ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING

National Conference on Advanced Innovation in Engineering and Technology (NCAIET-2015)

Alva's Institute of Engineering and Technology, Moodbidri

Vol. 3, Special Issue 1, April 2015

A Clustering Algorithm for Intrusion Detection using Hybrid Data Mining Technique

Vibha Rao

M.Tech CNE from Shree Devi Institute of Technology, Mangalore, India

Abstract: Intrusion Detection System (IDS) is a becoming a necessary component of any network in today's world of Internet. It is an important detection that is used as a countermeasure to preserve data integrity and system availability from attacks. The main reason for using data mining classification method for Intrusion Detection System is due to the enormous volume of existing and newly appearing network data that require processing. Data mining is the best option for handling such type of data. This paper focuses on a hybrid approach for intrusion detection system (IDS) based on data mining techniques. Clustering analysis is required to improve the detection rate and decrease the false alarm rate. Most of the previously proposed methods suffer from the low detection rate and high false alarm rate. This paper uses hybrid data mining approach that contains feature selection, filtering, clustering, divide and merge and clustering ensemble. The IDS with clustering ensemble is introduced for the effective identification of attacks to achieve high accuracy and detection rate as well as low false alarm rate.

Keywords: Intrusion detection system;datamining;false alaram rate; KDD CUP99 data set;detection rate.

I. INTRODUCTION

During the past few years, security of computer network has become main stream in most of everyone's lives. Today, most discussions on computer security is centered on the tools or techniques used in protecting and defending networks. Intrusion detection (ID) is the unrelenting active attempts in discovering or detecting the presence of intrusive activities. ID [1] relates to computers and network infrastructure encloses a broader scope and it also refers to all processes used in discovering unauthorized uses of network or computer devices. 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 attempts to compromise the confidentiality, availability, or integrity availability of a resource". For controlling intrusions, intrusion detection systems are introduced.

Data mining [2] is the process of discovering interesting knowledge such as patterns, anomalies and significant structures from large amounts of data stored in databases or warehouse. Data mining helps [10] to extract useful hidden information from the data warehouses. It also helps in predicting the feature trends and help the user to take the knowledge based decisions. Data mining methods are applied to the data in order to detect intrusion and non – intrusion behavior.

Task in the data mining categorized into summarization, classification, clustering, association and analysis. Summarization mainly deals with generalization of data. Classification determines the different attributes for the objects. Association derives the relationship among the objects. Analysis finds the some interesting patterns from the history of an object.

Data mining techniques are well suited for the application which deals with statical method, machine learning approach and data base oriented approach and also helpful in fraud detection, managing risk and market analysis.

II. IDS TAXONOMY

IDS broadly classified into two categories which are known as Host based IDS (HIDS) and Network based IDS (NIDS).

A host-based Intrusion Detection System (HIDS) monitors a computer system on which it is installed to detect an intrusion and responds by logging the activity and notifying the designated authority. A HIDS can be thought of as an agent that monitors and analyzes the system in order to provide security to the system.

A network-based Intrusion Detection System (NIDS) monitors traffic at selected points on a network or interconnected set of networks. The NIDS examines the traffic packet by packet in real time, in order to detect intrusion patterns. NIDS examines packet traffic directed towards potentially vulnerable computer systems on a network.

Intrusion Detection Systems are divided into two types according to the detection approaches: One approach is known as Misuse Detection and another one is Anomaly detection. Misuse detection finds intrusions by looking for activity corresponding to known techniques for intrusions. 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.

DOI 10.17148/IJIREEICE

A. KDDCUP99 IDS DATASET

The KDD CUP'99 data set [8] was created by processing the tcp dump portions of the 1998 DARPR Intrusion Detection System (IDS) evaluation data set, created by Lincoln Labs, U.S.A. Since one cannot know the intension of every connection on a real world network, the artificial data was generated using a closed network. The data set contains a total of 24 attack types that fall into 4 major categories: Denial of service (DOS), User to Root (U2R), Remote to Local (R2L) and probe. Each record is labeled either as normal, or as an attack, with exactly one specific attack type.

