The application of Fuzzy clustering number algorithm in network intrusion detection

GuoLang
Chengdu vocational technical college,
Chengdu, China 610041

***Abstract*—In view of the defects of K-means algorithm in intrusion detection: the need of preassign cluster number and sensitive initial center and easy to fall into local optimum, this paper puts forward a fuzzy clustering algorithm. The fuzzy rules are utilized to express the invasion features, and standardized matrix is adopted to further process so as to reflect the approximation degree or correlation degree between the invasion indicator data and establish a similarity matrix. The simulation results of KDD CUP1999 data set show that the algorithm has better intrusion detection effect and can effectively detect the network intrusion data.**

Keywords- K-means algorithm; Fuzzy clustering number; Intrusion detection

I. INTRODUCTION

With the development and popularization of network technology, network security issues become increasingly prominent. How to quickly and effectively detect all kinds of intrusion behaviors is very important to guarantee the security of system and network resource. Intrusion detection, as an important part of computer network security, has received wide concern of domestic and international scholars [1].

Intrusion detection technology is usually divided into two kinds: abnormal detection and misuse detection. Misuse detection is based on the known intrusion attack information to detect the intrusion behaviors in system, which relies on the study of training focus marking data samples. When facing the unknown attack, new marking data sample is needed to retrain the detecting system, so the price is very high. Abnormal detection [2] is to use normal behavior information in monitoring system as the basis of intrusion and abnormal activities in detection system. Its dependence of the detection system is relatively small, so in recent years it has become one of the hot topics in network security study.

Based on clustering analysis, abnormal detection is an unsupervised learning method. It puts the similar data into one clustering and dissimilar data into different classes so as to find the relationship between the attributes and find out data distribution pattern. Clustering method generally includes: classification method, level method, method based on density, method based on grid, and methods based on model [3]. As a classic division method, K-means algorithm has been used in intrusion detection, but the algorithm has faults like sensitive initial center, easy to fall into local optimum and the need for the user to preassign clustering number in advance according to prior knowledge. Based on the traditional algorithms study, this paper proposes a intrusion detection judgment method based on fuzzy clustering. The experimental results of intrusion detection data set show that that the method can achieve better intrusion detection effect.

II. FUZZY CLUSTERING ALGORITHM

In order to reflect the advantages of secondary fuzzy clustering simulation algorithm, the traditional fuzzy clustering method should be understood. The traditional fuzzy clustering methods are classified into two kinds: one is fuzzy equivalence matrix dynamic clustering method, the other is fuzzy ISODATA clustering method. They are both based on all attributive classification. Here the first method is used to explain the classification solving process of traditional method and the specific process is as follows:

First of all, establish initial Fuzzy matrix: by definition 1, select the sample data of data set in practical problems and transform them into initial numerical matrix. Assume that sample data object set A contains n objects $a_1, a_2, \ldots a_n$, and each object a_i has m properties. The corresponding attribute value is $a_{i1}, a_{i2}, \ldots a_{in}$; form a row vector $a_i = (a_{i1}, a_{i2}, \ldots a_{in})$, then all vectors groups $a_i (i = 1, 2, \ldots, n)$ constitute the initial matrix $A_{n\times m}$ (Fuzzy matrix).

Second, Fuzzy matrix standardization: by definition 2, similarity processing of the initial matrix is proceeded. In actual application, the process commonly use two formulas: Formula (1) and (2), in which the extremum standardization transformation of Formula (1) is the most simple and convenient method. By calculation, n×m components r_{ij} is got, i.e., standardization matrix $R_{n\times m}$'s element.

Extremum standardization transformation:

$$r_{ij} = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})}, \quad i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, m \qquad (1)$$

Matrix elements' standard deviation transformation:

Published by Atlantis Press, Paris, France.

$$ri_j = \frac{x_{ij} - \overline{x_j}}{S_j} \qquad (2)$$

In it,

$$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2} ,$$

$i = 1,2,...,n \quad j = 1,2,...,m$

Then, similar coefficients' standardization: the standardized matrix is further processed to reflect the approximation degree or correlation degree between index data, and to establish similar matrix $Q_{n \times n}$. Commonly available formulas are three: (3), (4), and (5), which are quantity product method, Euclidean distance coefficient method and Angle cosine method.

