International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Implementation of Improved K-Mean Algorithm for Intrusion Detection System to Improve the Detection Rate

Susheel Kumar Tiwari[1], Dr. Manish Shrivastava[2]
[1]PhD Research Scholar, Mewar University, Chittorgarh, Rajasthan, India
[2]Professor & Head (CSE) L.N.C.T Bhopal, Affiliated to R.G.P.V Bhopal, Madhya Pradesh, India

ABSTRACT

In Data mining there are lots of methods are used to detect the outlier by making the clusters of data and then detect the outlier from them. In general Clustering method plays a very important role in data mining. Clustering means grouping the similar data objects together based on the characteristic they possess. An improved K-means clustering algorithm is put forward on basis of the split-merge method for the purpose of remedying defects both in determination of value in K and in selection of initial cluster centre of traditional K-means clustering. At first , the concept of independence degree of date was incorporated into the experimental date subset construction theory , using independence degree to evaluate the importance of nature. Next ,the database is merged into several classes in respect of density of date points ,the combination of the minimum spanning tree algorithm and traditional K-means clustering algorithm is conducive to the achievement of splitting .Eventually ,the KDD Cup99 database is applied to conduct simulation experiment on the application of the improved algorithm in intrusion detection .The results indicate that the improved algorithm prevails over traditional K-means algorithm in detection rate and false alarm rate

Key words : Intrusion Detection System, K-Mean, Clustering

I. INTRODUCTION

Computer Security is used frequently, but the content of a computer is vulnerable to few risks unless the computer is connected to other computers on a network. As the use of computer networks, especially the Internet, has become pervasive, the concept of Computer security has expanded to denote issues pertaining to the networked use of computers and their resources. The major technical areas of computer security are usually represented by the initials confidentiality, integrity, and authentication or availability. "denial of service" attacks, which are sometimes the topic of national news, are attacks against availability. Other important concerns of computer security professionals are access control and no repudiation.

The main goal of intrusion detection is to detect unauthorized use, misuse and abuse of computer systems by both system insiders and external intruders. Among automated intrusion detection systems, a particular system for network intrusion detection, known as a network-based intrusion detection system (IDS), monitors any number of hosts on a network by scrutinizing the audit trails of multiple hosts and network traffic. It is usually comprised of two main components: an anomaly detector and a misuse detector [1][2]. The anomaly detector establishes the profiles of normal activities of users, systems, system resources, network traffic and/or services and detects intrusions by identifying significant deviations from the normal behavior patterns observed from profiles. The misuse detector defines suspicious misuse signatures based on known system vulnerabilities and a security policy. This component probes whether these misuse signatures are present or not in the auditing trails.

Clustering [3]is an unsupervised method which takes a different approach by grouping objects into meaningful subclasses so that members from the same cluster are quite similar and different to the members of different cluster. If we use clustering for intrusion detection then we have in anomaly detection model which is developed based on normal data and deviations are searched over this model. It is difficult to say that there are no attacks during the time traffic collected from the network. The unsupervised anomaly detection algorithm clusters[4] the unlabeled data instances together into clusters using a simple distance-based metric. Once data is clustered all of the instance that appear in small clusters are labeled an anomalies because normal instances should form large clusters compared to intrusions and malicious intrusions and normal instance are qualitatively different so they do not fall into same cluster.

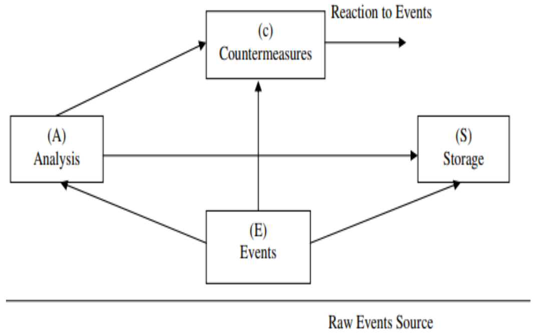

Figure 1. Common components of an Intrusion Detection Framework

II. LITERATURE SURVEY

A lot of research works have been carried out in the literature for intrusion detection and some of them have motivated us to take up this research. Brief reviews of some of those recent significant researches are presented below:

M. Jianliang, et al. [5] introduced the application on intrusion detection based on Kmeans clustering algorithm. K-means is used for intrusion detection to detect unknown attack and partition large data space effectively but it has many disadvantages degeneracy, cluster dependence.

Yu Guan, et al. [6] introduced Y-means algorithm which is a clustering method of intrusion detection. This algorithm is based on K-means algorithm and other related clustering algorithm. It overcomes two short comings of K-means, no of cluster dependency and degeneracy.

