An improved WaveCluster algorithm based on ICA[1]

Xiaoli Li
Modern Educational Technology Center
Nantong University
Nantong, China
nuonuoli1218@163.com

Min Luo
Computer School of Wuhan University
Wuhan, China
jsjgfzx@whu.edu.cn

***Abstract*—In this paper, we introduce an improved WaveCluster algorithm, which can deal with the high-dimensional data issue (IWCA algorithm for short). The algorithm combines the advantages of ICA algorithm that can reduce the dimension of database. By this means, we can not only cluster the high-dimensional data, but also reduce the time complexities. Finally, the experimental results on KDD Cup 1999 data sets show that IWCA algorithm can efficiently find clusters in high-dimensional data space.**

Keywords- ICA, wavelet, cluster, high-dimensional data

I. INTRODUCTION

Cluster analysis is an exploratory data analysis tool for solving classification problems. Its object is to sort cases (people, things, events, etc) into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Each cluster thus describes, in terms of the data collected, the class to which its members belong; and this description may be abstracted through use from the particular to the general class or type. Cluster analysis has many kinds of methods, such as partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods, and so on.

By comparing with several commonly cluster analysis methods, we find that these algorithms are not suitable for the high-dimensional data mostly. The time complexities of those minority algorithms which can be used in high-dimensional data can obviously advance along with the dimension ascension. So reducing the dimension of data space becomes an effective means in processing the high-dimensional problems.

There are some commonly technologies in reducing the dimension of data at present. Dimension-chooses and dimension-transforms are typical methods of them. They obtain a lower-dimensional subspace by choosing the dimension with good cluster attribute, or by analyzing the dimension of original data space. There are two popular algorithms in this domain. One is PCA (Principal Component Analysis), and the other is ICA (Independent Component Analysis). Both of them reduce simulation time by statistics and samples. The different between them is PCA supposes that the original data forms a Gaussian distribution, but ICA can obtain sample quantities on behalf of overall precisely, does not need such supposition. The sample quantities of ICA are fewer than that of PCA, and the accuracy is higher than that of PCA[1].

In this paper, based on to the analysis of ICA, we introduce a improved WaveCluster algorithm – IWCA. The experiments show that IWCA algorithm can not only cluster the high-dimensional data, but also reduce the time complexities compared with others cluster algorithms.

II. ICA ALGORITHM

ICA[2] is a linear transformation method which is based on higher-order statistical property of the signal in recent years. Through transformation, the output component is statistically independent as far as possible. It is conveniently to the recognition because that the number of the features reduces greatly by extracting the features.

Suppose that the objective function is J ,and the study rule is C , the feature extraction process of ICA is the process of causing J to obtain biggest or minimum value according to the study rule C .

A. Algorithm pretreatment

Suppose that the original data has the d -dimensional feature[3,4]. The expression is $X=\left(x_1,x_2,\cdots,x_d\right)^T$. Firstly, make the data form the symmetrical distribution about the zero, that is, to each x_i , make the average value be equal to 0.

$$E\{x_i\}=0 \tag{1}$$

So，

$$x_i \coloneqq x_i - E\{x_i\} \tag{2}$$

Then, carry on the processing of whitening to eliminate the relevance of second-order of the data. After whitening, the data is independent in statistics of second-order, and has the unit variance. Suppose that the transformation matrix for whitening is V , after whitening,

$$z=Vx,\quad E\{zz^T\}=I \tag{3}$$

Here, we use the PCA algorithm to whitening.

$$V=D^{-1/2}E^T \tag{4}$$

E is the matrix which is composed by the feature vector corresponding to the covariance matrix of x . D is the

1. Supported by the National Natural Science Foundation of China under Grant Nos. 90718006,90718005, 60743003

978-1-4244-3693-4/09/$25.00 ©2009 IEEE

corresponding feature value matrix. So, $E^T DE = E\{xx^T\}$. The goal is to unit by using E to divide $D^{1/2}$. Obviously, the feature vector is orthogonal.