A Denial of Service attacks temporarily interrupt or suspend services of host connected to the internet making resource either too busy or overflow.

Remote to Local attacks are the kind of intrusion attacks where remote intruder continuously send packets to local machine.

In user to root attacks hackers attempts to get the administrator privileges.

Probing is a attack in which the introducer scans the machine or network device in order to determine the vulnerabilities to compromise the system by gaining the knowledge of the configuration of the computer or network device.

In KDD CUP 99 dataset features [11] are grouped into four broad categories known as basic features, content features, time based traffic features and host based traffic features.

Basic features derived from the packet header. First six features of KDD CUP 99 dataset belongs to basic features.

Content features make use of domain knowledge in order to assess the payload of the original TCP packets. The number of failed login attempts is the example for content feature.

Time based traffic features deals with the properties that mature over a 2 second temporal window. The number of connections to the same host over the 2 second interval is the example for time based traffic feature.

Host-based traffic features make use of a historical window it is estimated over the number of connections instead of time. Host based features are therefore designed to assess attacks, which span intervals longer than 2 seconds.

III. RELATED WORK

Hybrid intrusion detection framework [3] is based on the combination of two classifiers one is Tree Augmented Naïve Bayes (TAN) and another one is Reduced Error Pruning (REP). The TAN classifier is used as a base classifier while the REP classifier is used as a Meta classifier. The Meta classification is the learning technique which is learns from the Meta data and judge the correctness of the classification of each instance by base classifier. The judgment from each classifier for each class is treated as a feature, and then builds another classifier to make the final decision.

The model for hybrid intrusion detection based on clustering and association [4] model first accepts the data from the dataset. Then based on that dataset data is divided into number of classes separately for this K-means clustering algorithm is applied. In the proposed work first need to access the data from the valid database like KDD CUP 99. Then the data is pre-processed. Pre process phase is also known as data audit phase. Because the data what we taken are not necessary support the properties of the proposed framework. Then check for redundant data so that only meaningful data going to processed. Then apply FP growth algorithm.

SVM Algorithm [5] computes clusters incrementally and to compute k-partition of a data set it uses k −1 cluster centers from the previous iteration of web log data. An important step in this algorithm is the computation of a starting point for the kth cluster center. This starting point is computed by minimizing the auxiliary cluster function. The SVM algorithm computes as many clusters as a data set contains with respect to a given tolerance.

Boosting approach [6] for intrusion detection system is proposed by making use of data mining concepts. Data mining is becoming an important component in intrusion detection system. The process of data mining consists of three stages: first is initial exploration, second one is model building and third one is deployment. This boosting approach results in balanced detection rate.

Clustering based approach for intrusion detection using fuzzy c-means [9] clustering algorithm is proposed for the incorporation of cluster features resulting from a fuzzy clustering into the training process is proved to be efficient for enhancing the strength of a base classifier.

IV. PROPOSED METHOD

Clustering is the process of grouping of similar objects. Each group is known as cluster. Clustering is the most important unsupervised learning technique it deals with finding a structure in a collection of unlabeled data. Cluster consists of members from same cluster as similar objects and members from different cluster are different from each other.

The proposed method is based on K-means Clustering algorithm. K-means clustering algorithm is the data mining technique which is used to detect the different types of attacks in a network while sending the data from

DOI 10.17148/IJIREEICE

source to destination by forming the clusters for different types of attacks. Fig.1 represents the proposed System Architecture which consists of the following blocks - feature selection, filtering, clustering, and normal and intrusion detection. The workings of the blocks are explained next. Fig.2 represents the K-means clustering algorithm. Fig.3 represents data flow of the proposed method.

