Semi-supervised Hierarchical Clustering Analysis for High Dimensional Data

Yuntao Qian, Xiaoxu Du, and Qi Wang

College of Computer Science, Zhejiang University
Hangzhou 310027, China

ytqian@zju.edu.cn, Duxiaoxu168@hotmail.com, wangqi@hz.cn

Abstract

In many data mining tasks, there is a large supply of unlabeled data but limited labeled data since it is expensive generated. Therefore, a number of semi-supervised clustering algorithms have been proposed, but few of them are specially designed for high dimensional data. High dimensionality is a difficult challenge for clustering analysis due to the inherent sparse distribution, and most of popular clustering algorithms including semi-supervised ones will be invalid in high dimensional space. In this paper, a semi-supervised hierarchical clustering algorithm for high dimensional data is proposed, which is based on the combination of semi-supervised clustering and dimensionality reduction. In order to achieve high harmony between dimensionality reduction and inherent cluster structure detection, the number of dimensions is reduced sequentially as the clusters are gradually formed in the hierarchical clustering procedure. The experimental results show the effectiveness of our method.
Keywords: Clustering analysis, dimensionality reduction, semi-supervised learning, high dimensional space

I. Introduction

As the data mining problem often involves a large amount of unlabeled data and a relative small amount of labeled data, consequently, learning with both labeled and unlabeled data that is known as semi-supervised learning, has become a topic of significant recent interest [1,2]. In this paper, we focus on semi-supervised clustering in which labeled data is employed to improve the clustering performance.

Some semi-supervised clustering algorithms have been presented [3,4], but few of them are specially designed for high dimensional data. In most of clustering applications such as image processing, pattern recognition, computational biology, and web information retrieval, the data need to be processed are always in high-dimensional space. High dimensionality not only makes computational cost very expensive, but also makes many popular clustering algorithms invalid due to sparse

density distribution. Therefore, the curse of dimensionality must be given a significant amount of research attention in semi-supervised clustering.

Dimensionality reduction is thought as an effective way to solve high dimensional problem. In most cases, dimensionality reduction is carried out as a preprocessing step, for example, linear/nonlinear discriminant analysis (LDA/NDA) and principal component analysis (PCA) are popular used in classification and clustering problems respectively [5]. However, PCA does not always enhance the quality of clustering since PCA based dimensionality reduction may destroy the inherent cluster structure in most cases. Recently spectral embedding and manifold learning such as spectral clustering [6], Neighbourhood component analysis [7], Isomap [8], and locally linear embedding [9], are used instead of PCA, which can estimate intrinsic dimensions that well reflect clustering structures. The foundation of these methods is dependent on the estimation of local density distribution, but this estimation is very difficult in high-dimensional space. Therefore, most of studies of high-dimensional data clustering always use more complicated schemes to incorporate dimensionality reduction into clustering procedure instead of using dimensionality reduction as preprocessing step, i.e., two problems of partitioning a data set and finding a reduced dimensionalities are solved at the same time.

Aiming at two principal clustering methods of partitioning clustering and hierarchical clustering, the existing dimensionality reduction based clustering methods can be also divided into these two classes. Friedman and Meulman propose a new partitioning criterion based on subsets of attributes, in which the distance between a pair of points is measured on a subset of attributes rather than on all of them simultaneously [10]. Roth and Lange use a classical EM algorithm to combine feature selection and clustering, in which the E-step is estimated by the fuzzy/probabilistic clustering based on Gaussian mixture model, and the M-step is formulated as a LDA problem where the fuzzy/probabilistic membership is used as the information of labeling [11]. Ding et al use cluster membership as the bridge to connect the clusters defined in the reduced dimensional space and those defined in the full dimensional space [12]. Aggarwal and Yu propose a generalized projected clustering method for high dimensional data [13], which finds the best projection for each cluster in such a way that the greatest amount of similarity among the points in that cluster can be detected. Hierarchical clustering technique is used in this method as a main frame, but the idea of partitioning clustering is also used to always associate a current cluster with each of the points. In general, partitioning technique based methods are very sensitive to data distribution. If a distribution does not conform to the presumed prototype, it will become less effective. And hierarchical technique based methods are more versatile since they can identify some irregularly shaped or non-globular clusters, but their computational cost is very high for large data set.