Zhou mingqiang, et al. [7] introduced a new concept of a graph based clustering algorithm for anomaly based clustering algorithm for anomaly intrusion detection. They used outlier detection method which is based on local deviation coefficient (LDCGB). Compared to other intrusion detection algorithm of clustering this algorithm is unnecessary to initial cluster number. **Chitrakar R and Huang chuanhe** [8] proposed a hybrid learning approach of combining k-medoids clustering and naive bayes classification. Because of the fact that k-medoids technique represents the real world scenario of data distribution the proposed algorithm will group the whole data into clusters more accurately than K-means such that it results in better classification.

Yang Jian [9] proposed an improved intrusion detection model based on DBSCAN which illustrates the idea of generating a cluster using density method and merging small clusters. The paper describes a more reasonable density-based clustering algorithm for intrusion detection (IIDGB), using rational method to calculate the distance and design the method of parameter selection.

Li Xue-yong and Gao Guo [10] proposed a new intrusion detection based on improved DBSCAN which uses an improved density based cluster algorithm to improve the drawbacks of earlier approach as proposed in [9] by using more rational method for calculating the distance and process of merging cluster. Some of the drawbacks which are overcome are high false alarm rate, generation of small clusters after the experiment.

Lei Li, et al. [11] introduced a novel rule-based intrusion detection system using data mining. They proposed an improvement over apriori algorithm by bringing the concept of length-decreasing support to detect intrusion. Association rules and sequence rules are the main technique of data mining.

Z. Muda, et al. [12] published new work of Intrusion detection based on K-means clustering and OneR classification; they proposed an approach which combines the techniques of K-means and OneR classification. The main goal of paper is to utilize K-means algorithm and to split and group data into normal and attack instances. The algorithm partition the dataset into k clusters according to an initial value known as seed point into each cluster's centroids or cluster centers. The mean value of each data contained within each cluster is called centroids.

Zhengjie Li, et al. [13] proposed anomaly intrusion detection method based on K-means clustering algorithm with particle swarm optimization. Particle swarm optimization (PSO) algorithm is an evolutionary computing technology which is based on swarm intelligence has good global search ability. The proposed algorithm has overcome falling into local minima and has relatively good overall convergence.

K. Wankhade, et al. [14] gave an overview of intrusion detection which is based on data mining techniques. They discussed the various data mining techniques which can be applied on intrusion detection system for the effective identification of both known and unknown pattern of attack in order to develop a secure information system.

H. Fatma and L. Mohamed [15] proposed a two stage technique to improve intrusion detection system based on data mining algorithm. They adopted a two stage technique in order to improve the accuracy of sensors. The first stage aim to generate meta-alerts through clustering and the second stage aims to reduce the rate of false alarms using a binary classification of the generated meta-alerts. For the first stage they used two alternatives, self-organizing map (SOM) with K-means algorithm and neural GAS with fuzzy cmeans algorithm and for the second stage they used three approaches, SOM with K-means algorithm, support vector machine and decision trees.

A.M. Chandrasekhar and K. Raghuveer [16] introduced a new concept of Intrusion detection technique by using K-means, fuzzy neural network and SVM classifiers. The proposed technique has four major steps: first one is to use Kmeans algorithm to generate different training subsets. Based on the obtained training subsets, different neuro-fuzzy models are trained. Then a vector for SVM classification is formed and lastly, classification using radial SVM is performed to detect intrusion has happened or not. The different approaches discussed above are proposed for intrusion detection using various techniques of data mining or other algorithm. They have many advantages over earlier approach but they are not good in all respect. In this paper we present a new approach for anomaly intrusion detection by using new medoid algorithm of K-medoid and certain modifications of it. The rest of the paper is organized as follows: section three discuss about the existing K-means algorithm in brief, section four discuss our proposed work in detail , section five discuss about the experimental methodology and results for our proposed work and finally in the last section we discuss our contributions and future work in this field.

III. PROBLEM IDENTIFICATION

The main drawback of traditional methods is that they cannot detect unknown intrusion. Even if a new pattern of the attacks were discovered, this new pattern would have to be manually updated into system. It is also capable of identifying new attacks to some degree of resemblance to the learned ones, the neural networks are widely considered as an efficient approach to adaptively classify patterns [11], but their high computation intensity and the long training cycles greatly hinder their applications, especially for the intrusion detection problem, where the amount of related data is very important.

