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

Meng Jianliang Shang Haikun Bian Ling
Dept of Computer,Computer Science and Technology College,North China Electric Power University, Hebei Baoding 071003
Shk.1984@163.com

ABSTRACT: Internet security has been one of the most important problems in the world. Anomaly detection is the basic method to defend new attack in Intrusion Detection .Network intrusion detection is the process of monitoring the events occurring in a computing system or network and analyzing them for signs of intrusions,defined as attempts to compromise the confidentiality. 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 unlabeled data. We use the K-means algorithm to cluster and analyze the data in this paper. Computer simulations show that this method can detect unknown intrusions efficiently in the real network connections.

KEYWORDS: cluster; intrusion detection;K-means algorithm;clustering analysis

I. INTRODUCTION

Along with the arrival of the information period, networked computers are playing very important roles in our daily life as well as in our business.As a result,more and more attention has been focused on the problem of internet security. Intrusion is defined as "the act of wrongfully entering upon,seizing,or taking possession of the property of another".We need effective intrusion detection systems to protect our computers from unauthorized or malicious actions.Nowadays lots of researchers turned into data mining techniques to attack the problem. Data mining can improve variants detection rate,control false alarm rate,and reduce false dismissals.Data mining based on intrusion detection systems can be roughly categorized into major two groups:misuse detection and anomaly detection.Network intrusion detection is the process of monitoring the events occurring in a computing system or network and analyzing them for signs of intrusions,defined as attempts to compromise the confidentiality. The intrusion attacks can be divided into four categories: Probe (e.g. IP sweep, vulnerability scanning), denial of service (DoS) (e.g. mail bomb, UDP storm), user-to-root (U2R) (e.g. buffer overflow attacks, root kits) and remote-to-local (R2L) (e.g. password guessing, worm attack).

Clustering is the method of grouping objects into meaningful subclasses so that the members from the same cluster are quite similar,and the members from different clusters are quite different from each other.Therefore clustering methods can be useful for classifying log data and detecting intrusions.

In an information system,the amount of normal connection data is usually overwhelmingly larger than the number of intrusions.Thus,the population of a normal cluster should be much larger than that of an intrusion cluster,and we may classify these clusters as 'normal' or 'intrusion' according to their population. 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 unlabeled data.Until now,the clustering algorithms can be categorized into four main groups:partitioning algorithm,hierarchical algorithm, density-based algorithm and grid-based algorithm.Partitioning algorithms construct a partition of a database of N objects into a set of K clusters.Usually they start with an initial partition and then use an iterative control strategy to optimize an objective function.

II. K-MEANS ALGORITHM

Clustering is the method of grouping objects into meaningful subclasses so that the members from the same cluster are quite similar,and the members from different clusters are quite different from each other. Until now,the clustering algorithms can be categorized into four main groups:partitioning algorithm,hierarchical algorithm, density-based algorithm and grid-based algorithm.Partitioning algorithms construct a partition of a database of N objects into a set of K clusters.[1,2]Usually they start with an initial partition and then use an iterative control strategy to optimize an objective function.

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

A. The idea of algorithm

Given the d-dimension data set $X = \{x_i | x_i \in R^d, i = 1, 2, \cdots N\}$ .The procedure follows a simple and easy way to classify a given data set through a certain number of clusters $W_1, W_2, \cdots W_k$, The main idea is to define *k* centroids:$C = c_1, c_2, \cdots, c_k$, one for each cluster, $c_i = \frac{1}{n_i} \sum_{x \in w_i} x$,where n_i is the number of the dataset in the cluster[3]. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to recalculate centroids as barycenters of the clusters resulting from the previous step.After we have these new

978-0-7695-3600-2/09 $25.00 © 2009 IEEE
DOI 10.1109/IFITA.2009.34

centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the centroids change their location step by step until no more changes are done. In other words centroids do not move any more.Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n_i} d_{ij}\left(x_j, c_i\right)$$

where $d_{ij}\left(x_j, c_i\right)$ is a chosen distance measure(Euclidean distance is chosen in this paper) between a data point x_j and the cluster center c_i , is an indicator of the distance of the data points from their respective cluster centers.

