Research of WaveCluster Algorithm in Intrusion Detection System

Mingwei Zhao
Dalian University of Technology
School of Electronic and Information Engineering
Innovation Mansion, 116024
Dalian, China
zhaomw@dlut.edu.cn

Yang Liu
Dalian University of Technology
School of Electronic and Information Engineering
Innovation Mansion, 116024
Dalian, China
victor0389@hotmail.com

Rong'an Jiang
Dalian University of Technology
School of Electronic and Information Engineering
Innovation Mansion, 116024
Dalian, China
rajiang@dlut.edu.cn

Abstract

In this paper, we introduce a clustering algorithm for Intrusion Detection based on WaveCluster algorithm and an entropy-based characteristics screening algorithm. WaveCluster algorithm has a low time complexity when the data are low-dimensional, but on the contrary, the actual network data are high-dimensional. So we reduce the dimension of the network data using characteristics screening before they are clustered. And the algorithm inherits the WaveCluster's advantage of multi-resolution, adaptive, and not requiring specific pre-determined parameters. We can rapidly and accurately identify arbitrarily shaped clusters at different scales and degree to find intrusion effectively. Experimental results on KDD Cup 1999 data sets show that the detection rate of the algorithm is higher than the algorithm in the reference. The time complexity of the algorithm is low.

1. Introduction

As individuals, corporations, and governments rely on the Internet for communication and work together more and more, the demand of solution of network security is also increasing rapidly. Users need to stop intruders at the same time guarantee the security of themselves and their partners. Although the widely used firewalls and various identity authentication systems could protect computer systems against unauthorized accesses, they are powerless to professional hackers and malicious acts of authorized users. Therefore, Intrusion Detection System is a useful complement to the firewall, it can detect, alarm, and expel intrusions before they harm computer systems. Now, Intrusion Detection System is considered to be the second security gate of computer systems behind the firewall. It provides protections of internal attacks, external attacks, and misuse through network monitoring, and greatly enhances the safety of the network.

An Intrusion Detection System can be mainly divided into data acquisition module, intrusions analyzing engine modules, emergency processing module, configuration management module, and related auxiliary module[1]. And the intrusions analyzing engine module plays a decisive role to the detection results. With network data transmission growing continuously, it costs too much to assort and label manually, so, various unsupervised detection methods, which are represented by clustering, have been used to the analysis of intrusions gradually. In the intrusions analyzing engine modules, K-means algorithm and K-medoid algorithm are used much, but these algorithms often require users to predetermine the parameter K, only suit for the clusters of convex shapes, and are more sensitive to the input order of data[2]. In this paper, WaveCluster algorithm is used, which has the characteristics of multi-resolution, thus it can find clusters in any shapes. And it is not sensitive to the input order of data. Compared to K-means algorithm, K-medoid algorithm and so on, it has a better clustering effect, can make a higher detection rate and a lower false positive, and would improve the performance of Intrusion Detection System.

WaveCluster can be defined as follows: the clustering analysis algorithm based on wavelet analysis. In WaveCluster algorithm, data is considered as multi-dimensional signals, and is transferred to the frequency domain using wavelet transform[3].

In WaveCluster algorithm, a spatial objects set O_i, $1 \le i \le N$, is given, and the purpose of the algorithm is to

978-0-7695-3508-1/08 $25.00 © 2008 IEEE
DOI 10.1109/CIS.2008.119

detect clusters and assign labels to the objects based on the cluster that they belong to. WaveCluster algorithm quantizes the feature space, and applies discrete wavelet transform on the quantized feature space. We use the algorithm in [4] to find connected components in the two-dimensional feature space, and it is the same to high-dimensional feature spaces. Finally, the algorithm labels cells and makes lookup table. The main idea of WaveCluster algorithm is to transform the original feature space by applying wavelet transform, and then find the dense regions in the new feature space. It generates clusters at different scales and degree, which can be chosen on the needs of users.

WaveCluster algorithm performs well on low-dimensional data. It is order insensitive to input objects. It does not need users to decide the parameter. And it automatically generates multilevel clustering results for users to choose, because of the characteristic of multi-resolution of the algorithm. These good characteristics of WaveCluster algorithm ensure its good performance in Intrusion Detection System.

