I.J. Mathematical Sciences and Computing, **2016, 4, 34-40**
Published Online November 2016 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijmsc.2016.04.04

***Available online at** http://www.mecs-press.net/ijmsc*

A Hybrid Approach based on Classification and Clustering for Intrusion Detection System

Jasmeen K. Chahal [a], Asst. Prof. Amanjot Kaur [a]

[a] *Department of CSE & IT, BBSBEC, Fatehgarh Sahib, 140406, India*

Abstract

Computer security plays an important role in everybody's life. Therefore, to protect the computer and sensitive information from the untrusted parties have great significance. Intrusion detection system helps us to detect these malicious activities and sends the reports to the administration. But there is a problem of high false positive rate and low false negative rate. To eliminate these problems, hybrid system is proposed which is divided into two main parts. First, cluster the data using K-Mean algorithm and second, is to classify the train data using Adaptive-SVM algorithm. The experiments is carried out to evaluate the performance of proposed system is on NSL-KDD dataset. The results of proposed system clearly give better accuracy and low false positive rule and high false negative rate.

Index Terms: Intrusion Detection System, high false positive rate, false negative rate, K-Mean, Adaptive-SVM, NSL-KDD.

1. Introduction

Computer Security is a critical issue in every sector. The online information exchange increases the threat of stealing the information. Also, many malicious programs like viruses, Trojan horses are installed online, which may alter the information or program on a computer. These malicious programs are called intrusions. Intrusion is any programs or activity which affects the integrity, confidentiality and availability of a resource.

Data integrity refers to ensure that the data cannot be changed by unauthorized person during transit. Confidentiality refers to protect the sensitive information from reaching the wrong people and puts that sensitive information to the right people. It is equivalent to privacy. Availability refers to guarantee to the reliable access to the information by authorized people [10].

Intrusion Detection System is a system which detects the intrusion by monitoring the system. IDS monitor & analyze the network or system and producing reports to the administration either malicious activity happens

* Corresponding author.
E-mail address: jasmeenkaurchahal@gmail.com, amanjot.kaur@bbsbec.ac.in

or not. It produces alerts to the administration about the malicious activity. FPR (False Positive Rate) and FNR (False Negative Rate) are two main aspects which evaluate the performance of Intrusion Detection System. False positive refers to no intrusion but IDS generates an alert. False Negative refers to a real intrusion but IDS never generates any kind of alerts. There are two kind of intrusion detection techniques namely anomaly based and signature based.

In Anomaly based technique, the user's behavior is stored in a database. If the current user's behavior is not similar and very different from the activities stored in the database, then the particular activity is said to be abnormal. This technique has the capability of finding new attacks.

In Signature based technique only well-known attacks are identified which affect the vulnerabilities of the system. In this technique, there is a library which is a collection of abnormal activities done by intruders.

Data mining is a process of analyzing data from large amount of data and discovering interesting patterns and useful information [11]. In IDS, information comes from various sources like data, network log data, alarm massages etc. The data sources are too complex and network traffic is very huge, therefore, the data analysis is very hard. Data mining has the capability to extract information from large databases. Therefore, it has great significance to use data mining techniques in intrusion detection system. By using data mining techniques, a model of IDS is obtained which helps to detect the abnormal and normal data. In the proposed system, two data mining techniques classification and clustering are used to detect the anomalous data. K mean clustering alone is not reliable to classify the data normal or anomalous. Therefore, we proposed a hybrid approach with adaptive SVM algorithm to increase the performance of intrusion detection system.

2. Related Work

S. Duque & Omar [2] proposed a K-Mean clustering on NSL-KDD dataset. The algorithm is applied on different five clusters. The best results are obtained when 22 clusters were used.

Also K-Mean clustering is used in hybrid approaches, like B. Sharma and H. Gupta [3] uses two techniques association rule & clustering. Apriori & K-Mean is used to detect the intrusions. The experiment is done on KDD'99 dataset. The performance measures are execution time (120ms), CPU Usage (74%) and memory usage (54%).