Dot product:

$$T_{ij} = \begin{cases} 1, i=j \\ \frac{1}{M}\sum_{k=1}^{m} x_{ik}x_{jk}, i \neq j \end{cases}, i = 1,2,...n; j = 1,2...,m \qquad (3)$$

$$M > \max_{i \neq j} \left| \sum_{k=1}^{m} x_{ik} \times x_{jk} \right|$$

In it,

Euclidean distance coefficient (value approximation degree) :

$$T_{ij} = \sqrt{\sum_{k=1}^{m}(r_{ik} - r_{jk})^2}, i = 1,2,...,n; j = 1,2,...,m \qquad (4)$$

Angle cosine coefficient (shape approximation degree) :

$$T_{ij} = \frac{\sum_{k=1}^{m} r_{ik} r_{jk}}{\sqrt{\sum_{k=1}^{m} r_{ik}^2 \sum_{k=1}^{m} r_{jk}^2}}, i = 1,2,...,n; j = 1,2,...,m \qquad (5)$$

Finally, λ level classification and clustering: by theorem 1, a given λ value (λ∈[0, 1]), when λ is big, the classification is fine; when λ is smal, the classification is coarse. After λ is selected, compare λ value with each component T_{ij} in similar matrix $Q_{n \times n}$. When $T_{ij} \geqslant \lambda$, set A_{ij} value to 1; When $T_{ij} \geq \lambda$, T_{ij} value is set to 0. The new value of T_{ij} is noted as T_{ij} , then the λ cut matrix formed by T_{ij} constitutes only 0 and 1. Divide R_λ into n vectors $T_i = (T_{i1}, T_{i2}, \ldots, T_{in})$ by line, $i = 1,2,\ldots,n$ and put the corresponding T_i with same components into one class and the corresponding T_i with different components a separate class, until all the classifications are completed. Choose the class required from λ level classification, the corresponding data object classification is the clustering result of the problem.

III. Experiment and Analysis

A. The determination of parameter r

Detection rate and false detection rate are important evaluation indicators to measure the effectiveness of intrusion detection method, in which detection rate is the percentage of correct detected intrusion data number and all the invasion data number of the test set; False detection rate is the percentage of normal data failure detection as the invasion data number and all the normal data number of the test set. Usually the intrusion detection calculates detection rate with the guarantee of having low detection failure rate. By experiment, good r value is chosen to ensure that the maximum distinguish between normal and intrusion behavior is proceeded with lower failure detection rate. 15000 normal data are chosen as a training set and are trained with the proposed automatic-determined clustering number algorithm. The parameters are set: each time extract 3000 articles randomly; differential evolution's population size N=20; the largest iteration number of the algorithm is 100 times. After training, randomly extract 10000 normal data as a test set; respectively choose different r values to experiment; corresponding to different r value, the change curve of failure detection rate is shown in Figure 1:

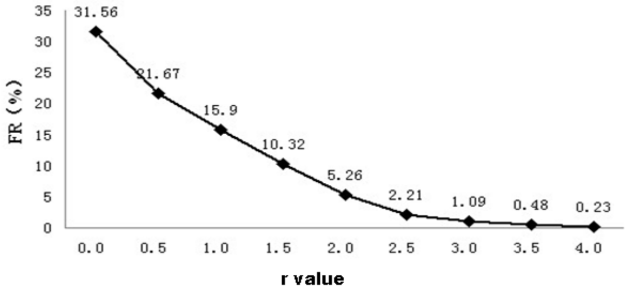

Figure 1. Failure detection rate (FR) change curve corresponding different r value

Figure 1 shows that when r >= 2.5, this paper's method has lower failure detection rate corresponding different training set, and can keep the false detection rate below 3%.

