International Journal of Hybrid Information Technology
Vol.8, No.6 (2015), pp.195-204
http://dx.doi.org/10.14257/ijhit.2015.8.6.19

Improved K-means Algorithm with the Pretreatment of PCA Dimension Reduction

Hongtao Liu, Chen Fang, Yu Wu, Ke Xu and Tian Dai

School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
Correspondence should be addressed to Hongtao Liu, liuht@cqupt.edu.cn

Abstract

The improvements we have made are to get the optimal K value, to obtain the initial cluster centers and to calculate the distance by the feature weight. Meanwhile, to cater to the characters of dataset, the IWK-means algorithm uses the principal component analysis (PCA) method as a pretreatment to reduce the dimension of dataset. Finally, the proposed method is experimentally validated on the datasets from the UCI Machine Learning Repository and compared with the existing clustering algorithms by the two evaluation criteria of Rand Index and Adjusted Rand Index

1. Introduction

At present the high-dimensional dataset has become the main dataset for scientific analysis, drifting away the trouble with programming based on high-dimensional dataset, it also takes up many computer resources and wastes a lot of time, because many dimensions in the high-dimensional dataset do not affect the outcome or the impact is very small. Therefore, the dimension reduction is particularly important in the treatment of high-dimension dataset. The basic principle of dimension reduction is to let sample points map into a low-dimensional space from the input space by linear or non-linear transformation, so we can obtain a compact low-dimensional representation of original dataset. The transformation can be mainly divided into linear or non-linear, supervised or unsupervised, global or local. Considering the time-consuming and space complexity and the characteristics of source dataset synthetically, we choose a simple and efficient algorithm of dimension reduction, which is the principal component analysis (PCA).

Catering to the existing deficiencies of the K-means algorithm and the characteristics of the high-dimensional dataset, we propose an algorithm which can improve the quality of clustering and choose an appropriate evaluation criterion to verify that the proposed algorithm has a better clustering effect in comparison with other clustering algorithms.

2. Description of the IWK-means Algorithm

In this section, a modified version of K-means algorithm based on the pretreatment of PCA dimension reduction, which is called IWK-means, is proposed to improve the clustering effect of K-means. Firstly, because most of the existing dataset has the characteristic of high-dimension, which might have a little effect on the experimental result, we use the PCA to reduce the dimensionality in order to save time and reduce the complexity in the data processing. Secondly, catering to the deficiencies of K-means cluster algorithm, we improve the distance cost function, which is proposed by Yongsen Li, to determine the optimal value of K, and then we propose an algorithm to calculate the initial cluster centers which are well separated. Finally, using the contribution degrees of each features, which are computed by PCA method, as the weights to calculate the distances in K-means algorithm.

ISSN: 1738-9968 IJHIT

2.1 The Pretreatment of PCA (Principal Component Analysis)

Principal component analysis (PCA) is also known as the Karhunen-Loeve Transform. In statistics, PCA is a technique for simplifying dataset. It is a linear transformation, which converts the data to a new coordinate system, so that the largest variance of any data's projection is in the first coordinate, which is called the first principal component, the second largest variance is in the second coordinate, which is called the second principal component, and so on. Principal component analysis is often used to reduce the dimension of dataset, meanwhile maintain the characteristic which has the largest contribution of the dataset. This is completed through keeping the low-level principal component and ignoring the high-level principal component, so that the low-level components are often able to retain the most important aspect of the data. However, this is not certain, but depended on the specific application.

Principal component analysis is a statistical method for dimension reduction, which by means of an orthogonal transformation that the original random vector whose components are related transform into a new random vector whose components are not related, the representation in the algebra is that the covariance matrix of original random vector transform into a diagonal matrix, the representation in the geometry is that the original coordinate transform into a new orthogonal coordinates, which point to P orthogonal direction that the scatter of their sample points is most open. And then to reduce the dimension of multidimensional variable system, so that it can be converted into a low-dimensional variable system by a high precision, then further transform the low-dimensional system into one-dimensional system by constructing an appropriate value function.

