IOSR Journal of Computer Science (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727
PP 34-39
www.iosrjournals.org

Intrusion Detection Model Based on Data Mining Technique

Sadia Patka
Computer Science and Engineering, Anjuman College of Engineering and Technology, India

ABSTRACT :*Intrusion Detection System (IDS) is becoming a vital component of any network in today's world of Internet. IDS are an effective way to detect different kinds of attacks in an interconnected network thereby securing the network. An effective Intrusion Detection System requires high accuracy and detection rate as well as low false alarm rate. Most of the previously proposed methods suffer from the drawback of k-means method with low detection rate and high false alarm rate. This paper presents a hybrid data mining approach for IDS encompassing feature selection, filtering, clustering, divide and merge and clustering ensemble. The main research method is clustering analysis with the aim to improve the detection rate and decrease the false alarm rate. A method for calculating the number of the cluster centroid and choosing the appropriate initial cluster centroidis proposed in this paper. The IDS with clustering ensemble is introduced for the effective identification of attacks. The KDD CUP 1999 data set is used to test the performance of the model. Experimental results shows that the system achieves high detection rate and low false alarm rate as compared to others existing methods.*

Keywords - *Intrusion detection system, data mining, clustering, k-means, ensemble, detection rate, false alarm rate*

I. INTRODUCTION

Security of network systems is becoming an important issue, as network attacks have increased in number over the past few years. It is essential to find an effective way to protect it as more and more sensitive information is being stored and manipulated online. The network based attacks can also be referred as some kind of intrusion. An intrusion can be defined as "any set of actions or a type of attack that attempt to compromise the confidentiality, availability, or integrity availability of a resource". For controlling intrusions, intrusion detection systems are introduced.

An Intrusion Detection System (IDS) [1] is a defense system that plays an important role to protect or secure a network system and its main goal is to monitor network activities automatically to detect malicious attacks. Intrusion detection system (IDS) is increasingly becoming a vital and critical component to secure the network in today's world of Internet.

Intrusion Detection Systems are divided into two types [2] according to the detection approaches: Misuse Detection and Anomaly Detection.

1.1 Misuse detection

Misuse detection first build pattern for malicious behavior and then identify intrusion based on this known pattern i.e. it finds intrusions by looking for activity corresponding to known techniques for intrusions.

The main advantage of misuse detection is its higher detection accuracy to all known attack. The shortcoming of this approach is that it can only detect intrusions that follow predefined patterns.

1.2 Anomaly detection

Anomaly detection defines the expected behavior of the network or profile in advance. Any significant deviations from such defined expected behavior are reported as possible attacks. But not all such deviations are attacks.

The main advantage of this approach is that it can examine unknown and more complicated intrusions. The shortcoming of this approach is its low detection rate and high false alarm rate.

Between these two approaches, [3,4,5] only anomaly detection has the ability to detect unknown attacks, since misuse detection can only detect intrusions which contain known patterns of attack.

Clustering techniques [6] can be useful for detecting intrusions from network data, since clustering methods can discovers complex intrusions over a different time period. Clustering is an unsupervised machine learning mechanism for discovering patterns and deals with unlabeled data with many dimensions.It is the process of assigning the data into groups of similar objects and each group is called as cluster. Each group consists of members from the same cluster that are similar and members from the different clusters are different from each other**.**

The remaining part of this paper is organized as follows. In Section 2, the related work in IDS is discussed. Section 3 presents the proposed method. Experimental results are given in Section 4. Finally, Section 5 concludes the paper.

II. RELATED WORK

Anomaly based IDS have the ability to detect new attacks, as any attack will differ from the normal activities. In order to detect attacks, a number of clustering based detection methods has been proposed.

K-means [7] is one of the simple portioning algorithms that solve the clustering problem. The procedure of K-means algorithm follows a very simple and easy way to classify a given data set through a certain number of k clusters that are fixed a prior.

A clustering algorithm that uses SOM and K-Means [8] for intrusion detection was proposed in which when the SOM finish its training process, K-means clustering is adopted to refine the weights obtained by training, and when SOM finish its cluster formation, K-means is applied to refine the final result of clustering.

