iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 Issue: IV Month of publication: April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4352

www.ijraset.com
Call: 08813907089 | E-mail ID: ijraset@gmail.com

The Application on Intrusion Detection based on K-means Cluster Algorithm

Sushant Narvekar[1], Abhishek Kamble[2], Prof. Girish Wadhava[3]
[3]*Assistant Professor,* [1, 2]*Department of Information Technology Vidyalankar Institute of Technology, Mumbai, India*

Abstract: In today's scenario networking is the most essential part of the communication. Individuals can do a lot of things on the internet. Its security has been one of the most important problems in the world. Network attacks have increased over the past few years, intrusion detection system (IDS) is increasingly becoming a critical component to protect the network. In recent years, many researchers are using data mining techniques for building IDS. A wide variety of data mining techniques have been applied to intrusion detections. In data mining, clustering is the most important unsupervised learning process used to find the structures or patterns in a collection of un-labelled data. In this paper, we present an Intrusion Detection method using K-means clustering to cluster and analyse the data, Neuro-fuzzy models, Support vector machine (SVM) and C4.5 algorithm. Computer simulations show that this method can detect unknown intrusions efficiently in the real network connections.

Keywords: K-means algorithm, intrusion detection system, cluster, clustering analysis

I. INTRODUCTION

Intrusion is defined as "the act of wrongfully entering upon, seizing or taking possession of the property of another".

Network security has become a critical issue owing to the incredible growth of computer networks usage. It becomes technically hard and economically expensive for the manufactures to secure the computer systems from external attacks. A Network Intrusion Detection System (NIDS) is a device (or application) that examines network and/or system activities for malicious activities or policy violations and produces reports to a Management Station. Intrusion detection is the process of monitoring the events occurring in a computer system or network. It is used for analysing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies, acceptable use policies, or standard security practices. Intrusion detection system detects attacks and anomalies in the network, and thus are becoming very important. IDS are useful in detecting successful intrusion, and also in monitoring the network traffic and the attempts to break the security. Intrusion detection is the practice of observing and examining the actions going on in a system in order to identify the attacks and susceptibilities. 99 data set that is applied in network intrusion detection. Large amount of data can be handled with the data mining technology. It is still in developing state, as it is growing rapidly it can become more effective. Intrusion detection methods started appearing in the last few years. Using intrusion detection methods, you can collect and use information from known types of attacks and find out if someone is trying to attack your network or particular hosts. The information collected this way can be used to harden your network security, as well as for legal purposes. Both commercial and open source products are now available for this purpose. Many vulnerability assessment tools are also available in the market that can be used to assess different types of security holes present in your networks.

II. EXISTING METHODS

Intrusion Detection System is used for analysing signs of possible incidents, which creates violations or imminent threats for computer security or standard security practices. In present, there are various techniques which are working effectively in intrusion detection system.

There are various methods used in Intrusion Detection System:

1) K-Means
2) Fuzzy Logic
3) Support Vector Machine(SVM)
4) C4.5 Decision Tree

III. OVERVIEW

K-means represents a type of useful clustering techniques by competitive learning, which is also proved to be promising techniques in intrusion detection.

©IJRASET (UGC Approved Journal): All Rights are Reserved

A. K-Means

The K-means algorithm, starting with k arbitrary cluster centers in space, partitions the set of giving objects into k subsets based on a distance metric. The centers of cluster are iteratively updated based on the optimization of an objective function. This method is one of the most popular clustering techniques, which are used widely, since it is easy to be implemented very efficiently with linear time complexity. The principle goal of employing the K-Means clustering scheme is to separate the collection of normal and attack data that behave similarly into several partitions which is known as *K*th cluster centroids. In other words, K-Means estimates a fixed number of *K*, the best cluster centroid representing data with similar behavior.

