IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-4, Special Issue-7, Dec 2016 ISSN: 2347-2693

Intrusion Detection System Based on Modified K-Means Clustering Algorithm

Nipjyoti Sarma
GIMT, Guwahati-17, Assam
nipjyoti_cse@gimt-guwahati.ac.in

Sabyasachi Roy
GIMT,Guwahati
sabyasachi95roy@gmail.com

Jyoti Nath
GIMT,Guwahati
jnath3151@gmail.com

Ashapurna Sarma
GIMT,Guwahati
ash.sarma93@gmail.com

Himakshi Bora
GIMT,Guwahati
himabora71@gmail.com

Available online at: www.ijcseonline.org

***Abstract*- Due to the growth of Information Systems, different types of electronic attacks are happening day by day. This leads to the security breach rising every day Therefore it is of utmost important to protect highly sensitive and private information by securing the data. An intrusion detection system (IDS) monitors network or system activities and for nasty activities produces reports to a management. It monitors network traffic and its suspicious behaviour against security. Different types of intrusion detection methodologies are available, but all the current IDS are not perfect. Now a day's Data mining concepts are used in the area of research in intrusion detection implementation. This paper tries to forward an idea of modifying the traditional K- means algorithm using fuzzy concept to prepare a model of intrusion detection system. The experiments have been done on the KDD Cup 99 dataset.**

Keywords— IDS, Data mining, KDD Cup, Clustering, Fuzzy, False Positive

I. INTRODUCTION

We can see rapid advancement of technologies in our daily life. With this rapid growth the cyber crime rates continues to increase. The valuable data always attract intruders and is liable for maximum attack on the network system [1]. To reduce the risk of network security from intruders we are introducing Intrusion Detection System. In this paper we are using data mining algorithms for intrusion detection. The main purpose of Intrusion Detection System (IDS) is to identify the objects that are attempting to manipulate the database such as web servers or entire network system. Data mining is the latest technology introduced in network security environment to find regularities and irregularities in large datasets [2].

We need a techniques to detect the intrusion both known and unknown types. Thus we need to introduce a technique that can detect the data that are departing from the normal ones. Anomaly detection does this. It also detects the new type intrusion [3].Clustering techniques are used for anomaly detection in this paper. It is an anomaly detection method. It is the process of grouping the objects according to their similar behaviour or patterns. The objects that are dissimilar are grouped in a different group. What it produces is the partial optimal solution, but is not the globally optimal solution [4]. Thus we need a modified K-Means algorithm for the satisfactory result. Improving K-means algorithm is the simple and easy approach to analysis the given set of data. In this paper KDD99 dataset is used for testing the algorithm. By using datasets from KDD99 we built a system which can detect the abnormal data by creating clusters. From these clusters we can detect the anomalous data.

In our work we choose a clustering approach to detect the intrusion which can find out the normal and attack data effectively.

II. LITERATURE REVIEW

Since we are dealing with large number of datasets, using data mining techniques will be more efficient to detect intrusion. Applying different techniques of data mining, we can achieve the detection methods for intrusion detection.

Various data mining techniques such as classification, two-stage techniques, clustering are used frequently to work on intrusion detection. In this section we are describing the various techniques on data mining that detects intrusion.

A. Classification

Classification is a data mining techniques which takes each instance of a data set and assigns it to a particular class [5]. It is a technique to detect the class data as normal and abnormal data. Analysing the data and listing those data which are in similar sequences in a column are normal class data and those which are dissimilar are labelled as abnormal class data.

B. Clustering

Clustering is the technique of grouping a set of data in such a way that objects of similar behaviour grouped in a cluster and dissimilar objects in another cluster.

K-means clustering is one of the most simplest techniques to solve clustering problem. The main goal is to utilize K-means clustering approach is to split and to group data into normal and attack instance[2]. It works on a dataset D which contains n objects and partitions these objects in k

clusters. This methods starts with selecting n objects from D. It calculates the cluster center by using the mean value of the objects in each cluster.

