ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE) Vol. 5, Issue 12, December 2019

Application of Cluster Algorithm Based on Kernel in Intrusion Detection

Xuecheng Liu[1]
College of Mathematics and Statistics, Taishan University, 271000, Tai'An, ShanDong, China
Email:controllxc@126.com

***Abstract*—This article proposed a kind of cluster algorithm based on kernel, and applied it in the intrusion detection, constructed one kind of new intrusion detection model. By using Mercer kernel function, we can map the data in the original space to a high-dimensional feature space in which we can perform clustering efficiently. Furthermore, the article used the data partition method in the initialization cluster center' choice, this cluster method has a big improvement in performance compared to the classic cluster algorithm, and it has quicker convergence rate as well as more accurate cluster. The results of simulation experiments show the feasibility and effectiveness of the kernel clustering algorithm.**

***Index Terms*—Network security, Intrusion detection, Clustering analysis, Kernel function**

I. INTRODUCTION

With the continuous development of network technology and network scale, the risk and opportunity of network intrusion is more and more, the risk of network security has become a global problem. So at present, how to quickly and efficiently find all kinds of new intrusion behavior have become very important for ensuring the security of system and network. The technology of intrusion detection is mainly divided into two categories, that is anomaly detection and misuse detection. The essence of misuse detection centers around using an expert system to identify intrusions based on a predetermined knowledge base. Anomaly detection is concerned with identifying events that appear to be anomalous with respect to normal system behavior.

For anomaly detection, people construct intrusion detection system using the choosing statistical characteristic mainly relying on their intuition and experiences. The essence of misuse detection centers around using an expert system to identify intrusions based on a predetermined knowledge base. As a result, misuse systems are capable of attaining high levels of accuracy in identifying even very subtle intrusions that are represented in their expert knowledge base; similarly, if this expert knowledge base is crafted carefully, misuse systems produce a minimal number of false positives. And the establishment of model we used in this method depends entirely on learning of training data set sample, so guarantee that the data set is clean is of vital importance to establish practical intrusion detection. In fact, a clean data set collected for the system learning is often not so easy, and the price is also high. So studying a kind of unsupervised intrusion detection method is very necessary.

II. IMPROVED KERNEL CLUSTERING ALGORITHM

Clustering analysis[1] is an unsupervised learning method, the goal of Clustering analysis is to divide a number of feature model into several collections in accordance with the 'similarity' between them, each set feature model is 'similarity' in accordance with some measure, and the different set feature model is not 'similarity' in accordance with some measure. These are a lot of kinds of clustering algorithm such as, K centric algorithm, maximum likelihood estimation algorithm and the method based on graph theory. K centric algorithm is one of the most commonly used simple algorithm, the method firstly select some metric distance as a similarity measure between the modes, and then determine the criterion function which can evaluate the quality of the clustering results, after the initial clustering center is given, find the best clustering results which can make the criterion function gain the extremism value using iterative method. But K centric clustering algorithm [3] has a strong dependence for different values of K and the class distribution, this make the clustering effect poor. In order to overcome the shortcomings of centurion clustering and improve the clustering effect, we use the idea of kernel learning method, and proposed an improved K -means clustering method based on kernel. First the samples for classification in the original will be mapped to a high-dimensional feature space (kernel space) [5], this makes the samples linear separable (or approximate linear separable), then carries on the centric clustering in the kernel space. If the K nonlinear mapping is continuous and smooth, then the topology and order of the sample will be maintained in the kernel space. So, the point that gather into a cluster in original space will also get together in the kernel space by nonlinear mapping, and it will highlight the different characteristics of the different class samples, makes the linear undivided sample become linear separable (or approximate linearly separable), and also cluster [7] in the kernel space through the definition of similarity metrics (Euclidean distance in kernel space between sample) and objective function. Then, if use kernel method[2], the algorithm can reduce the error of clustering for the undivided linear sample as well as the asymmetric distribution sample, and get better clustering effect. So this paper puts forward a kernel centric algorithm, and applies this clustering algorithm into actual data, the results show that our method not only improves the clustering effect, and also accelerates the speed of clustering and lowers the number of clustering iteration.