B. Feature Selestions

In order to make the output component independent as far as possible through linear substitution, we suppose that each x_i is the linear combination of m independent sub-feature $S = (s_1, s_2, \cdots, s_m)^T$ after the process of pretreatment, and each s_i has zero mean

$$X = AS \quad (5)$$

Namely,

$$x_i = a_1 s_1 + a_2 s_2 + \cdots + a_m s_m \quad (6)$$

In which, $A = [a_1, a_2, \cdots, a_m] \in \Re^{d \times m} (a_i \in \Re^d)$ is the transformation matrix of the Order, $r(A) = m$.

So, the likelihood estimator Y of S is the m-dimension feature value which we want:

$$Y = \hat{S} = WA \quad (7)$$

Here, make W be the inverse matrix of A, ($A = W^{-1}$), and $W = BD^{-1/2}E^T$, so

$$y = Wx = BD^{-1/2}E^T x \quad (8)$$

By formula (4), we sort the feature value from large to small in matrix D, and compose a new feature matrix D_m using first m maximum feature value., corresponding to the feature matrix E_m^T composed by these m feature vectors, so

$$V_m = D_m^{-1/2} E_m^T \quad (9)$$

Here, the dimension of the feature space is reduced from d-dimension to m-dimension.

III. WAVECLUSTER ALGORITHM

In WCA[5] (WaveCluster Algorithm), the author thought the data is the multi-dimensional signal. It can transformate the data to the frequency domain by the signal processing technology of wavelet transformation. Through wavelet transformation, we can obtain a transformation space according to the suitable core function. It is easy to distinguish data cluster by finding the crowded region in transformation space. WCA is suitable in any shape cluster, and it can process noise. It is not only insensitive to the data-input order, but also includes the multi-resolution attribute as the result of the application in signal processing domain.

A. WCA Algorithm

Given a set of spatial objects o_i, $1 \le i \le N$, the goal of the algorithm is to detect clusters and assign labels to the objects based on the cluster that they belong to. The main idea in WaveCluster is to transform the original feature space by applying wavelet transform and then find the dense regions in the new space. It yields sets of clusters at different resolutions and scales, which can be chosen based on user needs. The main steps of WaveCluster are shown in Algorithm 1.

Algorithm 1.

Input: Multi-dimensional data objects' feature vectors

Output: clustered objects

1) Quantize feature space, then assign objects to the units.
2) Apply wavelet transform on the quantized feature space.
3) Find the connected components (clusters) in the subbands of transformed feature space, at different levels.
4) Assign labels to the units.
5) Make the lookup table.
6) Map the objects to the clusters.

B. Properties of WaveClustes

- When the objects are assigned to the units of the quantized feature space at step 1 of the algorithm, the final content of the units is independent of the order in which the objects are presented. The next steps of the algorithm will be performed on these units. Hence, the algorithm will have the same results for any different order of input data, so it is order insensitive with respect to input objects.

- As it will be formally and experimentally shown later, the required time for WaveCluster to detect the clusters is linear in terms of number of input data, and it cannot go below that, because all the data should be at least read. After reading the data, processing time will be just a function of number of units in the feature space. Thus, it makes WaveCluster very efficient, specially for very large number of objects. WaveCluster will be specially very efficient for the cases where the number of units n and the number of feature space dimensions d are low.

- Different to other cluster algorithms based on grid, WaveCluster can cluster in multi-scales simultaneously. That is to say, the result of WaveCluster algorithm indicates the quantization of multi-feature spaces, and obtain the cluster collection. So we can choose the cluster according to user's request.

- WaveCluster finds the connected components in the average subband (LL) of the wavelet transformed feature space, as the output clusters. Average subband is constructed by convolving the low pass filter along each dimension and down sampling by two. So a wavelet transformed unit will be affected by the content of units in the neighborhood covered by the filter. It means that the spatial relationships between neighboring units will be preserved. The algorithm to find the connected components, labels each unit of feature space with respect to the cluster that it belongs to. The label of each unit is determined based on the

labels of its neighboring units. It does not make any assumptions about the shape of connected components and can find convex, concave, or nested connected components. Hence WaveCluster can detect arbitrary shape clusters.

- One of the effects of applying low pass filter on the feature space is the removal of noise. WaveCluster takes advantage of this property, and removes the noise and outliers from the feature space automatically.