The steps in the k-Means clustering-based Intrusion detection method are as follows:

1) *Step 1:* Select k random instances from the training data subset as the centroids of the clusters C1; C2;...Ck.
2) *Step 2:* For each training instance X:
 a) Compute the Euclidean distance *D(Ci,X),i = 1...k*
 b) Find cluster Cq that is closest to X.
 c) Assign X to Cq. Update the centroid of Cq.
3) *Step 3:* Repeat Step 2 until the centroids of clusters C1; C2; ...Ck stabilize in terms of mean-squared error criterion.
4) *Step 4:* For each test instance Z:

Compute the Euclidean distance *D(Ci,Z),i = 1...k.* Find cluster Cr that is closest to Z.

IV. K-MEANS ALGORITHM FOR INTRUSION DETECTION SYSTEM

A. Data pre-processing

As for continuous features, different features of raw data are on different scales. This causes bias toward some larger features over other smaller features. To solve the problem, a measurement is performed as follows: Firstly, calculate the

mean absolute deviation s_f:

$$s_f = \frac{1}{n}\left(\left|x_{1f} - m_f\right| + \left|x_{2f} - m_f\right| + \cdots + \left|x_{nf} - m_f\right|\right)$$

where $x_{1f}, x_{2f}, \ldots, x_{nf}$ are n measuring values of variants f, m_f is the mean value of the

$$m_f = \frac{1}{n}\left(x_{1f} + x_{2f} + \cdots + x_{nf}\right);$$

Secondly, calculate the standardized measurement:

$$z_{if} = \frac{x_{if} - m_f}{s_f}.$$

Then we can convert every instance in the training sets to a new one based on previous three formulas. It is a transformation of an instance from its own space to our standardized space, based on statistical information retrieved from the training sets, which can solve the problem above.

B. The application of algorithm

In order to apply the K-means algorithm to intrusion detection system, we design and realize the K-means algorithm analyse module [6],the process flow is shown in the graph :

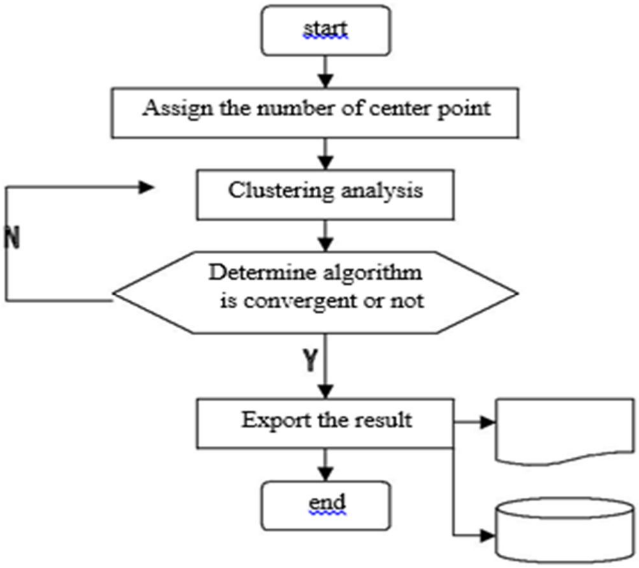

start
Assign the number of center point
Clustering analysis
N
Determine algorithm is convergent or not
Y
Export the result
end

Figure 1. The procedure flow

The output of this module includes mainly five parts:

a) Export the clustering results into the document. The results are composed of the iteration times, run time, the mean distance between each point and cluster centre, the mean distance between each point and cluster centre after clustering, the data number in each cluster and serial number in each cluster, etc.

b) Export the centre point of each cluster;

c) Show the allocation of new data, the result includes: the serial number of new data and the clustering serial number;

d) Show the allocation after adding new data. The output result include: the data number of each cluster after new data is added and the serial number in each cluster;

e) Add the new data into database.

C. Advantages of k-means

1) Simple, robust and easy to understand.

2) K-Means is computationally efficient and faster than hierarchical clustering provided large number of variables exists and k is kept small.

3) More efficient algorithm than k-mediod.

4) It gives tighter clusters than other clustering method.

D. Fuzzy Logic

Fuzzy Logic is a problem solving control structure approach that gives itself to implementation in the systems which are ranging from multichannel PC or Workstation acquisition and control systems. It can be engaged in hardware, software, or in both. It offers a simple manner to attain on a definite decision based upon indefinite, ambiguous, inaccurate, noisy, or absent input information.