Compared with the above methods, our semi-supervised hierarchical clustering algorithm for high dimensional data has two distinct features:

- Agglomerative hierarchical clustering technique is used to deal with the cluster structures with non-spheric shapes. The labeled examples are used as the initial seeds of cluster. In clustering procedure, the unlabeled examples that are closest

to the existed clusters, are assigned to the corresponding clusters and given a specific cluster label, which is different from the unsupervised hierarchical clustering algorithm that initializes each unlabeled examples as a cluster, then successively agglomerates these clusters. Since we only need to calculate the distance between the residual unlabeled examples and the existed clusters in each iterative phase, the computational cost can be saved.

- In clustering procedure, as the unlabeled examples are gradually transformed to the labeled examples by agglomerative operations, the number of dimensions is also gradually reduced by dimensionality reduction algorithm. Clustering and dimensionality reduction are run as an iterative sequential procedure. It is based on the idea that clustering and dimensionality reduction are dependent on each other, and they are an integrated system. Since both labeled and unlabeled examples exist in clustering procedure, a semi-supervised dimensionality reduction algorithm is proposed, by which the information contained in both of the labeled and unlabeled examples can be utilized to determine a reduced subspace.

This paper is organized as follows. In the next section, our method is introduced in detail. The empirical results are discussed in section III. Finally, we present the conclusions and summary in section IV.

II. Algorithms

The key issue of high dimensional clustering is formulating a suitable mechanism to keep functional harmony between dimensionality reduction and clustering. For clustering with unlabeled examples, although some unsupervised dimensionality reduction or feature extraction methods such as PCA, independent component analysis, factor analysis can preserve the main information of a data set according to their respective focuses, but their criteria are always not consist with the criterion of clustering, i.e., inherit cluster structures are destroyed severely in many cases. We think that some information concerning class labeling may be required as a bridge to reach functional harmony for clustering problem. Obviously, the cluster membership can be utilized instead of class labeling, but there is a dilemma whether to do clustering first or do dimensionality reduction first.

In order to solve this problem, we propose a semi-supervised hierarchical clustering algorithm, in which the dimensionality is gradually reduced with a semi-supervised dimensionality reduction algorithm as clusters are gradually formed. The validity of our method is based on the fact that the importance of the labeling information becomes more and more important as number of dimensions is gradually reduced.

A. Semi-Supervised Dimensionality Reduction

Given a data set $D = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\} \in \mathbf{R}^d \times \{\text{class/cluster}\}$ and n is the total number of data, d is the number of original dimensions, $\mathbf{x}$ is the data vector, and y is the class label. D can be divided into two sets $D = \{L \cup U\}$, where L is the set of labeled data with known class labels and U is the set of unlabeled data. The semi-supervised dimensionality reduction algorithm is an optimization based on these two

types of data, and the criterion of optimization is a combination of LDA and PCA that are popular dimensionality reduction techniques used for labeled and unlabeled data respectively.

For the set of labeled examples *L*, LDA is given by a linear transformation matrix $\boldsymbol{W} \in \boldsymbol{R}^{d\times l}$ maximizing the so-called Fisher criterion

$$J_F(\boldsymbol{W}) = \frac{\boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W}}{\boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}} \tag{1}$$

$$\boldsymbol{S}_b = \sum_{i=1}^{c} p_i (\boldsymbol{m}_i - \boldsymbol{m}_0)(\boldsymbol{m}_i - \boldsymbol{m}_0)^T \tag{2}$$

$$\boldsymbol{S}_w = \sum_{i=1}^{c} p_i E\{(\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^T \,|\, \boldsymbol{x} \in C_i\} \tag{3}$$

$$\boldsymbol{m}_0 = \sum_{i=1}^{c} p_i \boldsymbol{m}_i \tag{4}$$

Each example $\mathbf{x}$ in *L* belongs to a specific class C_i, and *c* is the number of classes; $\boldsymbol{m}_i$ and p_i are the mean vector and *a priori* probability of class C_i respectively; $\boldsymbol{S}_b$ and $\boldsymbol{S}_w$ are the between-class and within-class scatter matrices respectively. The purpose of LDA is to maximize the between-class scatter while simultaneously minimizing the within-class scatter in the reduced *l*-dimensional space.