C. Time Complexity

Let N be the number of objects in the database, where N is a very large number. Assume the feature vectors of objects are d-dimensional, resulting in a d dimensional feature space. Here we suppose that N is very large and d is very small. The time complexity of the first step of WaveCluster algorithm is $O(N)$, because it scans all the database objects and assigns them to the corresponding units, where each dimension A_i in the d-dimensional feature space will be divided into n_i intervals. Assuming n_i = n in each dimension of feature space, there would be $K = n^d$ units[6]. Complexity of applying wavelet transform on the feature space (step 2) will be $O(ldK) = O(dK)$, where l is a small constant representing the length of filter used in the wavelet transform. Because the algorithm supposes that d is very small, so we think d is a constant, and $O(dK) = O(K)$. If we apply wavelet transform to T decomposition levels, to sample in each level downward, the needed time is[7]:

$$O(K + \frac{K}{2^d} + \frac{K}{(2^d)^2} + \cdots + \frac{K}{(2^d)^T})$$

$$= O(K\sum_{i=0}^{T}\frac{1}{(2^d)^i}) = O(K\sum_{i=0}^{T}(2^{-d})^i)$$

$$= O(K\frac{1-(2^{-d})^{T+1}}{1-2^{-d}}) \le O(\frac{4}{3}K)$$

That is, the price of wavelet transform is less than $O(\frac{4}{3}K)$. This indicate that it is very effective to use the multi-resolution cluster. To find the connected components in the feature space, the required time will be $O(cK) = O(K)$, w h ere c is a small constant. Making the lookup table requires $O(K)$ time. After reading data objects, the processing of data is performed in steps 2 to 5 of the algorithm. Thus the time complexity of processing data (without considering I/O) would in fact be $O(K)$, which is independent of number of data objects (N). The time complexity of the last step of WaveCluster algorithm is $O(N)$. Since we assume this algorithm is applied on very large databases, that is $N \ge K$, so $O(N) > O(K)$, thus the overall time complexity of the algorithm will be $O(N)$.

IV. IWCA ALGORITHM

WCA request the input object's number be high. We can find the signal high frequency and low frequency part only by this. But the time complexity $O(K) = O(n^d)$ forms the exponential order increase along with the increase of dimension. It is harder to find connect unit because that the number of adjacency unit is larger. So, in this chapter, we introduce a improved WaveCluster algorithm – IWCA algorithm, in view of the time complexity of high-dimensional data（$N < K = n^d$）.

A. IWCA Algorithm

Algorithm 2.

Input: High-dimensional data objects' feature vectors

Output: clustered objects

1) Quantize d-dimensional feature space X, then assign objects to the units.
2) Make the number of space dimension reduced from d to m by using ICA Algorithm on the quantized feature space $(m \ll d)$, and obtain a new m-dimensional feature space Y. Y retains important information of the original feature space as far as possible.
3) Obtain feature space Y' by applying wavelet transform on the feature space Y.
4) Find the connected components (clusters) in the subbands of transformed feature space Y', at different levels.
5) Assign labels to the units.
6) Make the lookup table.
7) Map the objects to the clusters

（1）Quantization

The first step of algorithm is to quantize the d-dimensional feature space X, the unit of the object is decided by their characteristic value. In Chapter Ⅲ, there are $K = n^d$ units in the feature space. If $l_{ij} \le o_{ki} \le h_{ij}$, $1 \le j \le d$, then unit $c_i = \langle c_{i1}, c_{i2}, \cdots, c_{id} \rangle$ contains the object $o_k = \langle o_{k1}, o_{k2}, \cdots, o_{kd} \rangle$. Where $c_{ij} = \left[l_{ij}, h_{ij} \right)$ is right open-interval of A_j. To each unit, count the objects' accumulation. The number (or size) of these units is an important issue that affects the performance of clustering.

（2）ICA Process of reducing dimension

Carry on processing to the original d-dimensional feature space X by using ICA Algorithm, and obtain a new m-dimensional feature space Y.