E. Support Vector Machine(SVM)

Support vector machines (SVM) are learning machines that plot the training vectors in high dimensional feature space, labeling each vector by its class. SVMs classify data

By determining a set of support vectors, which are members of the set of training inputs that outline a hyper plane in feature space. Computing the hyper plane to separate the data points leads to a quadratic optimization problem. There are two main reasons that we used SVMs for intrusion detection. The first reason is that its performance is in terms of execution speed, and the second reason is scalability. SVMs are relatively insensitive to the number of data points, and the classification complexity does not depend on the dimensionality of the feature space.

F. C4.5 Decision Tree

In C4.5 first grows an initial tree using the divide-and conquer algorithm as follows:

1) If all the cases in S belong to the same class or S is small, the tree is a leaf labelled with the most frequent class in S.

2) Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets S1, S2,... according to the outcome for each case, and apply the same procedure recursively to each subset.

V. RESULTS OF THE EXPERIMENTS

The proposed method is evaluated over the KDD Cup 1999 data, which contains a wide variety of intrusions simulated in a military network environment [7] .Each sample in the data is a record of extracted features from a network connection gathered during the simulated intrusions. A connection is a sequence of TCP packets to and from various IP addresses. A connection record consists of 41 fields. It contains basic features about TCP connection as duration, protocol type, number of bytes transferred, domain specific features as number of file creation, number of failed login attempts, and whether root shell was obtained.

It provides 100,000 labelled data items, composed of 99,999 normal samples and 1,000 attack samples. In the algorithm, the number of clusters k=5 , assign the former two as the original clustering centre. The result is shown in the following table:

TABLE I. EXPERIMENT RESULTS OF K-MEANS

classes	true	false	detected rate	false alarm rate
1	87869	771	99.13%	0.87%
2	8522	10	99.88%	0.12%
3	594	5	99.81%	0.19%
4	5	1	96.15%	3.85%
5	195	8	96.06%	3.94%

The experiment results show that K-means algorithm is an efficient method for intrusion detection.

VI. CONCLUSION

In this paper, we present the K-means algorithm for intrusion detection. Experimental results on a subset of KDD-99 dataset showed the stability of efficiency and accuracy of the algorithm. With different setting, the detection rate stayed always above 96% while the false alarm rate was below 4%.The time complexity is low, which is ,N is the number objects in the database, k is the cluster number, and t is the iteration time of the algorithm. The analysis and experiment show that K-means algorithm has better global search ability. The results of simulations that run on KDD-99 data set show that the K-means method is an effective algorithm for partitioning large data set. Therefore, K-means algorithm will be widely used in intrusion detection fields in the future.

REFERENCES

[1] R.Agrawal,J.Gehrke,D.Gunopulos, and P. Raghavan.Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. Proc.ACM SIGMOD,June 1998: 94-105

[2] G. Milligan. AValidation Study of a Variable Weighting Algorithm for Cluster Analysis.J.Classification 1989: 53-71