A parallel clustering ensemble algorithm [9] for IDS achieve the high speed, high detection rate and low false alarm rate. The parallel clustering ensemble is based on evidence accumulation algorithm. The evidence accumulation combines the results of multiple clustering into a single data partition, and then detects intrusions with PEA algorithm.

A hybrid learning approach [10] by using a combination of K-means and naive bayes classification, cluster all data into the corresponding group before applying a classifier for classification purpose.

A hybrid anomaly detection system [11] was proposed which combine k-means, and two classifiers: k-nearest neighbor and naive bayes. Firstly, it performs the feature selection process from intrusion detection data set using an entropy based feature selection algorithm which selects the important attributes and removes the redundant attributes. The next step is cluster formation using k-Means and then it further classifies them by using a hybrid classifier.

This paper presents a clustering algorithm for unsupervised anomaly detection. The proposed method is based on K-means clustering, which is a typical clustering algorithm. It overcomes the drawbacks of K-means thereby dealing with noise and outliers and generates the strategy for calculating the number of the cluster centroid and choosing the appropriate initial cluster centroid automatically. To improve the performance of IDS and to achieve high accuracy and detection rate as well as low false alarm rate an Intrusion Detection System with clustering ensemble is proposed.

IOSR Journal of Computer Science (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727
PP 34-39
www.iosrjournals.org

III. Proposed Method

This section describes the proposed system architecture and algorithm for intrusion detection system (IDS) based on hybrid data mining techniques.

3.1 System Architecture

Clustering is a process of labeling data and assigning that data into groups of similar objects. Each group is called as cluster. It consists of members from the same cluster that are similar and members from the different clusters that are different from each other.

K-means is one of the simplest unsupervised learning clustering algorithms. Its procedure follows an easy way to classify a given data set through a certain number of k clusters that are fixed a priori. First k center locations (c1,……,ck) are initialized. Then each data point xi is assigned to its nearest cluster centre ci. Each cluster centre ci is updated as the mean of all data point xi that has been assigned closest to it. The positions of the k centers are recalculated until the centers no longer move.

The proposed method is based on K-means clustering, which is a typical clustering algorithm. It overcomes the drawbacks of K-means thereby employing a hybrid approach.

Fig. 1 depicts the system architecture for intrusion detection. It consists of feature selection, filtering, clustering, divide and merge, clustering ensemble and normal and intrusion detection.

Feature selection is important if the data set consist of large number of attributes. It consists of selecting features using an information gain feature selection method which selects the important attributes from the data set. A filter method is proposed to reduce the noise and isolated points on the data set. It calculates the sum of the distance of each point from every other point and also calculates the sum of the average distance. For any point, if sum of the distance is greater than the average distance then that point is considered as an outlier and it is removed from the data set.

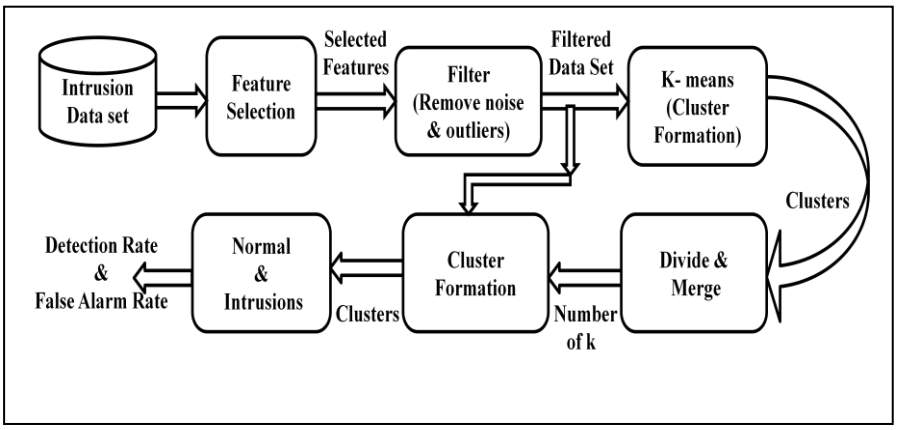

Fig.1. Proposed System Architecture

After applying filtering, initially the clusters are formed using K-means algorithm. The clusters that are formed by running the K-means algorithm are divided and merged again. By dividing and merging the clusters the number of k cluster centroids is calculated.The density of each point is calculated in filtered dataset to choose the appropriate initial centroids. These points are sorted as their density in descending order. Then the k points with the larger density are selected as the initial centroids.Again the clusters formation is done on the data set which is noise free using the calculated numbers of k cluster and the initial cluster centroids. Since the single clustering algorithm is difficult to get the great effective detection,