（3）Transform and Cluster

In the third step, discrete wavelet transform will be applied on the m-dimensional feature space Y which has been reduced dimension. Applying wavelet transform on the units $\{c_j : 1 \le j \le \xi\}$ results in a new feature space Y' and hence new units $\{t_k : 1 \le k \le \kappa\}$. Given the set of units $\{t_k : 1 \le k \le \kappa\}$, algorithm detects the connected components in the transformed feature space. Each connected component is a set of units $\{t_k : 1 \le k \le \kappa\}$ and is considered as a cluster. Corresponding to each resolution r of wavelet transform, there would be a set of clusters ζ_r, where usually at the coarser resolutions, number of clusters is less. Average subbands (feature spaces) give approximations of the original feature space at different scales, which help in finding clusters at different levels of details.

（4）Label and Make Look Up Table

Each cluster ω, $\omega \in \zeta_r$, will have a cluster number ω_n. In the fourth step of algorithm, it labels the units in the feature space that are included in a cluster, with its cluster number. That is, $\forall \omega \ \forall t_k$, $t_k \in \omega \Rightarrow l_{t_k} = \omega_n$, $\omega \in \zeta_r$, where l_{t_k} is the label of the unit t_k. The clusters that are found are in the transformed feature space and are based on wavelet coefficients. Thus, they cannot be directly used to define the clusters in the original feature space. IWCA algorithm makes a lookup table LT to map the units in the transformed feature space to the units in the original feature space. Each entry in the table specifies the relationship between one unit in the transformed feature space and the corresponding unit(s) of the original feature space. So the label of each unit in the original feature space can be easily determined. Finally, algorithm assigns the label of each unit in the feature space to all the objects whose feature vector is in that unit, and thus the clusters are determined. Formally, $\forall \omega \ \forall c_j$, $\forall o_i \in c_j$, $l_{o_i} = \omega_n$, $\omega \in \zeta_r$, $1 \le i \le N$, where l_{o_i} is the cluster label of object O_i.

B. Properties of IWCA

Our algorithm retains the related characteristics of original WaveCluster algorithm.

- It is order insensitive with respect to input objects.
- Processing time is just a function of number of units in the feature space.
- Choose the cluster according to user's request.
- It can detect arbitrary shape clusters.
- It can remove the noise and outliers from the feature space.

Moreover, as a result of ICA, we can cluster the high-dimensional data. It is impossible to traditional wavelet analysis.

C. Time Complexit

In IWCA algorithm, the time complexity in step 2 is changed into $O(n^m)$ from $O(K) = O(n^d)$, because we carry on the processing to reduce the dimension. m is very small, so $O(n^m) < O(N)$. That is, in our algorithm, the time complexity is $O(N)$, even to high-dimensional data cluster.

V. EXPERIMENTS

The experimental procedures are achieved in Matlab 7.0. The experimental platform is the computer with Intel Pentium Processor 1.60GHz CPU, 512MB Memory, and Windows XP Professional Service Pack 2 Operation System.

We take 100000 records as original samples from KDD Cup 1999[8] data sets, in which 97800 records are normal data, and 2200 records are intrusion data. The intrusion data is 2.2% of the total records, and is far less than the normal connection numbers. The process is divided into three stages. They are data pretreatment, training stage, and test stage.

Firstly, by using ICA algorithm, we carry on processing to reduce dimension of these 100000 records. And project it to low-dimensional data space under the premise of retaining the original data information (The original data has 41 dimension). Then, carry on processing to wavelet analysis of obtained low-dimensional data. Finally, 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 [9]. The results are shown in table Ⅲ and table Ⅵ.

TABLE I. EXPERIMENT RESULTS OF USING COHEN-DAUBECHIES-FEAUVEAU (2,2) WAVELET

	Correct	False	Total
Normal	97580	220	97800
DOS	248	192	440
U2R	367	113	480
R2L	257	143	400
Probe	688	192	880
Total	99140	860	100000

TABLE II. EXPERIMENT RESULTS OF USING COHEN-DAUBECHIES-FEAUVEAU (4,2) WAVELET

	Correct	False	Total
Normal	97552	248	97800
DOS	234	206	440
U2R	338	142	480
R2L	265	135	400
Probe	701	179	880
Total	99090	910	100000

TABLE III. DETECTION RESULTS OF THE ALGORITHM IN [9]

	Total detection rate		Classificated Intrusion detection rate			
	Detection rate	False Alarm Rate	DOS	U2R	R2L	Probe
Cohen-Daubechies-Feauveau (2,2) wavelet	99.14%	0.86%	56.36%	76.46%	64.25%	78.18%
Cohen-Daubechies-Feauveau (4,2) wavelet	99.09%	0.91%	53.18%	70.42%	66.25%	79.66%
Mean of Detect Rate	99.12%	0.88%	54.77%	73.44%	65.25%	78.92%

According to the result of the experiment by the algorithm in [9], we can see that the mean of detection rate is 99.12%, and the mean of false alarm rate is 0.89%.