It can be found that the complexity of WaveCluster algorithm increases in an exponential growth with the dimension of data[3]. And through analyzing the KDD Cup 1999 data sets which are often used in the field of Intrusion Detection as experimental samples, we can also find that the network data include 41 characteristics. Obviously, the source data captured by data acquisition module are high-dimensional data, and are not conducive to WaveCluster algorithm to process.

2. Improved WaveCluster algorithm

In this paper, we put forward a new characteristics-based WaveCluster algorithm, which integrates the original WaveCluster algorithm and an entropy-based characteristics screening algorithm of the source data.

First, the algorithm assesses the importance of the original characteristics set by constructing an entropy measurement based on the similarity of the objects in order to find the important characteristics subsets.

2.1. Definition of the similarity of the objects

Given n objects in d-dimensional feature space $x_i(x_{i1},x_{i2},\ldots,x_{id})^T$, i=1,…,n, the most commonly used measurement of the similarity is $L_p(p\in N)$ norm distance,

$$\forall x_i, x_j \in R^d, L_p(x_i, x_j) = \left\| x_i - x_j \right\|_p = \left(\sum_{k=1}^{d} \left| x_{ik} - x_{jk} \right|^p \right)^{1/p} \quad (1)$$

Taking into the peculiar behavior of the most L_p distance in high-dimensional space (it points in [5] that the furthest relative distance of the objects will weaken to 0 with the increasing of the dimension of the space when $p\geq 2$), L_1 distance is used to define the similarity of the objects in this paper. Given the set of the characteristics of d-dimensional space $F=\{X_1,X_2,\ldots,X_d\}$, a variant of L_1 distance can be defined as follows:

$$d(x_i, x_j) = \frac{1}{d} \sum_{k=1}^{d} \left| x_{ik} - x_{jk} \right| \quad (2)$$

In (2), d is used in the dimension standardization of the distance, so that the distance of the objects in different subspaces can be comparable. According to the distance measurement given by (2), the similarity can be defined as follows:

$$s(x_i, x_j) = e^{-\frac{d(x_i,x_j)}{\sigma}} \quad (3)$$

Clearly, $s(x_i,x_j)$ is the monotone decreasing function of the objects distance, that is, when $d(x_i,x_j)=0$, $s(x_i,x_j)$ gets the maximum value 1, and when $d(x_i,x_j)\to\infty$, $s(x_i,x_j)$ tends to the minimum value 0. Parameter σ is used to control the attenuation of the similarity, smaller σ is, the similarity will weaken faster with the distance increasing.

2.2. The entropy measurement based on the similarity of the objects

Given n objects in d-dimensional space and their similarity matrix $S=(s_{ij})_{n\times n}=(s(x_i,x_j))_{n\times n}$, for any two objects of these, e.g. $x_i,x_j\in R^d$, the entropy measurement is defined as follows:

$$H_{ij} = -\frac{1}{\ln 2}\left[\begin{array}{l} S_{ij} \times \ln S_{ij} \\ + (1 - S_{ij}) \times \ln(1 - S_{ij}) \end{array} \right] \quad (4)$$

We can find following natures of the entropy measurement based on the similarity of the objects.

(1)$H_{ij}\geq 0$, if and only if $s(x_i,x_j)=1$ or 0, it is equivalency, that is, when the similarity of the objects values maximum or minimum, the entropy is the smallest.

(2)$H_{ij}\leq 1$, if and only if $s(x_i,x_j)=0.5$, it is equivalency, that is, when the distance of the objects $d(x_i,x_j)=\sigma\ln 2$, the entropy is the largest. Here $\sigma = \overline{D} / \ln 2$, $\overline{D}$ is the average distance of the objects, when the distance of the objects approach to the average, the entropy is the largest. If all the objects in the feature space are considered, the total entropy is as follows:

$$E_H = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} H_{ij} \quad (5)$$

2.3. Entropy-based characteristics screening

Corresponding to every characteristic X_k,k=1,…,d in the original characteristics set F, the algorithm calculates the increase of the total entropy after excluding this characteristic as the importance measurement of characteristic X_k,