IV. PROPOSED APPROACH

We propose Artificial Intelligence based clustering algorithm for network intrusion detection. This k-means algorithm aims at minimizing a squared error function is given in Equation for the objective function.

$$J = \sum_{i=1}^{k}\sum_{i=1}^{n}\left\| x_i(j) - c_j \right\|^2$$

Where $\left\| x_i(j) - c_j \right\|^2$ is a chosen distance measure between a data point xj (j) and the cluster centre cj is an indicator of the distance of the n data points from their respective cluster centers. One of the main disadvantages to K-Mean algorithm is that it requires the number of clusters as an input to the algorithm. The algorithm is incapable of determining the appropriate number of clusters and depends upon the user to identify this in beforehand. For example, if you had a group of people that were easily clustered based upon gender while calling the k-means algorithm with k=3 would force the people into three clusters and when k=2 would provide a more natural fit. Likewise, if a group of individuals were easily clustered based upon home state and you called the k-means algorithm with k=20 then the results might be too generalized to be effective.

But finding the value of i that best suits of data is very difficult. Hence we moved on to hill climbing. Hill climbing is good for finding a local optimum (a good solution that lies relatively near the initial solution) but it is not guaranteed to find the best possible solution (global optimum) out of all possible solutions (search space) which can be overcome by using steepest ascent Modified Hill climbing finds globally optimal solution. The relative simplicity of the algorithm makes it a popular first choice amongst optimizing algorithms and it is widely used in artificial intelligence, in order to reach a good state from a start state. Selection of next node and starting node can be varied to give a list of related algorithms. This can often produce a better result than other algorithms when the amount of time available to perform a search is limited, such as with real-time systems. Artificial Intelligence approach based Hill climbing algorithm attempts to maximize (or minimize) a target function $f(x)$ where x is a vector of continuous and / or discrete values. In each iteration, hill climbing will adjust a single element in x and determine whether the change improves the value of $f(x)$. Then, x is said to be globally optimal

Artificial Intelligence approach based Hill Climbing aided k-means Algorithm steps are shown bellow.

Input: randk - random value of kΔk - A random move in cluster

Output: k - Number of clusters Pseudo code: Modified Hill Climbing Algorithm

```
do
l1: iter =true;
 ksolved ← randk;
l2: newsolution ← ksolved + Δk;
 if (f (newsolution) < f (ksolved ) then
solution ← newsolution;
ksolved ← solution; k←ksolved;
 if (algorithm converged and globally optimum) then
 output k;
 iter = false;
else goto l2 ;
else goto l1 ;
while (iter);
```

Input: E= { e1, e2…en } - Set of entities to be clustered k - number of cluster from Modified Hill Climbing Algorithm MaxIters - Limit of iterations

Output: C= {c1, c2…cn } - Set of clustered centroids

L= {l (e) e= {1, 2…n} - Set of cluster labels of E

Pseudo code:

```
Modified Hill Climbing aided k-means Algorithm
for each ci ϵ C
do ci ← ej ϵ E (E.g. random selection);
end
for each ei ϵ E do
L (ei) ← argmin Distance (ei, ci)j ϵ {1,…, k};
end changed ← false;
iter ← 0; repeat
for each ci ϵ C do
Update cluster (ci);
End
for each ei ϵ E do
minDist ← argminDistance (ei ,cj) jϵ {1…k};
if minDist ≠ l (ei) then;
l(ei) ← minDist;
changed ← true;
end
end
iter ← iter+1;
until changed=true and iter ≤ MaxIters;
```

In the above algorithm is the best K value is obtained by modified hill climbing and this value is utilized in k– means algorithm in order to form effective clusters with uniform cluster density. The following section deals with performance evaluation of implemented system

V. RESULT

Use KDDCUP99 data packets to verify the feasibility and effectiveness of Improved Means algorithm, choose 10000 DoS attack data, of which 5500 records are used as training data for training model, and the other 1700 records are used as testing data for testing the effectiveness of the intrusion detection model. The experiment adopts different cluster number k, cluster the training data at first to get the cluster center set, and then send the testing data into the anomaly detection system for intrusion detection, calculate detection rate of each data set at the same time, experiment results are shown in Table 1. It can be seen that the network intrusion detection model based on Improved K-means is feasible, and the improved algorithm is better than traditional K-means algorithm in detection ratio and false alarm ratio based on different cluster amount.