TABLE IV. EXPERIMENT RESULTS OF USING COHEN-DAUBECHIES-FEAUVEAU (2,2) WAVELET

	Correct	False	Total
Normal	97631	169	97800
DOS	298	142	440
U2R	378	102	480
R2L	302	98	400
Probe	723	157	880
Total	99332	668	100000

TABLE V. EXPERIMENT RESULTS OF USING COHEN-DAUBECHIES-FEAUVEAU (4,2) WAVELET

	Correct	False	Total
Normal	97641	159	97800
DOS	361	79	440
U2R	377	103	480
R2L	289	111	400
Probe	792	88	880
Total	99460	540	100000

TABLE VI. DETECTION RESULTS OF THE ALGORITHM IN THIS PAPER

	Total detection rate		Classificated Intrusion detection rate			
	Detection rate	False Alarm Rate	DOS	U2R	R2L	Probe
Cohen-Daubechies-Feauveau (2,2) wavelet	99.33%	0.67%	67.27%	78.75%	75.5%	82.16%
Cohen-Daubechies-Feauveau (4,2) wavelet	99.46%	0.54%	82.05%	78.54%	72.25%	90%
Mean of Detect Rate	99.4%	0.6%	74.66%	78.65%	73.88%	86.08%

According to the result of the experiment by the algorithm in this paper, we can see that the mean of detection rate is 99.4%, and the mean of false alarm rate is 0.6%. This shows that the IWCA algorithm in this paper could detect the intrusion data better, and keep a lower false positive.

VI. CONCLUSION

In this paper, we introduce a improved WaveCluster algorithm (IWCA), based on to the analysis of ICA. It can be used in data mining. The algorithm can cluster the high-dimensional data, and the time complexities of it still is $O(N)$. In paper, we carry on the simulation experiment by Matlab. The experimental results show that IWCA algorithm is more effective than other algorithms.

REFERENCES

[1] Gursoy, M.I.; Subast, A. A comparison of PCA, ICA and LDA in EEG signal classification using SVM. Signal Processing, Communication and Applications Conference, 2008. SIU20-22 April 2008 Page(s):1 – 4.

[2] A. Hyvärinen and E. Oja. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 2000, 13(4-5):411-430.

[3] Bell A J, Sejnowski T J. The "independent components" of natural scenes are edge filters[J]. Vision Res. 1997, Vol. 37, No. 23, P. 3327~3338.

[4] Hoyer P, Hyv_arinen A. Independent component analysis applied to feature extraction from colour and stereo images[J]. Network: Comput. Neural Systems, 2000, 11(3): 191~210.

[5] Gholamhosein Sheikholeslami , Surojit Chatterjee, Aidong Zhang, WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases, Proceedings of the 24rd International Conference on Very Large Data Bases, August 24-27, 1998, P.428-439.'

[6] Guowei Wu; Lin Yao; Kai Yao: An Adaptive Clustering Algorithm for Intrusion Detection. International Conference on Information Acquisition, 20-23 Aug. 2006 Page(s):1443 – 1447.

[7] Mingwei Zhao; Yang Liu; Rongan Jiang; Research of WaveCluster Algorithm in Intrusion Detection System. 2008 International Conference on Computational Intelligence and Security. Volume 1, 13-17 Dec. 2008 Page(s):259 – 263.

[8] KDD Cup 1999 data. University of Calfornia, Irvine. http://kdd.ics.uci.edu/databaseds/kddcup99/kddcup99.html,1999.

[9] Dangfeng Zhu, "Research of Intrusion Detection Technique Based on Wavelet Neural Networks: (Master's degree thesis)," Lanzhou, Lanzhou University, 2006.