B. The steps of algorithm

Step 1 (Initialization): Randomly choose k instances $c_1, c_2, \cdots\cdots, c_k$ from the data set X and make them initial cluster centers of the clustering space;

Step 2 (Assignment): Assign each instance to its closest center:if $d_{ij}\left(x_i, c_j\right) < \mathrm{d}_{\mathrm{im}}\left(x_i, c_m\right)$,where $m = 1, \cdots\cdots, k;\ j = 1, \cdots\cdots, k;\ j \neq m;\ i = 1, \cdots\cdots, n$,and then assign x_i to cluster c_j ;

Step 3 (Updating): Recalculate the centroids of clusters: $c_1^*, c_2^*, \cdots\cdots, c_k^*$;

Step 4 (Iteration): If $i \in \{1, \cdots, k\}, c_i^* = c_i$,then end the algorithm,and the current $c_1^*, c_2^*, \cdots, c_k^*$ represents the final cluster,otherwise assign $c_i = c_i^*$,and repeat Steps 2 and 3 until there is no more updating.

C. The advantages and disadvantages of K-means algorithm[4,5]

Advantages:the flexibility of K-means algorithm is better for large dataset,whose time complexity is o(tkn),where t represents iteration times of algorithm,k is the number of clusters,n is the number of data poins in the dataset.

Disadvantages:cluster number needs to be given at first,however this number is generally received after clustering.And the K-means algorithm can not deel with the data in categorical attribute,it is sensitive to isolated points.It can not discover non-ball shape clusters or clusters that are quite different from each other.It usually traps into local optimization instead of global optimization.In addition,the results of algorithm are not steady,that is to say,with the same input parameter,the clustering results are completely different.

III. K-MEANS ALGORITHM FOR INTRUSION DETECTION SYSTEM

A. Data preprocessing

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} , $\cdots$ x_{nf} are n measuring values of variants, m_f is the mean value of the variant f , that is $m_f = \frac{\left(x_{1f} + x_{2f} + \cdots + x_{nf}\right)}{n}$;

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 :

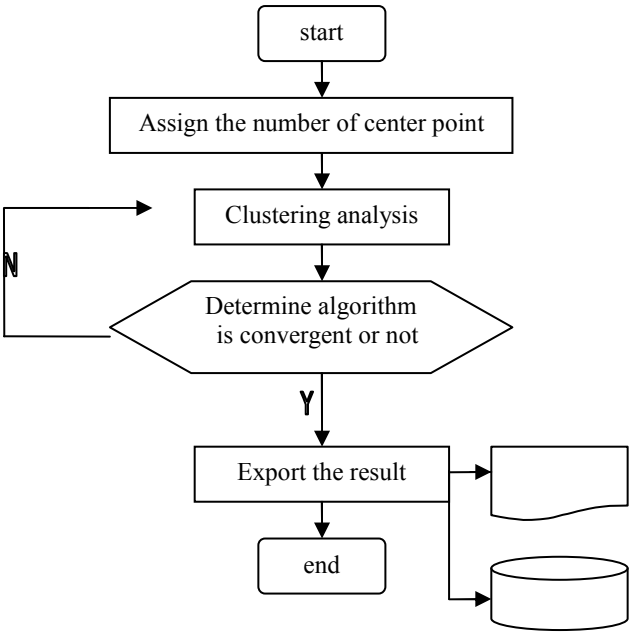

Figure 1. The procedure flow

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 center,the mean distance between each point and cluster center after clustering,the data number in each cluster and serial number in each cluster,etc;

b) Export the center 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.

IV. 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 labeled 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 center.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.

V. 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.

ACKNOWLEDGMENT

The author would like to thank Computer Research Section for the help in the expriment. And the author should also thank North China Electric Power University for the support for this work.

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