First mercer kernel use a nonlinear mapping Φ : $R^n \mapsto F, x \mapsto \Phi(x)$ to maple the sample S in the original space into a high dimensional kernel space F, this goal is to highlight characteristic differences between samples of different categories, and makes samples in kernel space become linearly separable, and then it can proceed the general C centric clustering in the high dimensional kernel space. In kernel space, the sample to be classified is:

$(\Phi(x_1),\Phi(x_2),...,\Phi(x_l))$, then the dot product form in the input space can be expressed in Mercer kernel in the feature space:

$$K(x_i,x_j)=((\Phi(x_i).\Phi(x_j))$$

(1)

A kernel function matrix $K_{i,j}=K(x_i,x_j)$ is composed of all the samples. Support vector machines use the Mercer kernel to constructor the decision function in the characteristic space; it corresponds to the input space of nonlinear functions. The main idea of kernel clustering method is to map the sample in the input space to the characteristic space using Mercer kernel, and make the sample after mapped has better clustering form.

In fact, if a function should satisfy the Mercer condition, it can be used as the Mercer kernel, and it can be decomposed into characteristic space form of dot product at the same time. Mercer condition can be described as: for any square integral function $g(x)$, are all satisfied:

$$\iint_{L_2\otimes L_2}K(x,y)g(x)g(y)dxdy\geq 0$$

(2)

Then it can find kernel characteristic value ($\phi_i(x),\lambda_i$) of characteristic function of the kernel function K, and the kernel function can be expressed as:

$$K(x,y)=\sum_{i=1}^{N_H}\lambda_i\phi_i(x)\phi_i(y)$$

(3)

The N_H is the dimension of the characteristic space and the nonlinear mapping function can be expressed as:

$$\Phi(x)=(\sqrt{\lambda_1}\phi_1(x),\sqrt{\lambda_2}\phi_2(x),...,\sqrt{\lambda_{N_H}}\phi_{N_H}(x))^T$$

(4)

Then formula (1) can be calculated by formula (3) and formula (4).

The Mercer kernel function [4] has three forms; there are linear kernel function, polynomial kernel function and Gaussian radial basis kernel. And the Gaussian radial basis kernel is the most common used:

$$K(x,y)=\exp(^{-\|x-y\|^2}\!\!\diagup_{\sigma^2}),\sigma\neq 0\in R$$

(5)

σ is the customization parameters.

Because the corresponding characteristic space of the Gaussian kernel function is infinite dimension, limited samples in the characteristic space must be linearly separable, so we can get the clustering algorithm based on kernel.

Assumes that the sample in the input space has been mapped to characteristic space $\Phi(x_1),...,\Phi(x_l)$, the Euclidean distance in the characteristic space can be expressed:

$$d_H(x,y)=\sqrt{\|\Phi(x)-\Phi(y)\|^2}$$
$$=\sqrt{\Phi(x).\Phi(x)-2\Phi(x).\Phi(y)+\Phi(y)\Phi(y)} \qquad (6)$$

In general, the expression of nonlinear function is unknown, so formula (6) can be written:

$$d_H(x,y)=\sqrt{K(x,x)-2K(x,y)+K(y,y)}$$

(7)

So we will regard the formula (7) as a similarity measure function of clustering.

Kernel cluster can adopt hard portion method, and use fuzzy clustering method and so. Here we use hard part -ion method only, the criterion of cluster is making the following objective function minimum:

$$J=\sum_{i=1}^{C}\sum_{\substack{j=1\\x\in C_i}}^{N_i}(K(x_j,x_j)-\frac{2}{N_i}\sum_{k=1}^{N_i}K(x_j,x_k)+\frac{1}{N_i^2}\sum_{k,p=1}^{N_i}K(x_k,x_p)) \qquad (8)$$

Here C is the number of cluster categories, N_i is the number of the C_i sample, the modulus of the center categories is:

$$\|W_i\|^2=\frac{1}{N_i^2}\sum_{k,p=1}^{N_i}K(x_k,x_p)$$

(9)

According to the above clustering criteria (8) and the measure of distance similarity, we can build kernel clustering algorithm as follow:

Step 1：determine the number of the clustering type C，$2\leq C\leq l$, and permissible error $E_{\max}$，k =1;

Step 2 ： determine the initialize cluster center $w_i(k),k=1,...,C$

In order to overcome the blindness in selecting the initialize center, we should make full use of existing data information of sample points. A way of preprocessing can be used in selecting of initialize center. First under the significance of certain norm, determine the farthest apart of the distance between two data points, and then adopt the technology of data segment, divide the data sample point set {$x_1,x_2,...,x_n$} into C segment evenly, and the center of each segment is regarded as the initialization center. This processing method can avoid the choice too close in the initialization center. Particular way is as follows:

(1)Undergo preprocessing，make:

$$M=\max_{1\leq i,j\leq l,i\neq j}\|x_i-y_i\|_2^2,d=\frac{M}{C},S_1=S,W=\varnothing$$

(2) in i=1，2,…,$C-1$，if $i<C$, then

$$W_i=\{x_j\,\Big|\,\|x_j-\overline{z_i}\|_2^2\leq d,x_j\in S_i\}$$

Which $\|\overline{z_i}\|_2^2=\max\{\|\overline{x_j}\|_2^2,x_j\in S_i\}$

$W=W\cup W_i$,

Make $S_{i+1} = {S_i}/{W_i}$;

(3) $W_C = {S}/{W}$;

(4) Compute the center $w_i = \frac{1}{W_i} \sum_{x_i \in W_i}^{l} x_i, i = 1,2,..., C$;

Step 3: construct the mapping function

$K_{xx} = K(x, x)$,

$K_{xw(k)} = K(x, W)$

$K_{ww} = K(W, W)$;

Step 4： $i = 1,2,..., C$, $j = 1,2,..., l$

According to the formula (7), calculate the distance between the center of the cluster and each sample point, which is:

$$d_H(x_j, w_i(k)) = \sqrt{K_{xx_{j,j}} - 2K_{xw_{j,i}}(k) + K_{ww_{i,i}}(k)}$$

Make

$$d_{ij} = \begin{cases} 1, d_H(x_j, W_i(k)) = \min\limits_p \{d_H(x_j, W_p(k))\} \\ 0, else \end{cases}$$

Step 5: modify the kernel matrix $K_{xw}(k), K_{ww}(k)$:

$$K_{xw_{\bullet,i}}(k+1) = \frac{\sum_{j=1}^{l} d_{ji} K(w, x_j)}{\sum_{j=1}^{l} d_{ji}},$$

$$K_{ww_{i,i}}(k+1) = \frac{\sum_{j=1}^{l}\sum_{p=1}^{l} d_{ji} d_{pi} K(x_j, x_p)}{(\sum_{j=1}^{l} d_{ji})^2}$$

Step 6: calculate the error.

$$e = \sum_{i=1}^{C} (K_{ww_{i,j}}(k+1) - 2\frac{\sum_{j=1}^{l} d_{ji} K_{ww_{j,i}}(k)}{} + K_{ww_{i,i}}(k)) .$$

Step 7：if $e < E_{\max}$ then end; else goto step 4.

Finally the result can be got, if $d_{ji} = 1$, then $x_j \in w_i$.

III. DETECTION ALGORITHMS

The clustering result need to mark their categories, by the algorithm of the first and the second assumption we can speculate, in the generated clustering result, if a cluster is gathered from normal data, then the number that it contains data should far greater than the number that the cluster is gathered from invasion data. So the cluster can be sorted simply according to the number that the cluster contains data, and set a number ***N***, those clusters which contain data more than *N* is considered as normal data, while the rest of the cluster is considered abnormal. The algorithm of identify cluster is described as follows:

Assume that C_i , $i = 1,..., num\_cluster$ is the generated clustering, *N* is the constant number between 0 and 1.

①Sort (C_i), sort the C_i according the number that the cluster contains data

② $j = 1$; $K = N \times num\_cluster$;

③ $repeat$ ；

④if $j < k$, then mark C_i as normal cluster；

⑤else label C_i as abnormal cluster;

⑥ $j++$;

⑦until $j > num\_cluster$.

This method is very simple, and easy to understand and implement, but its effectivement has a close relationship with the number of subcluster of normal behavior. If the number of divided normal behaviors is too small, each subcluster has its unique cluster center in characteristic space, it will reduce the relative amount of data in the single subcluster, these normal number will even less than the number of abnormal cluster. In this case, the normal data will mistakenly are regarded as abnormal, or the abnormal data will be regarded as normal data. In order to prevent the emergence of the problem, the capacity of all kinds of normal data should be increased as far as possible when the data set is generated, so in any of subcluster, each cluster can contain enough data in order to be distinguishing from the abnormal data.