B. Detection rate and false detection rate calculation

For the four types of attacks: DOS, R2L, U2R, PROBE, the occurrence frequency of DOS and PROBE attack is relatively higher and the centralized data size is bigger, so each time1500 articles are randomly chosen to test. Randomly choose 10 data group for detection rate calculation; U2R and R2L have less data, including 52 articles and 1126 articles, the attack data of these two types are all selected to test detection rate; For all types of attacks, each time randomly select 4000 articles data in proportion, including 3600 articles' DOS attack, 125 articles' R2L

Published by Atlantis Press, Paris, France.

attack, 250 articles' PROBE attack, and 25 articles U2R attack. According to the above way, randomly select 10 groups data to calculate the detection rate; But to false detection rate calculation, every time 10000 articles are randomly selected in test data, 10 groups data are randomly selected to calculate. In the above training set and if =2.5, the detection rate and false detection rate of the four different attacks and normal data are shown in Table 1:

TABLE I. DETECTION RATE AND FALSE DETECTION RATE

Test	DOS	PROBE	U2R	R2L	Attacks	FR
1	99.60	85.53	23.08	7.10	95.28	2.17
2	98.87	85.87			95.00	2.05
3	98.73	84.40			94.55	1.97
4	98.73	84.73			95.10	1.99
5	99.40	84.53			94.98	2.01
6	99.33	84.33			95.10	2.00
7	99.13	85.73			95.23	2.04
8	99.47	85.33			95.10	2.25
9	99.27	84.73			95.03	2.39
10	99.27	84.20			95.18	2.25
Mean value	99.18	84.94	23.08	7.10	95.06	2.11

Compare the testing results and the detection results of literature [1], [2], [3], the result is as follows:

TABLE II. COMPARE RESULTS WITH LITERATURE

References	DOS	PROBE	U2R	R2L	All Attacks	FR
Literature[1]	-	-	-	-	35.7-88	1.44-8.14
Literature [2]	-	-	-	-	28-93	0.5-10
Literature [3]	97.1	83.3	13.2	8.4	91.8	0.5
This paper's method	99.18	84.94	23.08	7.1	95.06	2.11

Table 2 shows, literatuer[1]'s highest detection rate is 88% with 8.14%error detection rate; literature [2] 's highest detection rate is 93% with 10%error detection rate; compared with literature [2] and[1], the new algorithm has higher detection rate in the guarantee of low error rate of detection, which is higher than that of the above two methods; Compared with the literature [3], although this paper algorithm's false detection rate is a little higher, it is in the acceptable range, and for frequent data model, such as DOS and PROBE attacks, it has obtained good test results, which is higher than that of literature [3], and improves the detection rate of all kinds of attacks. All algorithms have low U2R and R2L detection rate. The main reason is that the two behaviors have small differences from normal behavior and easy to be put as normal data object, which causes the difficulty of detection algorithm.

IV. CONCLUSION

Aiming at traditional clustering algorithm's shortages in intrusion detection such as difficult determination of the cluster number and cluster center is sensitive and easy to fall into local optimum, this paper puts forward a fuzzy clustering number algorithm and proposes an intrusion detection judgment method based on the algorithm. The experiment results of intrusion detection data set show that this algorithm can automatically determine the clustering number without human intervention and achieve good results. It is an effective intrusion detection algorithm and has very good practical value.

REFERENCES

[1] Fu De-sheng,Zhou Shu,Guo Ping.Design and Implementation of Distributed Network Intrusion Detection System Based on Data Mining[J].Computer Science,2009,36(3):103-105.

[2] XU R,DONALD W.Survey of clustering algorithms[J].IEEE Transactions on Neural Networks,2005,16(3):645-678.

[3] Price K V,Storn R M,Lampinen J A. Differential evolution:A practical approach to global optimization[M].Berlin Heidelberg:Springer,2005.

[4] Mukkamala S,Janoski G,Sung A H.Intrusion Detection Using Support Vector Machines and Neural Networks[C]//Proc.of IEEE International Joint Conference on Neural Networks.Washington D.C.,USA:[s.n.],2002:1702-1707.

[5] Portnoy L,Eskin E,Stolfo S J.Intrusion Detection with Unlabeled Data Using Clustering[C]//Proc.of ACM CSS Workshop on Data Mining Applied to Security.Philadelphia,PA,2001:123-130.

[6] Eskin E,Arnold A,Prerau M,etal.A geometric framework for unsupervised anomaly detection:Detecting intrustions in unlabeled data[J].Data Mining for Security Applications,2002,17(5):873-891.

[7] ELKAN C.Results of the KDD'99 classifier learning[J].ACM SIGKDD Explorations Newsletter,2000,1(2):63-64 .