Graph based clustering method is particularly suited for dealing with data that do not come from a Gaussian or a spherical distribution[6]. Considering the records of dataset as note and these notes as vertex of complete undirected graph. The value of the distance between these notes as weight of the edges can be calculated by [4]Euclidian distance. The described method in [7] mainly focus on how to classify the data on the boundary to be classified more accurately.

C. *Two-Stage Technique*

Two- stage [8] technique using SOM(self-organising map) with K-means and Neural gas with FCM to minimize the large volumes of detected alters in first stage and the second stage is to reduce the rate of false attack.

III. THE PROPOSED ALGORITHM

We assume a dataset of objects X and is defined by a set of attributes {A_1, A_2, A_3, A_4 ,.., A_m}, where, m is the dimension of the dataset. Each object P_i is defined by P_i =(P_{i1}, P_{i2},.....,P_{im}). The domain of P_i is P_i $\epsilon D_1 x$ D_2 x.....x D_m with D_i is the domain of attribute A_i. So, any object P_i = (p_{i1}, p_{i2},.....,p_{im}) means having attribute values $A_{i1}=p_{i1}$, $A_{i2}=p_{i2}$ and so on up to m dimensions, i.e. $A_{im}=p_{im}$. Any two objects P_i and P_j with $A_{ik}= A_{jk}$ for $1 \leq k \leq m$, does not mean that both P_i and P_j are same.
There may be n number of objects in the dataset X each having m dimensions. So, set of X can be defined by X={P_1, P_2, P_3, P_4 ,....., P_n}. We have to consider a cluster $C_i \subseteq X$ and an object $P_i \subseteq X$ and find out the similarity between them. The cardinality of the cluster C_i is given as $|C_i|$.
If there is an attribute A_i with Domain D_i in a Cluster C_i, then a fuzzy set $\tilde{A}_i$ in D_i for A_i can be defined as[9]
$\tilde{A}_i$ ={d, μ_{Ai}(d),| dϵ D_j}----------(1)
where, μ_{Ai}(d) is the membership function for the attribute A_i and is defined as
$\mu_{Ai}(d) = \frac{freq(d)}{|Ci|}$
Where, freq(d)= $\sum_{l=1}^{|Ci|} 1$ such that p_{li} = d
Here 0$\leq$ $freq(d) \leq |Ci|$ and hence, 0 $\leq$ μ_{Ai}(d) $\leq$ 1
For objects the above definitions would take the form

$\tilde{A}_i$ ={d, μ_{Ai}(d),| dϵ D_j}----------(1)
where, μ_{Ai}(d) is the membership function for the attribute A_i and is defined as μ_{Ai}(d) = freq(d)
Where, freq(d)= 1 if p_{li} = d or 0 otherwise.
Here, $freq(d) = 0 \; or \; 1$ and $\mu_{Ai}(d) = 0 \; or \; 1$
centroid of a cluster C_i having mixed type attribute as based on above definition of fuzzy membership function, termed as fuzzy set vector(FSV_i) of it and is as given below:
FSV_i = {V_{i1},V_{i2},V_{i3}.....V_{im}}
Where m is the total number of attributes and V_{ij} is defined as:

$$V_{ij} = \begin{cases} mean_i(A_j), if\ A_j\ is\ numerical\ attribute \\ \tilde{A}_i\ , if\ A_j\ is\ categorical\ attribute \end{cases}$$

----------(2)
Where for numerical attribute A_j the mean of the j^{th} attribute of cluster C_i is
$Mean_i(A_j) = \frac{1}{|Ci|} \sum_{k=1}^{|Ci|} p_{kj}$
And for categorical attribute the A_j, $\tilde{A}_i$ is calculated by using equation (1) above.
Let FSV_i is the fuzzy Centroid frequency vector for clusters C_i. Let y is the total number of dimensions of categorical attributes in the data-objects. Then the similarity measure between, j^{th} categorical attribute of p_k and j^{th} categorical attribute of $\widetilde{FSV}_{ij}$ is calculated by using the following equation.[9][10][11]

$$\partial_c(p_{kj}, FSV_{ij}) = \frac{(\tilde{p}_{kj} \cap \widetilde{FSV}_{ij})}{(\tilde{p}_{kj} \cup \widetilde{FSV}_{ij})} \quad \text{----------(3)}$$