$$I(X_k) = E_H(F - X_k) - E_H(F) \quad (6)$$

In (6), $E_H \in [0,1]$, $I(X_k) \in [-1,1]$. When the similarity of the objects is defined, dimension standardization of the distance has been done. So the total entropy in different feature subspaces are comparable. As the important characteristics that help show the structure of clusters are removed, the total entropy increases, and $I(X_k)>0$. In extreme cases, $I(X_k)=1$, that is, the distribution of data is only decided by characteristic X_k, and X_k has a good separability, once characteristic X_k is removed, the similarity of the objects are almost completely equal value; on the contrary, if some unimportant characteristics or noises are removed, the total entropy decreases, and $I(X_k)\leq 0$. In extreme cases, $I(X_k)=-1$, that is, characteristic X_k totally confused the original data distribution, once characteristic X_k is removed, the uncertainty of data distribution will decrease to 0. Consequently, by assessing the importance of each characteristic, those unimportant characteristics and noises are removed, and the important characteristics are chosen.

2.4. Description and analysis of the algorithm

Given a set of spatial objects o_i,$1\leq i\leq n$, the algorithm calculates entropy of each characteristic and finds main characteristics, reduces the dimension. And then, clusters are detected in the new low-dimensional feature space. The objects are assigned labels based on the cluster that they belong to. Wavelet transform is used as a clustering promoting tool.

Algorithm 1.

Input: Characteristic set and feature vectors of multi-dimensional data objects

Output: clustered objects

1. Calculate the total entropy of the original feature space and the entropy after removing each characteristic.
2. Get the subset of the important characteristics through removing those unimportant characteristics by the results of step 1.
3. Quantize low-dimensional feature space, then assign objects to the cells.
4. Apply wavelet transform on the quantized feature space.
5. Find the connected components (clusters) in the sub-bands of the transformed feature space, at different levels.
6. Assign labels to the cells.
7. Make the lookup table.
8. Map the objects to the clusters.

We analyze this algorithm, the time complexity of calculating the total entropy is $O(n^2)$, and the time complexity of calculating entropy after removing each characteristic is $O(d\times n^2)$. Obviously, when n is large, it is inefficient to calculate the importance of characteristics. So, we use a random sampling method to improve this problem. We take $m(m<<n)$ samples randomly from the original data set D as input of the algorithm, and the time complexity here becomes $O(d\times m^2)$. It can be found that the running time of the algorithm is only decided by sample size, and has nothing to do with the original set size. But if the original data distribution is destroyed in the process, that is unacceptable. The solution is mentioned in [6], that is, clustering results can stand the clustering property of original data accurately as long as the sampling rate is not less than a certain threshold, and when n is large, the sampling rate of 2.5% is suitable.

Assuming the feature vectors of objects are d-dimensional, we get a d-dimensional feature space. Here the number of data objects n is large but d is low. WaveCluster algorithm scans all the database objects first, and assigns them to the corresponding cells, the time complexity is $O(n)$. Assuming there are m cells in each dimension of feature space, there would be $K=m^d$ cells. Time complexity of applying wavelet transform on the quantized feature space (step 4) is $O(ldK)=O(dK)=O(K)$, where l is a small constant representing the length of the filter used in wavelet transform. If we apply wavelet transform for T levels of decomposition, downsample for each level, $d\geq 2$, the required time is

$$O\left(K + \frac{K}{2^d} + \frac{K}{(2^d)^2} + \cdots + \frac{K}{(2^d)^T}\right) = O\left(K\sum_{i=0}^{T}\frac{1}{(2^d)^i}\right)$$

$$= O\left(K\sum_{i=0}^{T}(2^{-d})^i\right) = O\left(K\frac{1-(2^{-d})^{T+1}}{1-2^{-d}}\right) \leq O\left(\frac{4}{3}K\right)$$

That means the cost to apply wavelet transform for multiple levels would be at most $O\left(\frac{4}{3}K\right)$. The time complexity of finding the connected components in the transformed feature space and assigning labels is $O(cK)=O(K)$, where c is a small constant. Making the lookup table requires $O(K)$ time. The last step of the algorithm's time complexity is $O(n)$. Because of the way of the algorithm finding the connected components, the number of clusters does not affect the time complexity of WaveCluster algorithm.

