7659, England, UK
Journal of Information and Computing Science
Vol. 2, No. 3, 2007, *pp*. 197-202

High Dimensional and Large Numbers of Data Clustering Method Based Sensitive Subspace *

Wencang Zhao +

College of Automation and Electronic Engineering, Qingdao University of Science & Technology, Qingdao 266042, China

(*Received 4 August 2006, Accepted 25 November 2006*)

Abstract. Clustering is the main method to analyse the large numbers of data, but when the data's dimension is higher, the consumed time increases exponentially. We put forward an effective clustering method for high dimensional and large numbers of data, which is based on the sensitive subspace consisting of the data set's sensitive dimensions. In order to build the sensitive subspace, we first estimate the probability density of each dimension, enhance its optional ability through extracting zero and smoothness processing, then through recognizing the number of the rallying points to gain the sensitive dimensions, and last do the Rival Penalized Competitive Learning (RPCL) clustering in the subspace. Moreover, we detected the red tide of hyper-spectral data using this method, which proved it could effectively get similar results with one-ninth time.

Keywords: Data clustering, Sensitive subspace, Probability density, High dimensional data, Large numbers of data, hyperspectral data

1. Introduction

The aerial remote sensing hyper-spectral data, whose dimension is 124 or 246, is a kind of high dimensional data. The data clustering is the main analysis method for the information extracting, but the efficiency of the conventional clustering methods is too low to deal with this kind of data set. Various clustering methods have been studied in order to solve the problem. [1] designed a scheme for clustering points in a high dimensional data sequence for the subsequent indexing and similarity search to save the clustering time; a new neural network architecture (PART) and a resulting algorithm was proposed in [2] to find the projected clusters for data sets in high dimensional spaces; and [3] used rough sets and fuzzy C-means clustering methods to reduce the dimension of the hyper-spectral band in order to quicken the analysis implementation.

The RPCL clustering method was put forward in [4]-[8], it had the ability of automatically allocating an appropriate number of units for an input data set, and it was propitious to analyse the large numbers of data set. When it is used to analyze the high dimensional data, the consumed time is exponentially longer as the data dimension increase.

This paper presents a quicker clustering method based the RPCL of the sensitive subspace to analyze this kind of data. In order to reduce the dimension of the data set, we calculate the probability density of every dimension and design a method to select the sensitive dimensions, then compose the sensitive subspace, and last, cluster the data set in the subspace. This method could not only enhance the efficiency of cluster implementation but also boost its convergence, and the results are as good as the results using whole data.

2. Convergence Method of RPCL

* Supported by the National High Technology Development 863 Program of China under Grant No. 2001AA630308 and the Doctoral Fund of Qingdao University of Science & Technology.
+ *E-mail* address: zhaocenter-journal@yahoo.com.cn

2.1. Overview of RPCL Algorithm

The RPCL is an unsupervised learning rule for clustering analysis [6]. Without pre-assigning the number of clustering, the RPCL randomly takes m data points as the seed $\{q_j\}_{j=1}^{m}$ among the data set $\{X_t\}_{t=0}^{N}$. For each data point X_t, the RPCL not only modifies the seed point of winner to adapt to X_t, but also penalizes the rival one by one smaller learning rate. The implementation course can be summarized as follows:

Step 1

Randomly take a data point X_t, and let the indicator

$$\mu_j = \begin{cases} 1, & \text{if } j = w = \arg\min_i \gamma_i \|X_t - q_i\| \\ -1, & \text{if } j = r = \arg\min_{i \neq w} \gamma_i \|X_t - q_i\| \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Where γ_i is the relative winning frequency of the seed point q_i on data set D.

Step 2

Update the seed point q_i by

$$q_j^{new} = q_j^{old} + \Delta q_j \tag{2}$$

with

$$\Delta q_j = \begin{cases} \alpha_w (X_t - q_j), & \text{if } \mu_j = 1 \\ -\alpha_r (X_t - q_j), & \text{if } \mu_j = -1 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $0 < \alpha_r << \alpha_w < 1$ are the learning rates for the winner and rival seed points.