Feature selection is used to describe the tools and techniques that are available in order to reduce the inputs into a manageable size so that the processing and analysis can be done easily. KDD cup'99 dataset consists of 41 features. Some examples for features in KDD cup'99 are duration, protocol type, service, flag, src_byte, dsc_byte etc. These features are going to select using information gain feature selection method which selects relevant features by removing the irrelevant features. Each record in the KDD cup'99 data set has 41 features and label assigned to each record as attack type or normal. Some example for attack types are back, buffer overflow, land, satan, smurf etc. The selected features are then filtered in order to remove noise. Filtering is done by calculating the sum of the distance of each point from every other point. Then for the filtered dataset k-means Clustering algorithm is applied to form clusters.

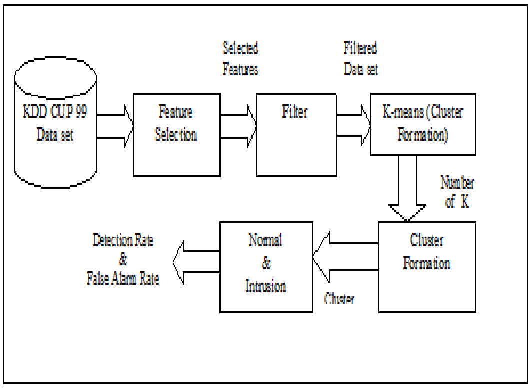


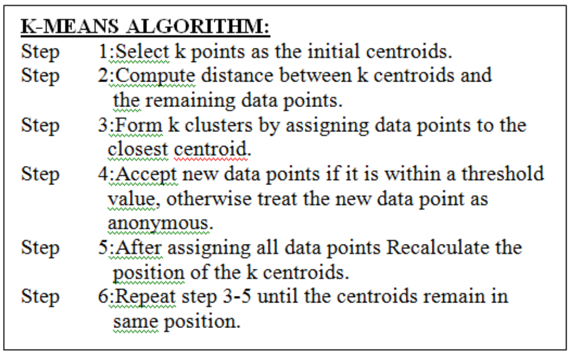
Fig. 1 Proposed System Architecture

K-means is one of the simplest learning algorithms which are used to solve the clustering problem. The working of k-means cluster is as follows [7].

K-means clustering algorithm takes set of data points and k clusters as input and places the k-centroids in random location in space. Then find the nearest centroid for each record in the dataset. This is done by making use of Euclidian distance formula to find the distance between data points and every cluster centroid. Then choose the cluster which as minimum distance to the centroid. Recalculate the centroids position by considering all the data points that belongs to cluster. The process is repeated until the centroids position is unchanged.

After applying k-means algorithm clusters are formed for normal and different types of attacks in the KDD cup'99 data set. The proposed system contains the large dataset and when new data set arrived its centroid is calculated by finding minimum distance between clusters. If the dataset not within a threshold level then treat the data set as new type of attack. If the dataset is within a threshold level then treat that dataset with respect to belonging cluster. This well helps to achieve high accuracy and detection rate by reducing false alarm rate.

K-means clustering algorithm is the simplest algorithm compare to other clustering algorithm hence it is well suited to solve the clustering issues related to large data set.



K-MEANS ALGORITHM:

```
Step    1:Select k points as the initial centroids.
Step    2:Compute distance between k centroids and
          the remaining data points.
Step    3:Form k clusters by assigning data points to the
          closest centroid.
Step    4:Accept new data points if it is within a threshold
          value, otherwise treat the new data point as
          anonymous.
Step    5:After assigning all data points Recalculate the
          position of the k centroids.
Step    6:Repeat step 3-5 until the centroids remain in
          same position.
```

Fig. 2 K-Means clustering algorithm

A. DATAFLOW DIAGRAM

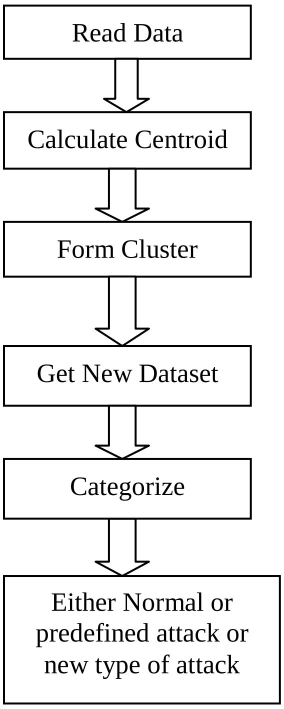


Fig. 3 Data flow diagram of proposed method

First read the data from dataset and then calculate the centroid for each dataset by making use of Euclidian distance formula. Then choose the cluster which as minimum distance to the centroid to form a cluster. For a

DOI 10.17148/IJIREEICE

new dataset centroid is calculated by finding minimum distance between data point with clusters. Then consider the cluster as normal or predefined attack or new type of attack.