Detection algorithm is described as below:

Suppose *x* is a network message detected:

① *x* will be standardization using the statistics data in preprocessing algorithm, that is $x \rightarrow x'$

② $j = 1$ ；

③repeat；

④calculate the distance between the center O_j of C_j and x', that is $dist(O_j, x')$;

⑤ $j++$ ；

⑥until $j > num\_cluster$

⑦ find the minimum $dist(O_{\min}, x)$ ， and gain the cluster $label$ that belongs $O_{\min}$;

⑧if $label$ is normal, and then *x* is normal data message; else it is abnormal data message.

This detect algorithm is very simple and rapid, so its efficiency is very high.

IV. SIMULATION EXPERIMENT

A. Sample description

The selected sample data is the authoritative test data in the current intrusion detection field, KDD CUP1999[6], derived from the 1998DARPA intrusion detection evaluation process. The data set provides a total of 4,900,00 records, to the provided every TCP/IP connection, in addition to some basic attributes(such as protocol types, the number of bytes sent etc.), also make use of domain knowledge to extend some of the properties(such as the number of login failure, the number of files generated operating, etc.), Some attribute information is calculated basis on the past 2 seconds, for example in the last 2 seconds the connection number that connect to the same server. Each connection has 41 qualitative or quantitative characteristics, 8 attributers in all characteristics are discrete variables, the rest are a continuous variables.

The intrusion data has 4 major categories, 24 small clusters. The major categories is: DOS （Denial of Service ）, for example ping of death; U2R(User to Root), for example eject; R2U(Remote to User), Remote to User guest; PROBING, Remote to User port scanning.

In order to meet the requirements of detection algorithm of two assumptions, the test data set need to some filtering. So this paper extracted from the test data set of 60638 records as the training sample set. 60032 records is normal data, the rest 606 records is invasion message, the proportion of all normal data reach 99%, accord with the first hypothesis requirement of detection algorithm. The intrusion type and the number of records is showed in the table One:

TABLE 1. INTRUSION TYPE AND THE NUMBER OF RECORDS IN TEST DATA SET

Major type	Small type and the records number
DOS	Total 284，nepture(141), smurf(143)
U2R	Total 68, buffer_overflow(22), loadmodule(2), perl(2), ps(13), xterm(13)
R2U	Total 131，ftp_write(3), guest_passwd(31), imap(1), multihop(18), named(17), sendmail(17), phf(2), warezmaster(29), xlock(9), xsnoop(4)
PROBING	Total 123，ispweep(30), nmap(19), portsweep(32), satan(42)

In the selection of test sample set, this paper have selected 4 sets of data, each set has 300,000 records. The first group and the second group is test data set, and the other two groups are selected from KDD CUP99 data which act as the training set(some selected data does not include in the training data, which are regarded as unknown intrusion data).

B. *Preprocessing*

Because the original test data include discrete and continuous variables in the attribute characteristics, some of them need to be processed.

For attribute characteristics of discrete variable, for example:

Assumption in a data set, the service attribute repeatedly emerge http, ftp, telnet, smtp 4 attributes, then the 4 attribute can be coded into 0001，0010，0100，1000, then the attribute of service is divided into four attributes service1，service2，service3，service4, where appear HTTP recording it will make service1=0，service2=0，service3=0，service4=1, where appear FTP recording it will make service1=0，service2=0，service3=1，service4=0, and so on, discrete attributes of service can be converted to continuity attribute variable.

According to the above method, other discrete attributes variable can also be handled, the benefits of this kind of process is to ensure that each record has the same distance.

For continuous attribute characteristics of the variables, different attribute have different metrics, so if don't preprocessing of the raw data, it will emerge that the large number eat the small number, for example, given the characteristics vector x_i = {1000，1，2}, x_j ={2000，2，1}, then

$$d(x_i, x_j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + |x_{i3} - x_{j3}|^2}$$

$$= \sqrt{|1000 - 2000|^2 + |1-2|^2 + |2-1|^2}$$

Apparently the whole characteristics are covered by the first characteristics data.