2.2 The Modified K-means Clustering Algorithm

2.2.1 K-means Clustering Algorithm: K-means clustering algorithm is an important means and method to divide or packet processing the spatial data. It divides the research object into several subsets by the spatial distance index according to similarity criterion, so that the difference of elements in the same subset is the least, while the difference of elements in the different subset is the largest. In the process of K-means algorithm, firstly we need to determine the K value by ourselves and randomly select K objects that each of the objects represents the mean or center of one cluster, for the remainder objects, we assign them to the nearest cluster by the distance between the object and the cluster center. Then calculating the mean of each cluster again to get the new cluster centers, and repeating this process until the clustering criterion function is convergent. The procedure of K-means clustering algorithm is introduced as following:

Step 1. Cater to the dataset $\left\{x^1, x^2, \ldots, x^n\right\}$, we randomly select *K* samples as the initial clustering centers $\left(z_1, z_2, \ldots, z_K\right)$;

Step 2. Find the nearest clustering center z_v for each sample x_i, and allocate it to the cluster whose center is z_v;

Step 3. Get the new clustering centers by calculating the mean;

Step 4. Calculate

$$D = \sum_{i=1}^{n}\left[\min_{r=1,2,\ldots,\mathrm{K}} d\left(x_i, z_r\right)^2\right] \tag{1}$$

Step 5. If the value of *D* is convergent, return $\left(z_1, z_2, \ldots, z_K, U\right)$ and terminate the algorithm, otherwise return to the step2.

2.2.2 Determine the Value of K Based On Distance Evaluation Function: To determine the optimal range of the cluster number *K*, $[K_{\min}, K_{\max}]$, $K_{\min}=1$ means the sample is uniform distribution that there is no obvious feature difference, so that the minimum *K* is 2. How to determine the $K_{\max}$, there is no clear theoretical guidance at present, the experimental experience of many scholars gets $k_{\max} \le \sqrt{n}$, where *N* is the sum of all objects. In general, a good clustering partition should reflect the intrinsic structure of dataset as much as possible, so that the similarity between samples in the same cluster is the least, while the similarity between samples from different clusters is the largest. Given this rule, a new calculation scheme will transform the distance cost function into the distance evaluation function for finding the best K value.

K-means clustering algorithm is a method for dividing the data. Usually there has a variety of distance functions which are used in clustering algorithm, here we use the Euclidean Distance:

$$d\left(x_i, x_j\right)=\left(\left|x_{i1}-x_{j1}\right|^2+\left|x_{i2}-x_{j2}\right|^2+\cdots+\left|x_{ip}-x_{jp}\right|^2\right)^{\frac{1}{2}} \tag{2}$$

where $x_i=\left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)$ and $x_j=\left(x_{j1}, x_{j2}, \ldots, x_{jp}\right)$ are two *p*-dimensional dataset.

As the general rule of spatial clustering algorithm, the division of class should let the similarity between samples in the same cluster is the least, while different cluster is the largest, namely the distance between any spatial object and the geometrical cluster center to which the object belongs is less than the distance between the object and other geometrical cluster centers, until the clustering criterion function converges. The clustering criterion function is:

$$E=\sum_{i=1}^{k}\sum_{p\in C_i}\left|p-m_i\right|^2 \tag{3}$$

where *E* is the sum of square-error of all research objects, *p* is the data object, m_i is the mean of C_i.

According to the basic idea above, we construct distance evaluation function and obtain the best K value by getting the minimum of distance evaluation function.

Sample dataset: $S=\{m_1, m_2, \ldots, m_n\}$, *K* is the number of cluster.

Definition 1 $I=\{S, K\}$ is a clustering space, the distance between classes is the sum of distance between the cluster center (the mean of samples in one cluster) and global center (the mean of all samples):

$$D_{out}=\sum_{i=1}^{k}\left|m_i-m\right| \tag{4}$$

D_{out} is the distance between classes, *m* is the mean of all samples, m_i is the mean of samples in the cluster C_i.

Definition 2 $I=\{S, K\}$ is a clustering space, the inner-class distance is the sum of distance between all samples and the center of each cluster:

$$D_{in}=\sum_{i=1}^{k}\sum_{p\in C_i}\left|p-m_i\right| \tag{5}$$

D_{in} is the inner-class distance, *p* is one spatial object, m_i is the mean of samples in the

cluster C_i .