[4] and [5] proposed hybrid approach of classification and clustering. Ravale and Nilesh et al. [4] proposed hybrid approach of K-Mean and RBF kernal function of SVM. The accuracy result of the hybrid approach is 93% and detection rate is 95%. Where, Chao and Wen et al. [5] proposed hybrid approach of K-Mean and K-NN. The accuracy result is better i.e. 99% in this work. Both hybrid approaches uses KDD'99 dataset.

Liang and Nannan et al. [7] proposed a system which is combination of K-Mean and Fuzzy C Mean (FCM) algorithms to eliminate false positive from the dataset DARPA 2000. The conclusion of the work is the effect of FCM algorithm is better than that of K-Mean clustering.

Zhengjie and Yongzhong [8] proposed hybrid approach of K-Mean & particle Swarm Optimization technique (PSO-KM). The detection rate of known attacks is 75.82% and of unknown attacks is 60.8%.

To improve the performance of SVM, Horng and Yang et al. [9] hybrid SVM with hierarchical clustering. The BRICH hierarchical clustering algorithm is used for feature selection procedure to eliminate unimportant features from dataset so that SVM classify the data more accurately. The accuracy rate of proposed system is 95.72% and false positive rate is 0.7%.

3. Proposed System

Literature review represents that many researchers done research on hybrid approaches of data mining to detect the intrusions and every hybrid approach has different accuracy, false positive and false negative rate. The proposed work is a combination of supervised and unsupervised approaches. K- Mean and adaptive SVM is supposed to give better solution to detect the anomalous data. Following is the description of proposed work:

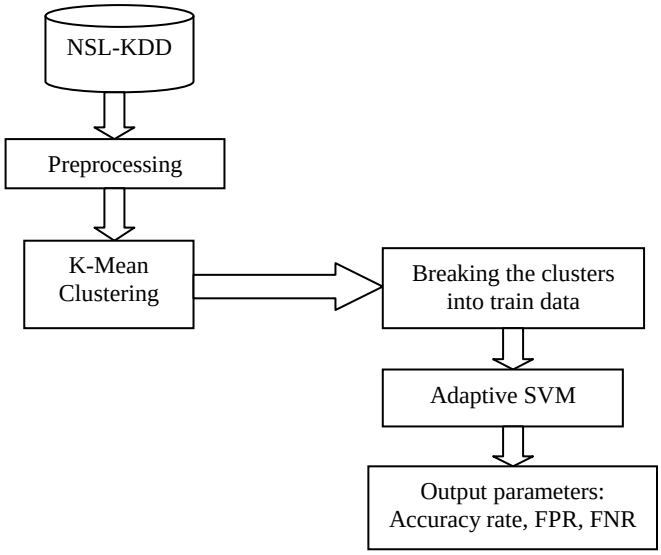

Fig.1. The Proposed Framework

3.1. Dataset Description:

NSL-KDD dataset [12], is extended version of KDD'99 dataset. There are 41 attributes in each record and it is consists of labels either it is a normal or anomalous. There are five classes of network connection vectors and they are categorized as one normal class and four attack classes. The 42nd attribute of the record represents the label either normal or abnormal. The four attack classes are further grouped as DOS, Probe, R2L, U2R [1] which are described below:

Table 1. Attack Type with Attack Classes

Attacks in Dataset	Attack Type
DOS	Worm ,Back, Apache2, Land, Processtable, Neptune, Pod, Smurf, Teardrop, Udpstorm, (10)
Probe	Sa int ,Satan, Portsweep, IPsweep, Nmap, Mscan (6)
R2L	Named, Warezmaster, Guess_Password, Ftp_write, Imap, Snmpgetattack, Phf, Multihop, Warezclient, Spy, Xlock, Xsnoop, Snmpguess, Httptunnel, Sendmail, (16)
U2R	Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps (7)

3.2. Preprocessing

Due to the difference between the format of data, it is necessary to preprocess it like to convert the character data into numeric data. In NSL-KDD dataset, three attributes are symbolic. These are:

1. Protocol_type: Defines the protocol used in the connection (e.g. TCP, UDP).
2. Service: defines which destination network service used (e.g. telnet, FTP).
3. Flag: defines the status of the connection (e.g. SF, REJ).