Step 3

Repeat Steps 1 and 2 until the extra seed points are driven far away from the data set D, or simply stop iterations when the indications are kept unchanged for all $X_t \in D$.

2.2. Convergence Method of RPCL for Hyper-spectral Data

We randomly took 10 points from the hyper-spectral data set as the seeds $\{q_j\}_{j=1}^{10}$ for the first RPCL implementation course. Then took the center of each cluster as the seed points $\{q_j\}_{j=1}^{n}$ for the next process and here the seed points would preferably reflect the real center of every cluster; we then ran it 39 times repeatedly. We summed the changed number of every cluster after each RPCL and showed the variational curve in fig.1. From the changing trend of the line, we could deduce a convergence method of RPCL for the high dimensional hyper-spectral data and then elicit the reasonable RPCL clustering times. For this experiment, the clustering process would be stabilized after 29 times.

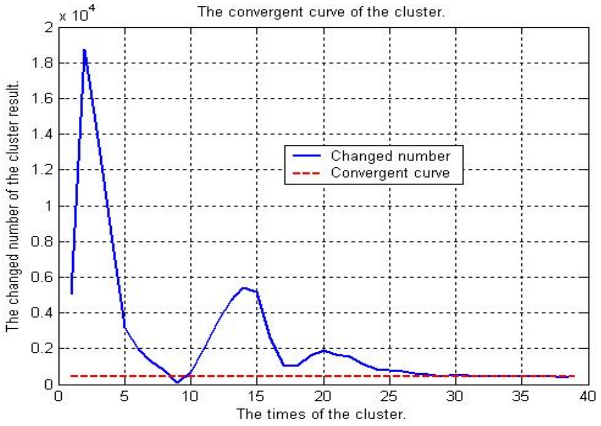


Fig. 1. The variational trendline of the sum of the changed number of every cluster after each RPCL course fluctuates greatly at the beginning and then converges at 450 after 29th RPCL.

3. RPCL Clustering Based Sensitive Subspace

3.1. Calculating the Sensitive Subspace of Hyper-spectral Data

The hyper-spectral data has more than 100 dimensions. If the data could be classified into several clusters, it should contain some dimensions that are sensitive to classify. So, we could select the sensitive dimensions to compose the sensitive subspace to do the RPCL clustering.

In order to save time and analyze expediently, we use the parzen window algorithm to compute every dimension's PDE. For example, the PDE of the dimension 24 and 94 are shown as fig2. In order to select the sensitive dimensions automatically, we do the following two steps to enhance their optional ability.

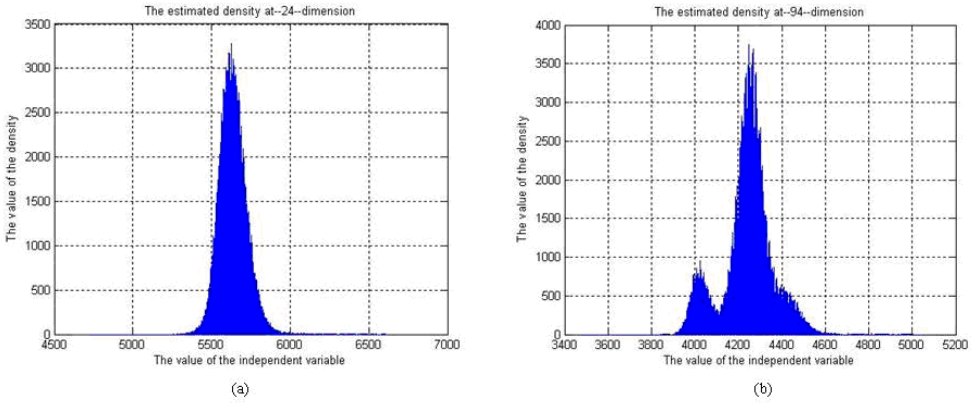


Fig. 2. The original PDE results

Step 1: Extracting 0 Processing

We check all the points' value of the PDE and delete the points whose value are 0, then get the PDE after the extracting processing. Fig3 shows the results after the processing. We will find that the dimensions' PDE are recognized easily than the original ones, but they have some fluctuations yet.