For the set of unlabeled examples *U*, PCA derives a relative small number of decorrelated linear combinations (principal components) of a set of random variables while retaining as much of the information from the original variables as possible. If we let the mean of the data set be zero, PCA is given by a linear transform $\boldsymbol{W} \in \boldsymbol{R}^{d\times l}$ maximizing the covariance matrix based criterion

$$J_P(\boldsymbol{W}) = \boldsymbol{W}^T \boldsymbol{R}_{\boldsymbol{xx}} \boldsymbol{W} \tag{5}$$

$$\boldsymbol{R}_{\boldsymbol{xx}} = E\{\boldsymbol{x}_k \boldsymbol{x}_k^T \,|\, x_k \in U\} \tag{6}$$

PCA can be converted to an eigenvalue problem of $\boldsymbol{R}_{\boldsymbol{xx}}$, and the *l* principal components in $\boldsymbol{W}$ are the eigenvectors corresponding to the *l* largest eigenvalues of $\boldsymbol{R}_{\boldsymbol{xx}}$.

If there exist a hybrid data set with labeled and unlabled examples, a good dimensionality reduction method should make full use of them together. Therefore, the following criterion can be naturally proposed for the hybrid data set $\boldsymbol{D}$

$$J_H(\boldsymbol{W}) = \frac{\boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W}}{\boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}} + \alpha \boldsymbol{W}^T \boldsymbol{R}_{\boldsymbol{xx}} \boldsymbol{W} \tag{7}$$

$\boldsymbol{W}$ can be obtained by maximizing the criterion $J_H(\boldsymbol{W})$. In right side of the above equation, the first term is the criterion of LDA and the second term is the criterion of PCA, and their importance is determined by the weighted variable α.

In order to reduce its computational cost, a simpler criterion is required. Li *et al*. proposed a new feature extraction criterion, the maximum margin criterion (MMC), for labeled data set, which has been proved efficient, accurate, and robust [14]. In this paper, we propose a MCC based semi-supervised dimensionality reduction algorithm.

MMC based dimensionality reduction for the labeled set $\boldsymbol{L}$, makes a example be close to those in the same class but far from those in different classes after the dimensionality reduction. The margin criterion is defined as

$$J_M = \frac{1}{2}\sum_{i=1}^{c}\sum_{j=1}^{c} p_i p_j d(C_i, C_j) \tag{8}$$

$$d(C_i, C_j) = d(\boldsymbol{m}_i, \boldsymbol{m}_j) - v(C_i) - v(C_j) \tag{9}$$

$$d(\boldsymbol{m}_i, \boldsymbol{m}_j) = (\boldsymbol{m}_i - \boldsymbol{m}_j)^T (\boldsymbol{m}_i - \boldsymbol{m}_j) \tag{10}$$

$$v(C_i) = tr(E\{(\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^T \,|\, \boldsymbol{x} \in C_i\}) \tag{11}$$

The margin criterion (8) can be easily simplified to

$$J_M = tr(\mathbf{S}_b - \mathbf{S}_w) \tag{12}$$

The reduced dimensional vector $\boldsymbol{W}$ is obtained by maximizing

$$J_M(\boldsymbol{W}) = tr(\boldsymbol{W}^T(\boldsymbol{S}_b - \boldsymbol{S}_w)\boldsymbol{W}) \tag{13}$$

For unlabeled data set $\boldsymbol{U}$, the margin criterion can be written as

$$J_P(\boldsymbol{W}) = tr(\boldsymbol{W}^T \boldsymbol{R}_{xx} \boldsymbol{W}) \tag{14}$$

The MMC criterion of semi-supervised dimensionality reduction for the hybrid set $\boldsymbol{D}$ is defined as

$$J_H(\boldsymbol{W}) = tr(\boldsymbol{W}^T(\boldsymbol{S}_b - \boldsymbol{S}_w + \alpha \boldsymbol{R}_{xx})\boldsymbol{W}) \tag{15}$$

Let $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \cdots \boldsymbol{w}_l]$ be an orthogonal matrix whose columns satisfy

$$\boldsymbol{w}_i^T \boldsymbol{w}_j = \delta_{ij} \quad (\boldsymbol{W}^T \boldsymbol{W} = \boldsymbol{I}_l) \tag{16}$$