V. K-MEDOIDS CLUSTERING ALGORITHM

K-medoids clustering algorithm is similar to k-means clustering algorithm but k-medoids are well suited for small dataset. In k-means clustering algorithm [12] the mean value of the objects in a cluster are calculated in order to find reference point for remaining objects in the dataset, but in k-medoids object in the center of the cluster treated as the medoid.

In K-medoids clustering algorithm the set of objects are chosen randomly to represent medoid in k clusters. Then assign the remaining objects in the dataset to the nearest medoid. Next the medoid object is replaced by non-medoid object until the quality of the cluster remains same.

By making use of k-medoids algorithm robustness can be achieved since medoids are assigned as a center of clusters. Complexity in k-medoids is more compare to k-means clustering algorithm.

VI. CONCLUSION

Intrusion detection is an area growing in relevance as more and more sensitive data are stored and processed in networked systems. On the bases of previous clustering methods none of the existing method results in low false alarm rate. In order to overcome this proposed intrusion detection system is implemented by making use of data mining concepts. The proposed method consists of feature selection, Filtering, cluster formation, clustering units, in order to achieve the high accuracy and detection rate and low false alarm rate.

REFERENCES

[1]. Guy Bruneau," The History and Evolution of Intrusion Detection", 2001

[2]. Tan, Steinbach, Kumar," Introduction to Data Mining", 2004

[3]. Maradul Dhakar and Akhilesh Tiwari, "A Novel Data Mining based Hybrid Intrusion Detection Framework", Journal of Information and Computing Science Vol. 9, No. 1, 2014, pp.037-048

[4]. Manish Somani and Roshni Dubey, "Hybrid Intrusion Detection Model Based on Clustering and Association", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol.3. Issue 3, March 2014.

[5]. Leena Bramhe and Prof. Shuwesh Shukla,"A Novel Approach for Improve Detection Rate in Anomaly based Intrusion Detection System", International Journal of IT, Engineering and Applied Sciences Research (IJIEASR), Vol. 2. No. 5 May 2013.

[6]. Snehlata S. Dongre and Kapil K. Wankhade, "Intrusion Detection System Using New Ensemble Boosting Approach", International Journal of Modeling and Optimization, Vol. 2, No. 4, August 2012.

[7]. Mrutyunjaya Panda, Manas Ranjan Patra,"Some Clustering Algorithms to Enhance the Performance of the Network Intrusion Detection System", Journal of Theoretical and Applied Information Technology, 2005-2008 pp.795-801.

[8]. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani,"A Detailed Analysis of the KDD CUP 99 Data Set", Proceeding of the 2009 IEEE symposium on computation intelligence in security and defense Application.

[9]. Huu-Hoa Nguyen, Nouria Harbi and Jerome Darmont, "An Efficient Clustering –Based Classification Approach for Intrusion Detection".

[10]. Mrs. Nidhi Singh, Mrs. Nitya Khare, "Efficient Data Mining Techniques To Enhance Intrusion Detection System", International Journal of Latest Research in Science and Technology, Volume 3, Issue 4: Page No.122-125. July-August 2014.

[11]. H. Günes Kayacık, A. Nur Zincir-Heywood, Malcolm I. Heywood, "Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets".

[12]. Dr.T.Velmurugan, "Efficiency of K-means and K-medoids Algorithms for Clustering Arbitrary Data Points". Int.T.Computer Technology & applications, Vol 3 (5), 1758-1764.