Table 1. Clustering results with 100 clusters with time efficiency

Cluster	Algorithm	
	K-Mean	Proposed Approach
	Time (ms)	Time (ms)
20	84	65
40	82	75
60	97	85
80	110	95
100	115	101

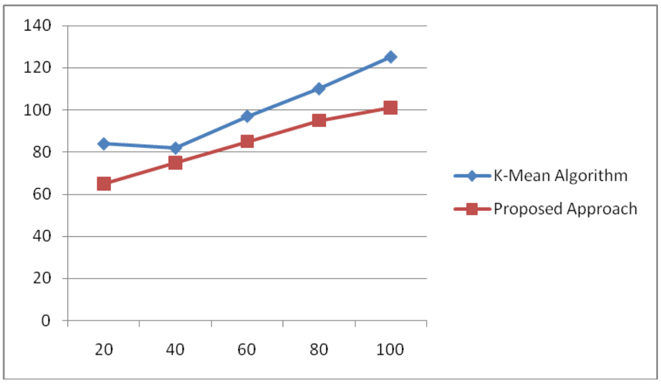

Figure 2. The number of clusters vs. Computation Time

VI. CONCLUSION

According to the characteristics of network intrusion data, aiming at the problems existed in the current intrusion detection researches, this paper proposes up a network intrusion detection method based on the fusion algorithm combining with information entropy and K-means, experiment results show that the fusion algorithm has improved the detection ratio and reduced the false alarm ratio compared with traditional K-means algorithm. However, the implementation of the fusion algorithm did not consider the algorithm execution efficiency, which requires the further study.

VII. REFERENCES

[1]. J. Anderson, "Computer security threat monitoring and surveillance", 1980.

[2]. Dorothy E. Denning, "An intrusion-detection model", IEEE Transactions on software engineering, pp. 222–232, 1987.

[3]. Kemmerer, R., and Vigna, G. "Intrusion Detection: A Brief History and Overview." IEEE Security & Privacy, v1 n1, Apr 2002, p27-30.

[4]. S. Staniford-Chen, S. Cheung, R. Crawford., M. Dilger, J. Frank, J. Hoagland, K. Levitt, C.Wee, R. Yip, D. Zerkle . "GrIDS- A Graph-Based Intrusion Detection system for Large Networks." Proc National Information Systems Security conf, 1996.

[5]. M.Jianliang, S.Haikun and B.Ling. The Application on Intrusion Detection based on K-Means Cluster Algorithm. International Forum on Information Technology and Application, 2009.

[6]. Yu Guan, Ali A. Ghorbani and Nabil Belacel. Y-means: a clustering method for Intrusion Detection. In Canadian Conference on Electrical and Computer Engineering, pages 14, Montral, Qubec, Canada, May 2003.

[7]. Zhou Mingqiang, HuangHui, WangQian, "A Graph-based Clustering Algorithm for Anomaly Intrusion Detection" In computer science and education (ICCSE), 7th International Conference ,2012.

[8]. Chitrakar, R. and Huang Chuanhe, "Anomaly detection using Support Vector Machine Classification with K-Medoids clustering" In Internet (AH-ICI), 3rd Asian Himalayas International conference, 2012.

[9]. Yang Jian, "An Improved Intrusion Detection Algorithm Based on DBSCAN", Micro Computer Information, 25,1008-0570(2009)01-3- 0058-03, 58-60,2009.

[10]. Li Xue-yong, Gao Guo- "A New Intrusion Detection Method Based on Improved DBSCAN", In Information Engineering (ICIE), WASE International conference, 2010.

[11]. Lei Li, De-Zhang, Fang-Cheng Shen, " A novel rule-based Intrusion Detection System using data mining", In ICCSIT, IEEE International conference, 2010.

[12]. Z. Muda, W. Yassin, M.N. Sulaiman and N.I.Udzir, "Intrusion Detection based on K-Means Clustering and OneR Classification" In Information Assurance and Security (IAS), 7th International conference, 2011.

[13]. Zhengjie Li, Yongzhong Li, Lei Xu, "Anomaly intrusion detection method based on K-means clustering algorithm with particle swarm optimization", In ICM, 2011.

[14]. 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. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.2, March 2014 38

[15]. H. Fatma, L. Mohamed, "A two-stage technique to improve intrusion detection systems based on data mining algorithms", In ICMSAO, 2013.

[16]. A.M. Chandrasekhar, K. Raghuveer, "Intrusion detection technique by using K-means,fuzzy neural network and SVM classifiers", In ICCCI, 2013.

[17]. Margaret H. Dunham, "Data Mining: Introductory and Advanced Topics",ISBN: 0130888923, published by Pearson Education, Inc.,2003.