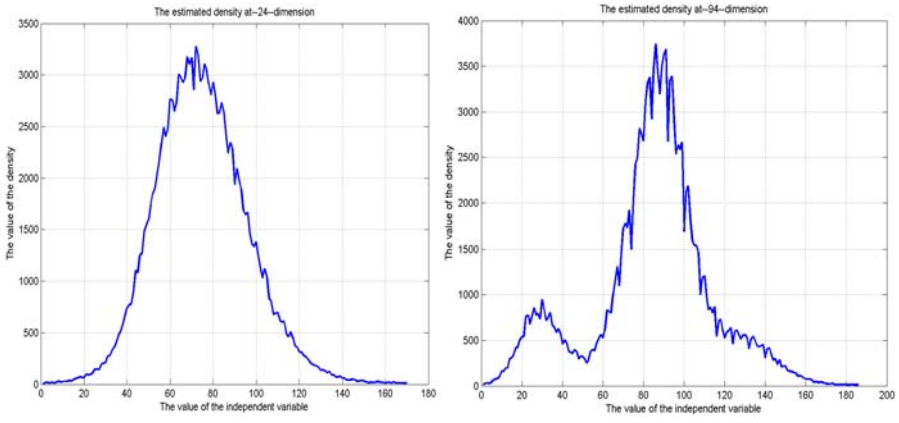
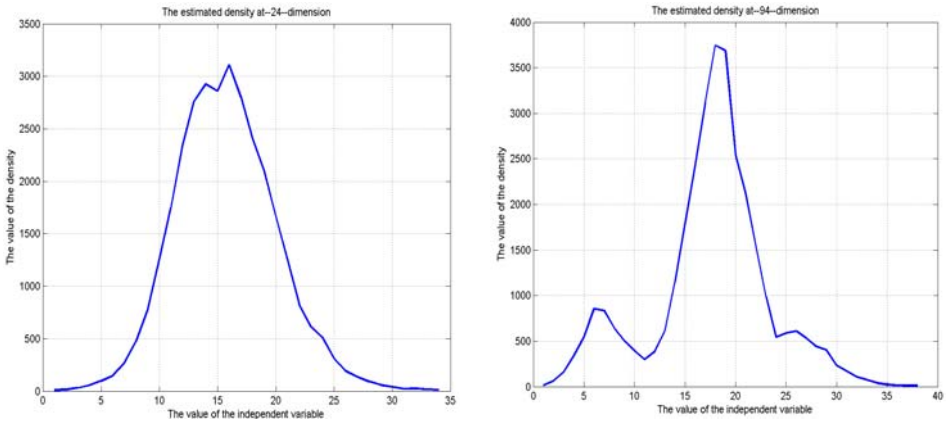


Fig. 3. The PDE results after the extracting 0 processing.

Step 2: Smoothness Processing

In order to smooth the PDE's curve, we use the down symbol method to do so. Here we down one symbol every five ones. The finally results of the dimensions 24 and 94 are as fig4 shows. Now we may find that it is very easy to recognize the rallying points' number automatically by the configurational feature of the PDE's wave shape [10].



(a) (b)

Fig. 4. The PDE results after the smoothness processing. (a) is the dimension that can not be separated easily, and (b) is the one that is sensitive and it could compose the sensitive subspace of the high dimensional data set.

We analyze all the 124 dimensions of the data set using this method, and gain the high dimensional data's sensitive dimensions that are 60-70, 77-85 and 94-95 (tot: 22). The 22 dimensions make up of the data's sensitive subspace. Now the dimension of the hyper-spectral data is reduced from 124 to 22, so the method can cut down the dimension effectively.