3.3. K-Mean Clustering

K-Mean Clustering [2] [3] [4], is a technique which groups the similar data based on the behavior. K- Mean

is an unsupervised task, i.e. data doesn't specify what we are trying to learn. Many researchers use K-Mean clustering in the hybrid approaches to detect the anomalous data. In proposed system, K-Mean clustering works as a pre-classification phase which groups objects based on the feature value into number of disjoint clusters.

Algorithmic steps are:

Step 1: Choose the number of centroids objects from dataset as the initial centroids.
Step 2: Then, calculate the Euclidean distance between each data point and the centroids.
Step 3: If the data point is closest to the centroid, then leave it and do not make any change in its position. But if the data point is not closest to the centroid, then move it to its closest one.
Step 4: Recalculate the centroid of both modified clusters.
Step 5: Repeat step 3 until we get the steady centroids.

In other words, its objective is to find [4]:

$$M = \sum_{a=1}^{k} \quad \sum_{b=1}^{n} \quad d_{ab}\ (x_b\ , y_a\ ) \tag{1}$$

Where, $d_{ab}\ (x_b,\ y_b)$ is an eculidean distance between the data point x_b and the centroid y_a.
Euclidean distance is:

$$d(x_b\ , y_a) = \parallel\ x_b\ - y_a\ \parallel \tag{2}$$

3.4. Adaptive SVM

Adaptive SVM (Support Vector Machine) aims to adapt two or more classifiers of any kind to new datasets [16]. The problem is how to select the best classifier for adaptation. The solution to this problem is to select the classifier with best parameters after estimating the performance of each classifier on the sparsely labeled dataset. The general problem of binary classification task is considered on original dataset D^o, which made up of majority of unlabeled instances D_u^o and limited number of labeled instances D_l^o,therefore, the original dataset is:

$$D^o = D_l^o\ \cup D_u^o$$

There are one or more subordinate datasets D_1^s……,D_M^s which is different from the original dataset. The subordinate classifier f_k is used to train each of the subordinate datasets D_k^s. We have,

$$D_l^o = \{(x_i\ , y_i)\}_{i=1}^{N}$$

where, x_i is the i_{th} data vector and $y_i \in \{-1, +1\}$ is its binary label. Data vector x always include a constant 1 as its first element, such that, $x_i \in R^{d+1}$, where d is the number of features. There exist multiple subordinate datasets as D_1^s, D_M^s with $D_k^s = \left\{\left(x_i^k, y_i^k\right)\right\}_{i=1}^{N_k}$, where $x_i^k \in R^{d+1}$ and $y_i^k \in \{-1, +1\}$. The subordinate dataset description is different from the original dataset. The subordinate classifier $f_k^s(x)$ has been trained from each subordinate dataset D_k^s, which gives us the result of prediction of data label through the sign of its decision function, i.e. $\hat{y} = f^x(x)$.

The traditional SVM trains the $f(x)$ from the labeled dataset D_l^o. Adaptive SVM is used to adapt a combination of multiple existing classifiers $f_1^s(x)$ …… , $f_M^s(x)$ to the new classifier.

The traditional SVM trains the $f(x)$ from the labeled dataset D_l^o. The decision boundary is determined by the kernel function $(x\ , x') = \langle \Phi(x), \Phi(x')\ \rangle$, where $\Phi(x)$ is a feature vector. The kernel function is the inner product of two projected feature vectors. Delta function is used in adaptive SVM in the form of $\Delta f(x) =$

$w^T\ \Phi(x)$ on the basis of $f^s(x)$:

$$f(x) = f^s(x) + \Delta f(x) = f^s(x) + w^T\ \Phi(x) \tag{3}$$

where, w are the parameters predicted from the labeled data D_l^o. As defined earlier, the objective is to make a group of subordinate classifiers and adapt this group to new classifier $f(x)$. By using Eq.(3), the adapted classifier's form is:

$$f(x) = \sum_{k=1}^{M} t_k\ f_k^s(x) + \Delta f(x) = \sum_{k=1}^{M} t_k\ f_k^s(x) + w^T\ \Phi(x) \tag{4}$$

where, $t_k \in (0,1)$ is the weight of each subordinate classifier $f_k^s(x)$, which sums to one as $\sum_{k=1}^{M} t_k = 1$.