Then we solve the following constrained optimization

$$\begin{aligned} &\max \quad \sum_{k=1}^{l} \boldsymbol{w}_k^T (\boldsymbol{S}_b - \boldsymbol{S}_w + \alpha \boldsymbol{R}_{xx}) \boldsymbol{w}_k \\ &\text{subject to} \quad \boldsymbol{W}^T \boldsymbol{W} = \boldsymbol{I}_l \end{aligned} \tag{17}$$

$J_H(\boldsymbol{W})$ is maximized when $\boldsymbol{W}$ is composed of the first l largest eigenvectors of $\boldsymbol{S}_b - \boldsymbol{S}_w + \alpha \boldsymbol{R}_{xx}$ [15]. Compared with maximizing the criterion (7), the criterion (17) is easier to be optimized.

B. Clustering Algorithm

As the labeled data set L can be used as the seeds of the cluster to build an initial clustering result in the hierarchical clustering procedure, the latter agglomerative operations are simplified to a procedure of the unlabeled examples being assigned to the existing clusters. Therefore, in every iterative clustering phase, the examples are classified into two sets: a current labeled set that includes the examples having been existed in the clusters and a current unlabeled set that includes the examples having not been clustered. At the same time, the semi-supervised dimensionality reduction procedure is completed with these two types of data sets in each iterative phase, and the next agglomerative procedure can be done in the new dimensional space. After the iterative phases, finally all unlabeled examples are assigned to the clusters, and the number of dimensions is decreased to the prior defined value. The detailed algorithm for semi-supervised hierarchical clustering is described as follows.

Algorithm: semi-supervised hierarchical clustering

(Number of clusters: c
Number of data points: n, which contains n_1 labeled points and n_2 unlabeled points, and $n_1 + n_2$
Number of original dimensions: d
Number of reduced dimensions: l)

{$l_k \Rightarrow$ the number of current dimensions; $C_k = \{C_{k1}, \cdots C_{kc}\} \Rightarrow$ current clusters; $n_{2k} \Rightarrow$ the number of current residual unlabeled points}

1. Pick all labeled points as the seeds of clusters, and each initial cluster in $C_0 = \{C_{01}, \cdots C_{0c}\}$ includes the examples with the same class label.
2. While $n_{2k} > 0$

2.1 begin Clustering process
for j=1: Δm
find an unlabeled point $\mathbf{x}_i$ that is closest to one of the current clusters C_{ki}
merge $\mathbf{x}_i$ into C_{ki}
end

$C_k = new\{C_{k1}, \cdots C_{kc}\}$

$n_{2k} = n_{2k} - \Delta m$

end

2.2 begin dimensionality reduction process
reduce the number of dimensionality from l_k to $l_k = l_k - \Delta d$ with the semi-supervised dimension reduction algorithm. The

points in $C_k = \{C_{k1}, \cdots C_{kc}\}$ are used as labeled data, and the residual unlabeled n_{2k} points are used as unlabeled data.

end

3. end

{In each iterative phase, the number of dimensions and unlabeled points are reduced by Δd and Δm respectively}

C. *Parameter Setting*

In our algorithm, some parameters should be set in advance.

- The weighted variable α that determines the importance of the PCA criterion relative to the LDA criterion. Generally the value of α can be given as 1. If $\alpha >> 1$, semi-supervised dimensionality reduction degenerates to PCA, while if $0 < \alpha << 1$, it degenerates to LDA.
- The number of dimensions and the number of unlabeled points, which should be reduced in each iterative phase. If we suppose that the number of dimensions is reduced by Δd in each iterative phase, the number of iterations can be determined by $k = (d - l)/\Delta d$, and the number of unlabeled labeled examples that are assigned to the current clusters is $\Delta m = n_2 / k$ in each iterative phase. In our experiments, Δd is valued by 1. If we want to save computational time, i.e., decreasing the number of iterations, Δd can be valued by a large positive integer.

III. Experimental Evaluation

We illustrate the semi-supervised dimensionality reduction based hierarchical clustering on several data sets of UCI machine learning repository. The performance measure is based on the matching between the clustering result and the classification benchmark

$$\text{Correct Rate} = \frac{\sum_{i=1}^{c} \max_{Y_j \text{ has not been matched with } C_1 \cdots C_{i-1}} \{C_i \cap Y_j\}}{\sum_{i=1}^{c} C_i} \tag{18}$$

$Y = \{Y_1, \cdots, Y_c\}$ is the classification benchmark.