Where $(\tilde{p}_{kj} \cap \widetilde{FSV}_{ij})$ and $(\tilde{p}_{kj} \cup \widetilde{FSV}_{ij})$ means the set intersection and set union operations between two fuzzy sets $\tilde{p}_{ij}$ and $\widetilde{FSV}_{ij}$.
Objects containing both categorical and numerical attributes can be inserted in one cluster by finding out the similarity between the cluster Centroid and the object and by adding both the similarity of numeric and categorical part.

IV. NEW ALGORITHM DESCRIPTION

Input: the number of clusters k, of data object n, first h data object.
Output: the number of cluster k, centre cluster i W which make the smallest values of similarity function.
(a) Initialize the k centre points
(b) Prepare K fuzzy set vector from the k points where k= 1,2 ,3 ,.....
(c) Set the data objects not participated in clustering in the nearest cluster fuzzy set vector by counting the the similarity functions using equation (2) and (3) above.
(d) Repeat the above steps(from a) to c)) to update the fuzzy set vector of the clusters already made.
(e) Repeat the above steps (from a) to d)) until improvement remaining same for the new cluster's fuzzy set vector.

V. EXPERIMENTAL RESULTS

In order to evaluate the proposed algorithm first of all we have downloaded the kdd cup 99 data set[12] . It is the most wildly used data set for the evaluation of anomaly detection methods. This data set is prepared by Stolfo et al. and is built based on the data captured in DARPA'98 IDS evaluation program. DARPA'98 is about 4 gigabytes of compressed raw (binary) tcpdump data of 7 weeks of network traffic, which can be processed into about 5 million connection records, each with about 100 bytes. The two

weeks of test data have around 2 million connection records. KDD training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labelled as either normal or an attack, with exactly one specific attack type. The computer-generated attacks fall in one of the following four categories: Denial of Service Attack (DoS), User to Root Attack (U2R), Remote to Local Attack (R2L), and Probing Attack [13]. For the convenient of simulation of experiments, this paper adopts 5 groups' of10000 sample records which are selected randomly. The normal data record of each sample is about 98%.We have selected only 15 kinds of essential attributes of the data to form clustering, out of which 12 are numerical and other three are categorical attributes. In the experiment we have only evaluate the fault detection rate to measure the test result. It is given in table 1[14].

TABLE 1

Sample Subset	Record No.	Fault Detection Rate %
Sample 1	10,000	72%
Sample 2	10,000	75.4%
Sample 3	10,000	80%
Sample 4	10,000	82.6%
Sample 5	10,000	84.2%

According to the table given above, we can observe that using our new and this improved clustering algorithm to execute intrusion detection can effectively separate out the normal data and the attack data, therefore so it has the very high possibility of detection and the algorithm is strength as after giving so many input also its is sound. The complexity is very less which is in the order of input dataset O(n).It removes the numerical only limitations of the k means algorithm and the experiment proved that the algorithm the paper proposed is effective.

VI. CONCLUSION

Use of clustering algorithm in the detection of fault in the network intrusion system is very useful as it can concentrate mainly in the abnormal detection rate. Traditionally k means clustering algorithm has been used in the area but due to the robustness and finding the means only it is not possible to apply this algorithm directly in the intrusion data as it is collection of heterogamous data. Therefore several mixed type clustering approaches are used already to detect intrusion in the IDS area. We have also proposed a novel idea whereby clustering all the categorical and numerical data of the dataset, this algorithm can find out the true positives, false positives in the dataset. For huge set of data this algorithm takes time to execute if input set is not pre-processed. An experimental result says that this new clustering algorithm cam effectively cluster intrusion data by overcoming the limitations of traditional k means algorithm.