Objective Function: To learn the parameter w of delta function Δf, the function is:

$$\min_w \frac{1}{2} ||\ w\ ||^2 + C \sum_{i=1}^{N} \xi_i \tag{5}$$

Such that, $\xi_i \geq 0$

$$y_i f^s(x_i) + y_i\ w^T \Phi(x_i) \geq 1 - \xi_i, \forall (x_i, y_i) \in \mathrm{D_l^o}$$

where, $\sum_{i=1}^{N} \xi_i$ measures total classification error of adapted classifier $f(x)$ and $||\ w\ ||^2$ is a regulization term that is inversely related to margin between training examples of two classes. The cost factor C in A-SVMs balances the contribution between the subordinate classifier (through the regularizer) and the training examples. The larger C is, the smaller the influence of the auxiliary classifier is.

The objective function of adaptive SVM model which is able to leverage multiple subordinate classifiers $(f_1^s(x) \ldots \ldots \ldots f_M^s(x))$, is achieved by replacing $f_1^s(x)$ with $\sum_{k=1}^{M} t_k f_k^s(x_i)$ in Eq.(5):

$$\min_w \frac{1}{2} ||\ w\ ||^2 + C \sum_{i=1}^{N} \xi_i$$

Such that, $\xi_i \geq 0$,

$$y_i \sum_{k=1}^{M} t_k f_k^s(x^i) + y_i\ w^T \Phi(x_i) \geq 1 - \xi_i, \forall (x_i, y_i) \in \mathrm{D_l^o}$$

4. Results

The experimental results are evaluated from the proposed framework in Fig.1. on NSL-KDD dataset. The hardware requirements used by proposed system are 2.8 GHz processor, 2 GB RAM, 200 GB Hard disk and implementation tool is MATLAB R2013a. As described in section 2.1 NSL-KDD dataset has 42 attributes in each record and 42nd attribute is labeled either normal or attack. As in [2], the accuracy of 22nd cluster is highest among all the clusters. Table 2. And Fig. 2. shows that the accuracy of the proposed system is much more as compared to individual data mining technique. Performance measures of proposed system are:

Accuracy = TP+TN/TP+TN+FP+FN
FNR = FP/FN+TN
FPR = FP/FP+TN

where,

FNR – False Negative Rate
FPR – False Positive Rate
TP – True Positive
TN – True Negative
FP – False Positive
FN – False Negative

Table 2. Comparison of K-Mean and Proposed Hybrid Work

Performance parameters	**Approaches**	
	K-Mean	**Hybrid (K-Mean, Adaptive SVM)**
Accuracy	81.61%	98.47%
FPR	4.03%	0.53%
FNR	98.14%	76.63%

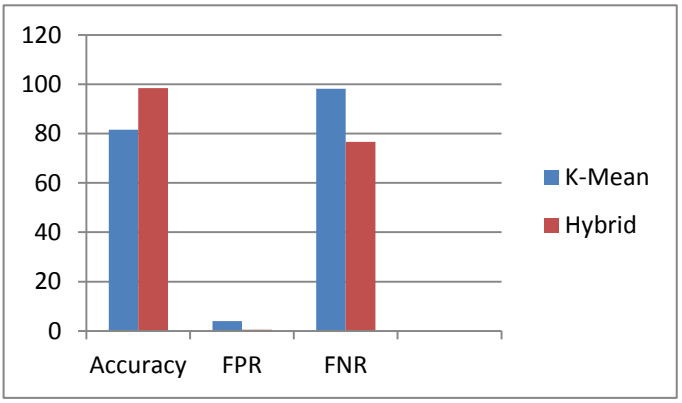

Fig.2. Graphical Analysis of results of K-Mean and Adaptive SVM

5. Conclusion

Detection of attacks is important issue in computer security field. Therefore, there should be a reliable approach to improve the performance of intrusion detection system. Single data mining technique is not sufficient to improve the performance of IDS. Therefore, we propose a hybrid approach of clustering and classification. K-mean clustering is used to make the clusters of dataset. Then, adaptive SVM is used to classify the train set. NSL-KDD dataset is used to evaluate the results. The results shows that the accuracy of hybrid approach (K-Mean, Adaptive SVM) is better than K-Mean technique.