For the semi-supervised clustering, we randomly select a part of examples in the dataset to build a labeled set *L*, and the rest examples are used as the unlabeled set *U*. The proportion of labeled examples in the whole data size is defined as 5% in our experiments.

In the following experiments, we compare three clustering methods: semi-supervised hierarchical clustering on the original dimensional space without dimensionality

reduction (M1), semi-supervised hierarchical clustering on the reduced dimensional space that is derived by unsupervised dimensionality reduction method - PCA (M2), and our semi-supervised hierarchical clustering (M3). The mean and minimum distance measures are chosen as the distance measure between an unlabeled point and a cluster.

- Pima Indians Diabetes Diagnoses data set (table 1): This data set includes 2 classes, 765 instances, and 8 attributes.
- Johns Hopkins University Ionosphere data set (table 2): This data set includes 2 classes, 351 instances, and 34 attributes.
- Optical Recognition of Handwritten Digits data set (table 3): This data set includes 10 classes, 3822 instances, and 64 attributes.

From the above three experiments, it is shown the performance of our method is always better than the other two methods. The value of α affects the clustering performance, so this parameter needs to be carefully chosen. In general, $0.1 < \alpha < 1$ is suitable, that is the labeling information should be given a large weight relative to unlabeling information.

IV. Conclusions

In this paper, the problem of semi-supervised clustering for high dimensional data set is discussed. The proposed algorithm is based on semi-supervised hierarchical clustering frame in which the clusters are formed gradually from a small amount of labeled examples as seeds by assigning unlabeled examples to the existed clusters according to their distances. In the hierarchical clustering procedure, dimensionality reduction is incorporated, and the number of dimensions is reduced gradually as the final clusters are formed. The criterion of dimensionality reduction is dependent on both the labeled data in the current clusters and the unlabeled data that have not been assigned to the current clusters. Through the iterative clustering – dimensionality reduction – clustering procedure, the harmony between clustering and dimensionality reduction is reached, and these two tasks are integrated into a harmonious system. The experimental results also demonstrate the effectiveness of our method. However, how to automatically determine suitable values for the parameters in our methods, and how to improve the computational effectiveness for large scale data sets, are need to be further studied in the future.

Table 1. Comparative experimental results on Pima Indians Diabetes Diagnoses data set

Method: clustering methods
Proportion: the proportion of the labeled examples in a data set
Dimension: the final reduced number of dimensions
α : weighted variable
Measure: the type of distance measure
Accuracy: clustering performance measure (correct rate)

Method	Proportion	Dimension	Measure	α	Accuracy
M1	5%	8	mean		63.28%

M2	5%	2	mean		63.15%
M3	5%	2	mean	0.1	69.01%
M3	5%	2	mean	1	67.34%
M3	5%	2	mean	10	63.15%
M1	5%	8	min		70.70%
M2	5%	2	min		69.40%
M3	5%	2	min	0.1	73.18%
M3	5%	2	min	1	72.53%
M3	5%	2	min	10	70.57%

Table 2. Comparative experimental results on Johns Hopkins University Ionosphere data set

Method	Proportion	Dimension	Measure	α	Accuracy
M1	5%	34	mean		71.79%
M2	5%	5	mean		72.65%
M3	5%	5	mean	0.1	86.61%
M3	5%	5	mean	1	82.93%
M3	5%	5	mean	10	73.22%
M1	5%	34	min		68.38%
M2	5%	5	min		69.23%
M3	5%	5	min	0.1	85.47%
M3	5%	5	min	1	83.48%
M3	5%	5	min	10	70.66%

Table 3. Comparative experimental results on Optical Recognition of Handwritten Digits data set

Method	Proportion	Dimension	Measure	α	Accuracy
M1	5%	64	mean		90.14%
M2	5%	8	mean		89.22%
M3	5%	8	mean	0.1	91.29%
M3	5%	8	mean	1	91.31%
M3	5%	8	mean	10	89.64%
M1	5%	64	min		90.37%
M2	5%	8	min		88.63%
M3	5%	8	min	0.1	91.78%
M3	5%	8	min	1	91.45%
M3	5%	8	min	10	91.31%

References

[1] Blum, A., Mitchell, T. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory* (Madison, WI, 1998), 92-100.