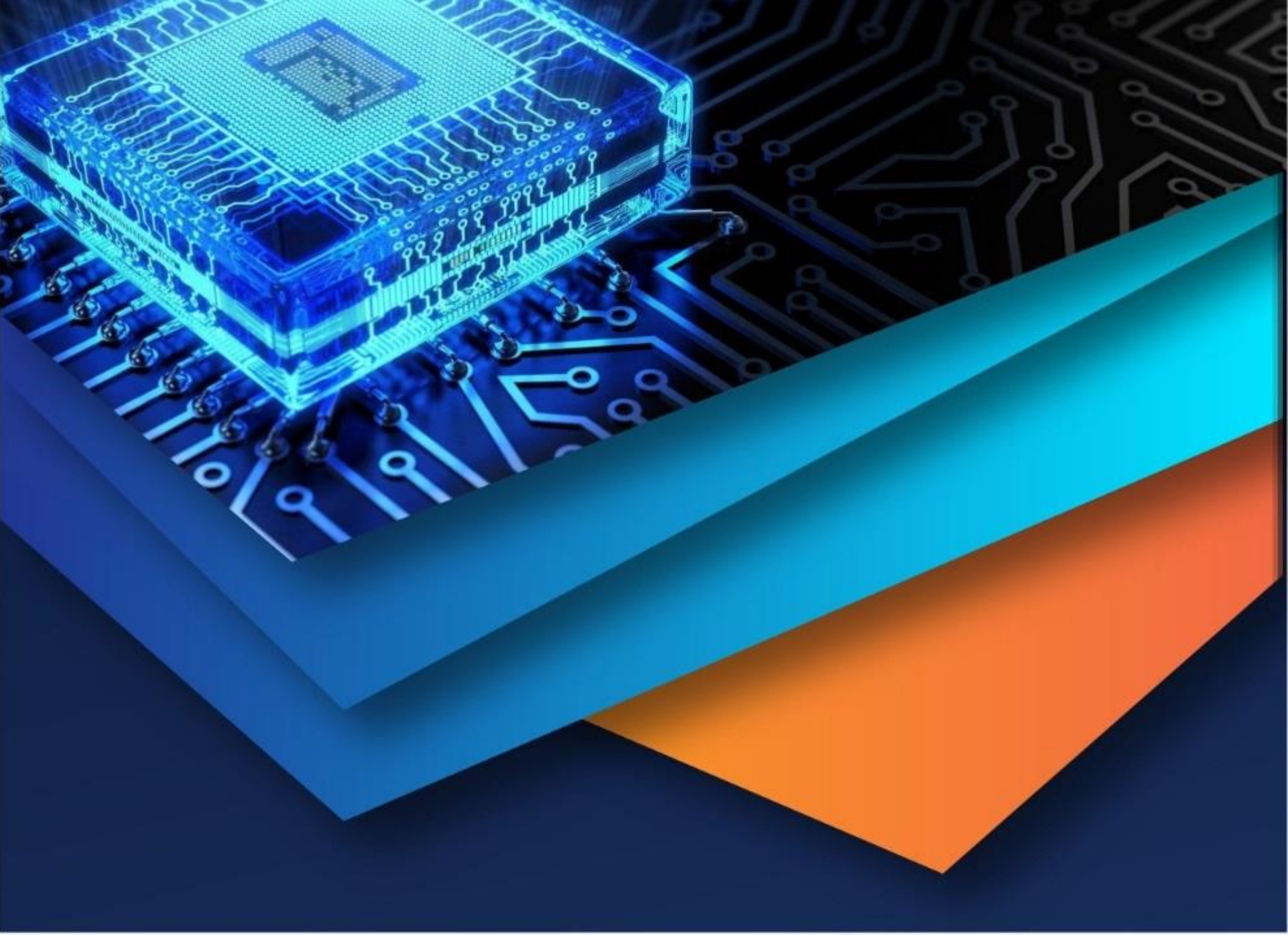
[3] Bradley,Fayyad.Refining Initial Point for K-means Clustering.Proceedings of the Fifteenth International Conference on Machine Learning,1998

[4] K. Alsabti, S. Ranka, and V. Singh.An Efficient k-means Clustering Algorithm.Proc. First Workshop High Performance Data Mining, Mar 1999

[5] Jiangtao Ren, Xiaoxiao Shi.An. Improved K-Means Clustering Algorithm Based on Feature Weighting [J].Computer Science, 2006, 33(7): 186-187

[6] Guowei Wu,Lin Yao,Kai Yao.An Adaptive Clustering Algorithm for Intrusion Detection. International Conference on Information Acquisition August 20 - 23, 2006, Weihai, Shandong, China

[7] E.Eskin,A.Arnold,M.Prerau,L.Portnoy,S.Stolfo.A.Geometric .Framw ork for Unsupervised Anomaly Detection:Detecting Intrusions in Unlabeled Data.Application of Data Mining in Computer Security,Kluwer,2002.



doi crossref

10.22214/IJRASET

INDEX COPERNICUS

45.98

JISRAF

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

ISI

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)