the clustering ensemble is introduced by varying the value of k for the effective identification of attacks to achieve high accuracy and detection rate as well as low false alarm rate.

3.2 Algorithm

Combining the above studies, the anomaly detection method based on improved K-means algorithm is proposed as follows:

Step 1: Scan all the records in dataset A, and calculate D_i and H of each record.
Step 2: For each record x_i in dataset A, if $D_i > H$, then label record x_i as outlier.
Step 3: Delete outliers in dataset A, a new dataset B with no noise is obtained.
Step 4: Select the initial value of k, let k=M.
Step 5: In dataset B, run algorithm K-means and obtain k clusters.
Step 6: Merge and divide the clusters.
Step 7: Output the value of k.
Step 8: For each record x_i in dataset B , Calculate the sum of the distances from point x_i to every other point.
Step 9: Sort all the records in set B by their sum in descending order.
Step 10: Choose the k num of points having highest density as initial centroids.
Step 11: Perform Cluster Formation again to detect normal and anomaly records.

IV. Experiment and Results

4.1 Performance Measures

The KDD CUP 99 10% intrusion dataset is used by the proposed method to test the performance of the model. It contains 4,94,021 instances and 41 features. The data set contains total of 5 classes – normal, Denial of service (Dos), User to Root(U2R), Remote to User(R2L), Probe. There are total of 24 attack types (connections) that fall into 4 major categories: DOS, U2R, R2L, Probe.

An Intrusion Detection System (IDS) requires high accuracy and detection rate as well as low false alarm rate. In general, the performance of IDS is evaluated in term of accuracy (AC), detection rate (DR), and false alarm rate (FAR) as in the following formula:

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (1)$$

$$\text{Detection Rate} = (TP) / (TP+FP) \quad (2)$$

$$\text{False Alarm Rate} = (FP) / (FP+TN) \quad (3)$$

Table 1 shows the categories of data behavior in intrusion detection for normal and intrusions (attacks) in term of true negative, true positive, false positive and false negative.

Table 1: General Behavior of Intrusion Detection Data

Actual	**Predicted Normal**	**Predicted Attack**
Normal	TN	FP
Intrusions	FN	TP

- True positive (TP) means attack data detected as attack.
- True negative (TN) means normal data detected as normal.
- False positive (FP) means normal data detected as attack.
- False negative (FN) means attack data detected as normal.

4.2 Performance Evaluation and Comparison

The experimental results show that the proposed hybrid method outperformed the compared algorithms. It achieves the detection rate of 96.53% and 0.3 false alarm rate for improved k-means and the detection rate of 97.32% and 0.2 false alarm rate for ensemble of improved k-means.

Table 2 represent the results across all category classes obtained from the proposed hybrid approach, Improved K-Means and Ensemble Improved K-means using KDD CUP 99 10% data set.

Table 2: Detection Rate Obtained By Proposed Hybrid Approach

Class	**Improved K-means**	**Ensemble Improved K-means**
Normal	97.99 %	98.99 %
DOS	97.97 %	98.97 %
U2R	96.07 %	98.03 %
R2L	93.33 %	93.33 %
Probe	97.29 %	97.29 %

Table 3 depicts the comparison of proposed approach with existing approaches in terms of detection rate. The comparison shows that the proposed approach detect better percentage of attacks than the existing approaches.