[2] Szummer, M., Jaakkola, T. Partially labeled classification with Markov random walks. In *Advances in Neural Information Processing Systems 15* (*NIPS'2002*) (Vancouver, British Columbia, Canada, December 9-14, 2002). MIT Press, 2003.

[3] Basu, S., Bileoko, M., Mooney R. J. Comparing and unifying search-based and similarity-based approaches to semi-supervised clustering, In Proceedings of Int'l Conf. Machine Learning (ICML'2003)(Washington DC, August 2003). 42-49.

[4] Xing, E.P., Ng, A. Y., Jordan, M. I., Russell S. Distance metric learning, with application to clustering with side-information, In Advances in Neural Information Processing Systems 16 (NIPS'2003) (Vancouver and Whistler, British Columbia, Canada, December 8-13, 2003). MIT Press, 2004.

[5] Fukunaga, K. Introduction to Statistical Pattern Recognition. Academic Press, San Diego, California, 1990.

[6] Ng, A.Y., Jordan, M.I., Weiss, Y. On spectral clustering: Analysis and an algorithm. In Advances in Neural Information Processing Systems 15 (NIPS'2002) (Vancouver, British Columbia, Canada, December 9-14, 2002). MIT Press, 2003, 849-856.

[7] Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R. Neighbourhood components analysis. NIPS'2004, Vancouver, British Columbia, Canada, December 13-16, 2004.

[8] Tenenbaum, J.B., Silva, V., Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. Science, 290(December 2000), 2319-2323.

[9] Roweis, S.T., Saul, L.K. Nonlinear dimensionality reduction by local linear embedding. Science, 290(December 2000), 2323-2326.

[10] Friedman, J. H., Meulman, J.J. Clustering objects on subsets of attributes. http://www-stat.stanford.edu/~jhf/. 1998.

[11] Roth, V., Lange, T. Feature selection in clustering problems. In Advances in Neural Information Processing Systems 16 (NIPS'2003) (Vancouver and Whistler, British Columbia, Canada, December 8-13, 2003). MIT Press, 2004.

[12] Ding, D., He, X.F., Zha, H.Y., Simon, H. Adaptive dimension reduction for clustering high dimensional data. In Proceedings of 2nd IEEE Int'l Conf. Data Mining (Maebashi, Japan, Dec 2002). 147-154.

[13] Aggarwal, C.C., Yu, P.S. Finding generalized projected clusters in high dimensional spaces. SIGMOD' 2000, Dallas, Texas, USA, May 16-18, 2000, 70-81.

[14] Li, H.F., Jiang, T., Zhang, K.S. Efficient and robust feature extraction by maximum margin criterion. In Advances in Neural Information Processing Systems 16 (NIPS'2003) (Vancouver and Whistler, British Columbia, Canada, December 8-13, 2003). MIT Press, 2004.

[15] Guo, Y.F., Li S.J., Yang, J.Y., Shu, T.T., Wu, L.D. A generalized Foley-Sammon transform based on generalized fisher discriminant criterion and its application to face recognition, Pattern Recognition Letters, 24(January 2003). 147-158.

Yuntao Qian received the B.E. and M.E. degrees in automatic control from Xi'an Jiaotong University in 1989 and 1992 respectively, and the Ph.D. degree in signal processing from Xidian University in 1996. From 1996 to 1998, he was a postdoctoral fellow in Northwestern Polytechnical University. In April 1998, he joined the College of Computer Science of Zhejiang University where he became a Professor in 2002. From 1999 to 2001, he was a visiting scholar at the Centre of Pattern Recognition and Machine Intelligence, Concordia University, Canada, and at Department of Computer Science, Hong Kong Baptist University. He has published more than 50 technical papers in academic journals and conference proceedings. His present research interests include data clustering analysis, pattern recognition, image processing, wavelet theory, and neural networks.

Xiaoxu Du is a MS student in College of Comupter Science of Zhejiang University since 2003. He obtained his bachelor degree in Management Information System from Huazhong Uninversity of Science and Technology. His major research interests include face recognition, clustering analysis.