3. Experiments

The experimental procedures are achieved in Matlab 6.5. The experimental platform is the computer with Intel Celeron 2.53GHz CPU, 512MB Memory, and Windows XP Professional Service Pack 2 Operation System.

We take 10000 records as original samples from KDD Cup 1999 data sets, in which 9700 records are normal data, and 300 records are intrusion data. All four main types of attacks are included in the 300 records of intrusion data. The intrusion data is 3% of the total records, and is far less than the normal connection numbers.

We select Cohen-Daubechies-Feauveau(2,2) wavelet function and Cohen- Daubechies-Feauveau(4,2) wavelet function to compare the algorithm in this paper and the algorithm in [7] using the experimental data. The results are shown in table 1 and table 2.

Table 1. Detection results of the algorithm in [7]

Wavelet Functions	Detection Rate	False Positive
Cohen-Daubechies-Feauveau(2,2)	98.15%	1.65%
Cohen-Daubechies-Feauveau(4,2)	97.98%	1.76%
Average	98.07%	1.71%

Table 2. Detection results of the algorithm in this paper

Wavelet Functions	Detection Rate	False Positive
Cohen-Daubechies-Feauveau(2,2)	99.21%	1.03%
Cohen-Daubechies-Feauveau(4,2)	98.80%	1.54%
Average	99.01%	1.29%

We can see from the results of the experiments that the algorithm in this paper has a detection rate of 99.01%, which is higher than the algorithm in [7]. The false positive of the algorithm in this paper is 1.29%, which is lower than the algorithm in [7]. These show that the algorithm in this paper could detect the intrusion data better, and keep a lower false positive.

Both of the algorithms are all need a pretreatment process. The pretreatment of the algorithm in this paper is to reduce the dimension of the source data, but due to the sampling method, it spends less time actually. The pretreatment of the algorithm in [7] is to train the neural network, it needs more time relatively. So the algorithm in this paper has a better time consuming result than the algorithm in [7] if the pretreatment process is included. But the neural network training of the algorithm in [7] do not need to do every time, the dimension reduction of the algorithm in this paper is necessary in each time. Generally speaking, in a short period, the algorithm in this paper is better in time consuming result, but it is opposite in the long term.

4. Conclusion and prospect

In this paper, we put forward a new characteristics-based WaveCluster algorithm, which integrates the original WaveCluster algorithm and an entropy-based characteristics screening algorithm of the source data. The algorithm overcomes the original WaveCluster algorithm's problems on clustering of the high-dimensional data in the network transmission. And it takes advantage of the original WaveCluster's good characteristics, increases the detection rate of Intrusion Detection System. With the rapid development of network applications, network data transmission also increases rapidly. In order to detect the intrusion data more timely and effectively, the next step of the work is mainly on the parallel improvement of the algorithm, so that the algorithm can run faster and also reduce the cost of client resources.

5. References

[1] Khaled Labib, "Computer security and intrusion detection," *Crossroads*, vol 11, no. 1, pp. 2-2, Aug. 2004.

[2] Roy Gelbard, Orit Goldman, Israel Spiegler, "Investigating diversity of clustering methods: An empirical comparison," *Data & Knowledge Engineering*, vol 63, no. 1, pp. 155-166, Oct 2007.

[3] Sheikholeslami G, Chatterjee S, Zhang A, "WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases," *Proceedings of the 24th VLDB conference*, New York, 1998, pp. 428-439.

[4] Horn B. K. P., Robot Vision, Massachusetts: Massachusetts Institute of Technology Press and the McGraw-Hill Book Company, 1986.

[5] Hinneburg A., Charu C., Aggarwal C. C., et al. "What is the nearest neighbor in high dimensional spaces?," *Proceedings of 26th International Conference on VLDB*, Cairo, Egypt, Morgan Kaufmann Publishers Inc, 2000.

[6] Guha S., Rastogi R., Shim K., "CURE: an efficient clustering algorithm for large databases," *Proceedings of the 17th ACM SIGMOD International Conference on Management of Data*, Seattle, Washington, ACM Press, 1998.

[7] Dangfeng Zhu, “Research of Intrusion Detection Technique Based on Wavelet Neural Networks: (Master’s degree thesis),” Lanzhou, Lanzhou University, 2006.