In order to solve this problem, the characteristics values of the data must be standardized, and can be make the following change:

(1)Calculate the mean absolute deviation S_f

$$S_f = \frac{1}{N}\sum_{i=1}^{n}(x_{if} - m_f) \qquad (10)$$

Among them, $x_{1f}, \ldots, x_{nf}$ is n attribute characteristic value of f , m_f is mean value of f , that is:

$$m_f = \frac{1}{N}\sum_{i=1}^{n} x_{if} \qquad (11)$$

(2)Calculate standardized characteristic attribute value

$$Z_{if} = \frac{x_{if} - m_f}{S_f} \qquad (12)$$

This average absolute deviation S_f has a better robustness than the standard deviation σ_f

According to the above three formula, by calculate their characteristics attributes, each message can get new data. Thus by using the statistical algorithm, the raw data can be mapped to a standard characteristics space, and can reduce the problem described above. The specific algorithm is as follows:

Assume that x_i is a raw network message in the training data set, n is the number of the whole training data set:

① $i = 1$;
②repeat;
③select $x_i \in Z$, According to the formula (10), (11), calculate S_f and m_f ;
④ $i++$;
⑤until $i = n$;
⑥ $i = 1$
⑦repeat;
⑧ According to the formula (12) calculate Z_{if} , the characteristics attribute value will be converted into standard form value, that is $x \to x'$;
⑨ $i++$;
⑩until $i = n$.

C. *Test result analysis*

About intrusion detection experiment result evaluation, it can refer to two main indexes: DR is detection rate and FAR is false alarm rate. They are defined as follows:

DR=the number of detected intrusion records/the number of total test record;

FAR=the number of the normal records which are mistaken for intrusion/the number of normal records in the total test records

In the test data set, the 4 group data set will be selected, each data set includes 10,000 records, and each set data is meeting the premise: that is the number of normal is much larger than the number of intrusion behavior:

TABLE 2. DATA SET DISTRIBUTION

Data types	number	Data types	number
Normal	8465	ftp_write	2
Buffer_ overflow	12	teardrop	94
land	2	smurf	289
warezmaster	73	nentune	272
snmpgetattact	191	rootkit	8
httptunnel	33	ipsweep	160
Guess_passwd	356	mscan	41

In the experiments, the parameter of the clustering algorithm is: $\sigma^2 = 0.5$, $E_{\max}$ =0.001, the following is the result of test model:

TABLE 3.DETECTION RESULTS

Date Set	DR (%)	FAR (%)
1-st group	82.84	1.51
2-ed group	78.05	0.72
3-rd group	83.17	0.69
4-th group	78.16	0.51

From the above results, the detection rate of our algorithm is very high, and the false alarm rate is very low. At the same time, the detection rate of abnormal samples is lower than the normal sample in the experiment, so the recognition ability of the system can be further improved by increasing the number of abnormal data.

V. CONCLUSION

This paper puts forward the application in intrusion detection based on algorithm of kernel clustering, and the simulation experiment results confirmed the feasibility and effectiveness of the method. This method map the input space into the high-dimensional characteristics space by using the Mercer kernel, and cluster in the characteristics space. Because of the mapping of the kernel function, the original characteristics which unable to show will be displayed significantly, thus out algorithm can be able to clustering smoothly. And on the choice of initial clustering center the method of data piecewise is used, this clustering method has great improvement than the classical clustering algorithm in performance, and has faster convergence speed and more accurate.

REFERENCES

[1] Wenxing Hong, Siting Zheng, Huan Wang, “A Job Recommender System based on user Clustering,” Journal of computers, Vol,8, No.8,August,pp. 1965,2013.

[2] Sch lkopf B, Mika S, Burges C et al. “Input space versus feature space in kernel-based methods,” IEEE Trans Neural networks, Vol.10(5), pp.1000-1017,1999.

[3] ZHANG Rong, RUDNICKY A I. “A large scale clustering scheme for kernel k-means,” Paattern Recognition, Vol 4,pp. 289-292,2002.

[4] GIROLAMI M. “Mercer kernel based clustering in feature space,” IEEE Trans on Neural Networds, Vol.13(2), pp.780-784,2002.

[5] Kong Rui, Zhang Guoxuan, Guo Li. “Kernel-based K-means Clustering,” Computer Engineering, Vol.30(11), pp.12-14,2004.

[6] Jianhua Zhao, Weihua Li, “Intrusion detection based on improved SOM with optimized GA,” Vol.8, No 6,pp.1456-1463,Jun 2013.

[7] C.F.Chien and L.F. Chen, “Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry,” Expert System with Application, vol.34(1), pp.280-290,2008.

Xuecheng Liu is a lecturer. He obtained his first degree of Bachelor of Management at the University of Jinan and Master of Science Degree at GuiZhou University. His major fields of study are Network Information Security.