4. Experiments and Analysis

We did 124 RPCL clustering experiments using different data sets whose dimensions were ranging from 1 to 124 and recorded their consumed time. Fig.5. shows the variable curve of these consumed times. We can see the consumed time exponentially increase along with the data's dimension increasing and the time manifold quicken after the dimension of than 60. We did many experiments on the hyper-spectral data obtained in different time and places through the sensitive subspace's calculating method, and found that the dimension of the subspace was less than 40 despite of the original data with 124 or 246 dimension. So, the method based on the sensitive subspace can save more than 1000 seconds during one time RPCL implementation.

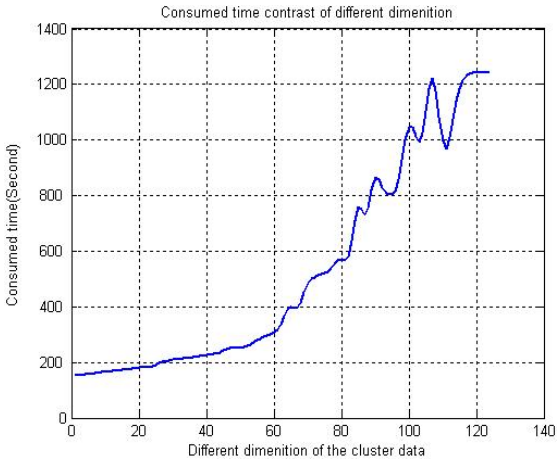


Fig. 5. The consumed time of one times RPCL implementation exponentially increases along with the increasing of the high dimensional data's dimension.

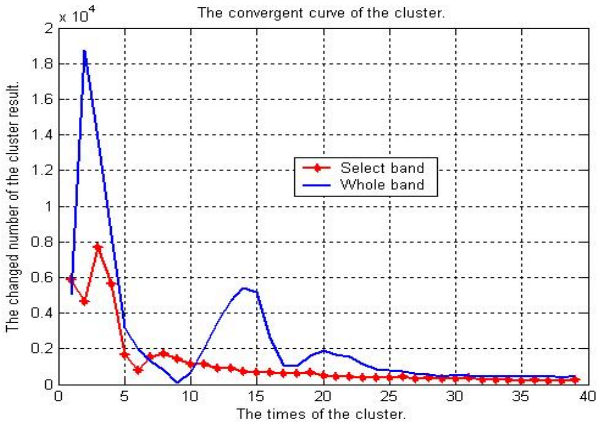
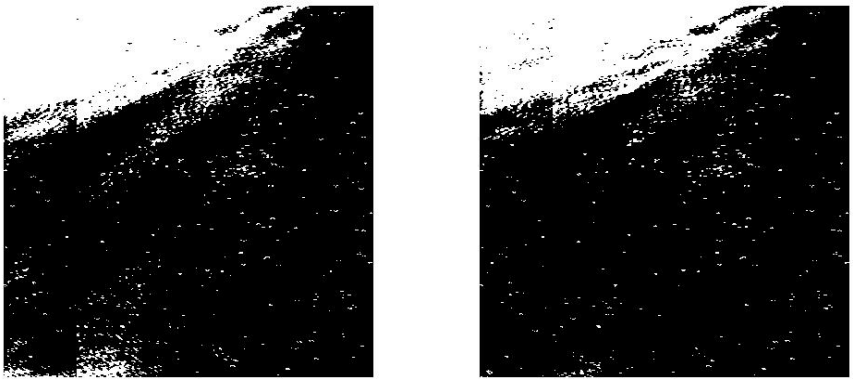


Fig. 6. The convergence courses of RPCL based sensitive subspace and original data, the dashdotted line is the one based sensitive subspace, which is relatively evener than the original data one's and converges at 20th, and the other one converges at 29th.