REFERENCES

[1] Shakil Ghalib, Kapil Dewan "Review on Intrusion Detection Approaches with Machine Learning" ©2015,IJARCSSE All Rights Reserved.

[2] Hitesh Chandra Mahawari and Mahesh Pawar , "A Study of Various Methods to find K for K-Means Clustering", International Journal of Computer Sciences and Engineering, Volume-04, Issue-03, Page No (45-47), Mar -2016

[3]. Vidit pathak, Dr. Ananthanarayana V.S. "A Novel Multi-Threaded K-Means Clustering Approach for Intrsion Detection "978-1-4673-2008-5/12/$31.00_2012 IEEE.

[4].Li Tian, Wang Jianwen"Research on Network Intrusion Detection System Based on Improved k-means Clustering Algorithm"978-0-7695-3930-0/09 $26.00 _ 2009 IEEE.

[5] Kapil Wankhande, Sadia Patka"An Overview of Intrusion Detection Based on Data Mining Techniques"978-0-7695-4958-3/13 $26.00©2013 IEEE DOI 10.1109/CSNT.2013.134.

[6] Pasquale Foggia, Gennaro Percannella,Carlo Sansone andmario vento,"A Graph-Based Clustering Method and It's Applications"F.Male et al. (Eds.):BVAI 2007, LNCS 4729, pp.277-287,2007.©Springer.Verlag Berlin Heidelberg 2007.

[7] Zhou mingqiang, Huang Hui, Wang Qian,"A Graph-based Clustering Algorithm for Anomaly Intrusion Detection "978-1-4673-0242-5/12/$31.00©2012 IEEE.

[8] Hachi fatma, Limam mohamed, "A two-stage technique to improve intrusion detection system based on data mining algorithms"978-1-4673-5814-9/13/$31.00©2013 IEEE.

[9] Enakshmi Nandi and Debabrata Sarddar, "A Modified MapReduce-K-Means Clustering Based Load Distribution Method for Wireless Sensor Network in Mobile Cloud computing", International Journal of Computer Sciences and Engineering, Volume-04, Issue-08, Page No (107-110), Aug -2016,

[10] Nilam Kolhe, Harshada Kulkarni, Ishita Kedia and Shivani Gaikwad, "Comparative Analysis of Cluster based Boosting", International Journal of Computer Sciences and Engineering, Volume-03, Issue-10, Page No (60-70), Oct -2015

[11] LIU Hai-tao, WEI Ru-xiang and JIANG Guo-ping" Similarity measurement for data with high-dimensional and mixed feature values through fuzzy clustering" in Proceedings of IEEE conference,2012 ,pp-617-621

[12] KddCup-99 Team, The third international knowledge discovery and data mining tools competition, Available from http://kdd.ics.uci.edu/databases/ kddcup99.html, 2002

[13] Dwipen Laskar, Nipjyoti Sarma, A Review On Intrusion Detection SystemUsing Data Mining Techniques,2013, MIT publications, Moradabad , India

[14] Li Tian, Wang Jianwen, 2009 International Forum on Computer Science-Technology and Applications, 978-0-7695-3930-0/09 $26.00 © 2009 IEEE, DOI 10.1109/IFCSTA.2009.25.

BIOGRAPHY

Mr.Nipjyoti Sarma, working as an Assistant Professor at GIMT, Guwahati has published many research papers in the areas of Data Mining, Image Processing. His area of interest is in Data Mining.

Sabyasachi Roy, pursuing B.Tech in Computer Science and Engineering department from Girijananda Chowdhury Institute of Management and Technology. His area of interest is in Data Mining.

Jyoti Nath, pursuing B.Tech in Computer Science and Engineering department from Girijananda Chowdhury Institute of Management and Technology. His area of interest is in Data Mining.

Ashapurna Sarma, pursuing B.Tech in Computer Science and Engineering department from Girijananda Chowdhury Institute of Management and Technology. Her area of interest is in Data Mining.

Himakshi Bora, pursuing B.Tech in Computer Science and Engineering department from Girijananda Chowdhury Institute of Management and Technology. Her area of interest is in Data Mining.

.