References

[1] L. Dhanabal, S.P. Shantharajah, "A study of NSL-KDD Dataset for Intrusion Detection System based on Classification Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, Vol.4, Issue 6, pp. (446-452), June 2015.

[2] S. Duque, N.B Omar, "Using data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)", Proceedings of Science direct: Procedia Computer Science 61, pp. (46-51), 2015.

[3] B. Sharma and H. Gupta, "A design and Implementation of Intrusion Detection System by using Data Mining", IEEE Fourth International Conference on Communication Systems and Network Technologies, pp.700-704, 2015.
[4] U. Ravale, M. marathe, P. Padiya, "Feature Selection based Hybrid Anomoly Intrusion Detection System using K Means and RBF Kernal Function", Proceedings of Science Direct: International Conference on Advanced Computing Technologies and Applications (ICACTA), pp. 428-435, 2015.
[5] W. C. Lin, S. W. Ke, C. F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors", Proceedings of Science direct: Knowledge-Based Systems, pp. 13-21, 2015.
[6] J. Haque, K.W. Magld, N. Hundewale, "An Intelligent Approach for Intrusion Detection based on Data Mining Techniques", Proceedings of IEEE, 2012.
[7] Liang Hu, Taihui Li, Nannan Xie, Jiejun hu, "False Positive Elemination in Intrusion Detection based on Clustering", IEEE International Conference on Funny System and Knowledge Discovery (FSKD), pp. 519-523, 2015.
[8] Zhengjie Li, Yongzhong Li, Lei Xu, "Anomoly Intrusion Detection Method based on K-Means Clustering Algorithm with Particle Swarm Optimization", IEEE International Conference of Information Technology, Computer Engineering and Management Sciences, pp. 157- 161, 2011.
[9] S. J. Horng, M.Y. Su, Y. H. Chen, T. W. Kao, R. J. Chen, J. L. Lai, C. D. Perkasa, "A novel intrusion detection system based on hierarchical clustering and support vector machines", Proceedings of Science direct: Expert Systems with Applications, pp. 306-313, 2011.
[10] Whatistarget.com/definition/confidentiality-integrity-and-availability-CIA
[11] J. Han, M. Kamber, "Data Mining: Concepts and Technnologies", Third Edition.
[12] http://nsl.cs.unb.ca/NSL-KDD/
[13] Dae-Ki Kang and Doug Fuller et al., "Learning Classifiers for Misuse and Anomaly Detection Using a Bag of System Calls Representation", IEEE Workshop on Information Assurance and Security United States Military Academy (2005).
[14] K. Shivshankar E., "Combination of Data Mining Techniques for Intrusion Detection System", IEEE International Conference on Computer, Communication and Control (IC4-2015).
[15] Jain Patik P and Madhu B.R., "Data Mining based CIDS: Cloud Intrusion Detection System for Masquerade attacks [DCIDSM]", IEEE 4th ICCCNT (2013).
[16] J.Yang, R.Yan, A.G.Hauptman, "Cross-Domain Video Concept Detection Using Adaptive SVMs", Proceedings of ACM, MM'07, Augsburg, Bavaria, Germany, 2007.

Authors' Profiles

Jasmeen K. Chahal is presently pursuing her Master degree in E- Security from BBSBEC, FGS. She holds a Bachelor degree in Computer Science Engineering from BBSBEC, FGS in 2014. Her research focus is on Data Mining and Intrusion Detection System.

How to cite this paper: Jasmeen K. Chahal, Amanjot Kaur,"A Hybrid Approach based on Classification and Clustering for Intrusion Detection System", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.2, No.4, pp.34-40, 2016.DOI: 10.5815/ijmsc.2016.04.04