Table 3: Comparison in Terms of Detection Rate

Approaches	**Normal**	**DOS**	**U2R**	**R2L**	**Probe**
K-means	87.99 %	87.98 %	86.27 %	86.66 %	86.48 %
K-means +KNN	95.87 %	97.14 %	92.80 %	87.09 %	93.10 %
K-means+ KNN + Naive Bayes	96.03 %	98.43 %	95.15 %	92 %	97.67 %
Ensemble Boosted Decision tree	99.56 %	99.91 %	21.15 %	90.23 %	93.13 %
Improved K-means	97.99 %	97.97 %	96.07 %	93.33 %	97.29 %
Ensemble Improved K-means	98.99 %	98.97 %	98.03 %	93.33 %	97.29 %

V. CONCLUSION

On the basis of the previous clustering method it can be concluded that none of the existing clustering methods has high detection rate and very low false alarm rate. Hence a hybrid data mining approach for intrusion detection system is proposed.The main research method is clustering analysis with the aim to achieve high detection rate and very low or zero false alarm rate.

The proposed method which is a hybrid approach consists of feature selection which selects the important attributes from the data set. A filter method helps in reducing noise and outliers on the data set. Divide and merge helps in calculating the k number of the cluster centroids. By the more accurate method of finding initial k clustering centers, the intrusion detection model with clustering ensemble is presented that achieve high accuracy and detection rate as well as very low false alarm rate.

The experimental results show that the proposed method outperformed the compared algorithms. It achieves the detection rate of 96.53% and 0.3 false alarm rates for improved k-means and the detection rate of 97.32% and 0.2 false alarm rates for ensemble of improved k-means.

REFERENCES

[1] V. K. Pachghare, Parag Kulkarni, Deven M. Nikam, "Intrusion Detection System Using Self Organizing Maps", In Proceedings of IAMA 2009, IEEE, 2009.
[2] Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir, "Intrusion Detection based on K-Means Clustering and OneR Classification", In Proceedings of 7th International Conference on Information Assurance and Security (IAS), IEEE, 2011, pp.192-197.
[3] Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir, "Intrusion Detection based on K-Means Clustering and Naïve Bayes Classification", In Proceedings of 7th International Conference on IT in Asia (CITA), IEEE, 2011.
[4] Shaik Akbar, Dr.K.Nageswara Rao, Dr.J.A.Chandulal, "Intrusion Detection System Methodologies Based on Data Analysis", In International Journal of Computer Applications (0975 – 8887) Volume 5– No.2, August 2010, pp.10-20.
[5] Deepthy K Denatious, Anita John, "Survey on Data Mining Techniques to Enhance Intrusion Detection", In Proceedings of International Conference on Computer Communication and Informatics (ICCCI -2012), Jan. 10 – 12, 2012, Coimbatore, INDIA, IEEE.
[6] Kapil Wankhade, Sadia Patka, Ravindra Thool, "An Overview of Intrusion Detection Based on Data Mining Techniques", In Proceedings of 2013 International Conference on Communication Systems and Network Technologies, IEEE,2013, pp.626-629.
[7] Yang Zhong, Hirohumi Yamaki, Hiroki Takakura, "A Grid-Based Clustering for Low-Overhead Anomaly Intrusion Detection", IEEE, 2011, pp.17-24.
[8] WANG Huai-bin, YANG Hong-liang, XU Zhi-jian, YUAN Zheng, "A clustering algorithm use SOM and K-Means in Intrusion Detection" In Proceedings of 2010 International Conference on E-Business and E-Government, IEEE, 2010, pp.1281-1284.
[9] Hongwei Gao, Dingju Zhu, Xiaomin Wang, "A Parallel Clustering Ensemble Algorithm for Intrusion Detection System'' In Proceedings of 2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science, IEEE, 2010, pp.450-453.
[10] Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir, "Intrusion Detection based on K-Means Clustering and Naïve Bayes Classification", In Proceedings of 7th International Conference on IT in Asia (CITA), IEEE, 2011.
[11] Hari Om, Aritra Kundu, "A Hybrid System for Reducing the False Alarm Rate of Anomaly Intrusion Detection System", In Proceedings of 1st Int'l Conf. on Recent Advances in Information Technology (RAIT-2012),IEEE, 2012.