs better in relation to accuracy, detection rate, and false alarm rate.

Table 6. Comparison between accuracy and detection rate.

	Proposed hybrid learning approach	10-fold cross-validation+Naïve Bayes
Accuracy (%)	**91.5**	90.3
Detection rate	**90.1**	82.7
False alarm rate	**6.36**	13.13

And finally, in table 7, the improvement in our method is specified.

6. Conclusion and future work

In this paper, we proposed a hybrid learning approach through a combination of K-Medoids clustering, Selecting Feature using SVM, and also Naïve Bayes classifier. The KDD CUP'99 benchmark dataset was used for evaluation. The experimental results obtained showed that our proposed approach was an efficient one. In this method, a new training dataset is created by K-Medoids clustering and Selecting Feature using SVM. Then its performance is evaluated by the Naïve Bayes classifier. The results obtained showed that the proposed method performed well in terms of accuracy, detection rate, and also false alarm rate. An interesting aspect that can be developed in the future is to consider a hybrid approach that performs better in detecting the R2L, U2R, and Probe attacks. Another emphasis to put on the research work was to find a new way to choose the number of clusters and also the initial cluster medoids.

Table 7. Improvement of proposed method in comparison with others.

	K-Medoids + GFR + Naïve Bayes	K-Medoids + total features + Naïve Bayes	10-fold cross-validation + Naïve Bayes
Accuracy (%)	5.7	6.4	1.2
Detection rate (%)	3.74	5.05	7.4
False alarm rate (%)	1.56	2.4	6.77

References

[1] Lin, W.-C., Ke, S.-W. & Tsai, C.-F. (2015). CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. Knowledge-based systems, vol. 78, pp. 13-21.

[2] Elejla, O. E., Belaton, B., Anbar, M. & Alnajjar, A. (2016). Intrusion Detection Systems of ICMPv6-based DDOS attacks, Neural Computing and Applications, pp. 1-12.

[3] Aslahi-Shahri, B., Rahmani, R., Chizari, M., Maralani, A., Eslami, M., Golkar, M., et al. (2015). A hybrid method consisting of GA and SVM for intrusion detection system, Neural Computing and Applications, vol. 27, no. 6, pp. 1669-1676.

[4] Ravale, U., Marathe, N. & Padiya, P. (2015). Feature selection based hybrid anomaly intrusion detection system using K means and RBF kernel function, Procedia Computer Science, vol. 45, pp. 428-435.

[5] Esmaily, J., Moradinezhad, R. & Ghasemi, J. (2015). Intrusion detection system based on Multi-Layer Perceptron Neural Networks and Decision Tree. 7th Conference on Information and Knowledge Technology (IKT), Urmia, Iran 2015.

[6] Anita, S. C., & Gupta, S. (2015). An effective model for anomaly IDS to improve the efficiency. International Conference on Green Computing and Internet of Things (ICGCIoT), Noida, India, 2015.

[7] Guo, C., Zhou, Y., Ping, Y., Zhang, Z., Liu, G. & Yang, Y. (2014). A distance sum-based hybrid method for intrusion detection, Applied intelligence, vol. 40, no. 1, pp. 178-188.

[8] El Moussaid, N. & Toumanari, A. (2014). Overview of intrusion detection using data-mining and the features selection. International Conference on Multimedia Computing and Systems (ICMCS), Marrakech, Morocco, 2014.

[9] Aziz, A. S. A., Hassanien, A. E., Hanaf, S. E.-O. & Tolba, M. F. (2013). Multi-layer hybrid machine learning techniques for anomalies detection and classification approach, 13th International Conference on Hybrid Intelligent Systems (HIS), Gammarth, Tunisia, 2013.

[10] Ihsan, Z., Idris, M. Y. & Abdullah, A. H. (2013). Attribute Normalization Techniques and Performance of Intrusion Classifiers: A Comparative Analysis, Life Science Journal, vol. 10, no. 4, pp. 2568-2576.

[11] Li, Y., Xia, J., Zhang, S., Yan, J., Ai, X. & Dai, K. (2012). An efficient intrusion detection system based on support vector machines and gradually feature removal method, Expert Systems with Applications, vol. 39, no.1, pp. 424-430.

[12] Mukherjee, S. & Sharma, N. (2012). Intrusion detection using naive Bayes classifier with feature reduction, Procedia Technology, vol. 4, pp. 119-128.

[13] Murty, P. S. R., Murty, R. K. & Sailaja, M. (2016). Exploring the Similarity/Dissimilarity measures for unsupervised IDS. International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, India, 2016.

[14] Shahamat, H. & Pouyan, A. A. (2015). Feature selection using genetic algorithm for classification of schizophrenia using fMRI data, Journal of AI and Data Mining, vol. 3, no. 1, pp. 30-37.

[15] Kaur, R., Kumar, G. & Kumar, K. (2015). A comparative study of feature selection techniques for intrusion detection. 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2015.