Furthermore, the convergence of the RPCL clustering method, based on the sensitive subspace, is superior to the course using the original data; fig 6 shows the convergence courses. The dashdotted line, based the sensitive subspace, is relatively straighter than the original data line and converges at 20th RPCL, but the convergence line based on the original data converges at 29th RPCL. So, the method not only needs fewer RPCL runs, but also has little error during most of the clustering.

Finally, the classification results of the RPCL clustering based on the sensitive subspace are primarily the same as the ones of the RPCL used the original data, which are showed in fig 7. Here (a) are the results of the method based the sensitive subspace, which ran 20 RPCL implementation and (b) is the ones of the original data, which ran 29 RPCL implementation. In fig 7, the white areas denote the red tide scopes and the others are the ocean water areas. In this experiment, the consumed time of our method is 20×180 seconds and the other one's time is 29×1240 seconds; so, this method can save 32 360 seconds.



(a) (b)

Fig. 7. The clustering results of the RPCL based the sensitive subspace compare with the ones using the original data. Fig (a) is the results of the method based sensitive subspace which needs 20 times RPCL and fig (b) is the ones of the method using original data which need 29 times RPCL. The dissimilarity of the results is slim, but the first ones may save more than 1000×9 seconds time.

5. Conclusions

In order to suit the demand of the analysis of hyper-spectral data set in large numbers, we put forward a quickly clustering method based the RPCL of sensitive subspace. There are two contributions in this letter. The first contribution concerns a method which calculates the changed number of the cluster after each RPCL clustering implementation to judge the convergence of the classification process. The second contribution is a proposed method to automatically recognize the sensitive dimensions of the data set to compose the sensitive subspace and reduce the data's dimension, and then the RPCL clustering based on the

sensitive subspace is done. The experiments proved that the convergence rule could decide the times of the cluster course effectively, and the RPCL based the sensitive subspace could not only save nearly 9 times of time, but also gets primarily the same results as that of the RPCL using the original data.

6. References

[1] S. L. Lee, C. W. Chung. On the effective clustering of multidimensional data sequences. *Information Processing Letters*. 2001, **80**: 87-95.

[2] Y. Q. Cao, J. H. Wu. Projective ART for clustering data sets in high dimensional spaces. *Neural Networks*. 2002, **15**: 105-120.

[3] H. Shi, Y. Shen, Z. Y. Liu. Hyperspectral Band Reduction Based on Rough Sets and Fuzzy C_means Clustering. *Journal of Electronics & Information Technology*. 2004, **26**: 619-624.

[4] L. Xu, A. Krzyzak, E. Oja. Rival Penalized Competitive Learning for Clustering Analysis, RBF Net. And Curve Detection. *IEEE Transaction on Neural Networks*. 1993, **4**: 636-649.

[5] L. Xu. Bayesian Ying-yang machine, clustering and number of clusters. *Pattern Recognition Letters*. 1997, **18**: 1167-1178.

[6] Y. M. Cheung, L. Xu. Rival Penalized Competitive Learning Based Approach for Discrete-valued Source Separation. *International Journal of Neural Systems*. 2000, **10**: 483-490.

[7] L. Xu. BYY harmony learning, structural RPCL, and topological self-organizing on mixture models. *Neural Networks*, 2002, **15**: 1125-1151.

[8] Y. M. Cheung. Rival Penalization Controlled Competitive Learning for Data Clustering with Unknown Cluster Number. *Proceedings of 9th International Conference on Neural Information Processing.* Singapore, 2002, 18-22.

[9] R. O. Duda, P. E. Hart, D. G. Stork. *Pattern Classification*, 2nd edi. Indianapolis: Wiley-Interscience, 2001, 134-135.

[10] W. C. Zhao, G. R. Ji. Ocean Red Tide Recognition Method Based Absorbing and Reflecting Crest of Hyper-spectral Images. *Acta Oceanologica Sinica*, (In press).