Definition 3 $I=\{S,K\}$ is a clustering space, the K value accesses the optimum when $D_{in}=D_{out}$, we define the distance evaluation function:

$$F(S,K)=\left|\frac{D_{in}}{D_{out}}-1\right|=\left|\frac{\sum_{i=1}^{k}\sum_{p\in C_i}|p-m_i|}{\sum_{i=1}^{k}|m_i-m|}-1\right| \tag{6}$$

when we get the $Min\{F(S,K)\}$, the *K* is the optimum.

2.2.3 Choose the Initial Cluster Centers: According to the deficiencies of K-means clustering algorithm which are mentioned above, the random selection of the initial cluster centers in K-means algorithm makes the effect of each clustering has randomness, on that basis, we improve this aspect of K-means algorithm in this section. This algorithm is based on choosing two attributes from the p attributes which can best describe the change of the dataset as the two axes. The original dataset via the process of PCA dimension reduction, we choose the two dimensions which have the largest contribution as the horizontal and vertical coordinate of the initial cluster centers. After that, we take the mean of all data points which is in the selected coordinate as the center of dataset:

$$m=[\bar{x}_I,\bar{x}_{II}] \tag{7}$$

Where $\bar{x}_I$ is mean of the variable which is selected as the main axis, $\bar{x}_{II}$ is defined similarly. The Euclidean distances

$$d_{im}=\left((x_{iI}-\bar{x}_I)^2+(x_{iII}-\bar{x}_{II})^2\right)^{\frac{1}{2}},\ i=1,2,\ldots,n \tag{8}$$

are calculated between each data point and the center. And the data point with the highest distance from the center will be selected as the first initial cluster center, which is named c_1. Then we calculate the Euclidean distance

$$d_{ic_1}=\left((x_{iI}-x_{c_1I})^2+(x_{iII}-x_{c_1II})^2\right)^{\frac{1}{2}},\ i=1,2,\ldots,n \tag{9}$$

between each data points and c_1 to select the data point for the second initial cluster center which has the farthest distance to c_1, we name it c_2.

To select a next $c_k\ (k\geq 3)$ for the rest initial cluster centers, d_{ic_k} is calculated between each data points and c_{k-1}. The $d_{ik}\ (k\geq 3)$ is the sum of distances between each points and the (k-1)th initial cluster centers. For example, d_{i3} is calculated as following:

$$d_{i3}=d_{ic_1}+d_{ic_2},\ i=1,2,\ldots,n \tag{10}$$

This accumulation scheme can avoid the nearest data points being chosen as the initial cluster centers in different clusters. This ensures that the next initial cluster center is far away from the previous ones as much as possible. The process is repeated until the number of initial cluster center is equal to the value of K.

2.2.4 Feature Weighted K-means: We get the contribution degree of each attributes by the PCA dimension reduction, so we suppose that the contribution degree can be used in the K-means algorithm with the form of weight. As we know, the clustering criterion

function is:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{11}$$

where m_i is the center of the cluster C_i, p is any point in the cluster C_i.

Huang, *et al.*, modified the criterion with an unknown weight w_v:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \sum_{v=1}^{M} w_v^{\beta} |p - m_i|^2 \tag{12}$$

where v is the attribute of dataset, M is the sum of attribute, β is a parameter that describe the influence ratio of the weight to the distance. The criterion can also be rewritten by Minkowski distance:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \sum_{v=1}^{M} w_v^{\beta} |p - m_i|^{\beta} = \sum_{i=1}^{k} \sum_{p \in C_i} \sum_{v=1}^{M} |w_v (p - m_i)|^{\beta} \tag{13}$$

where $\beta = 1$ means that it is CityBlock distance, $\beta = 2$ means that it is Euclidean distance. So we use the contribution degree as the weight w_v.

2.3 The Evaluation Criteria

To compare the clustering effect, we will use the evaluation criteria of the Rand index and the Adjust rand index.

The Rand index in statistics, and in particular in data clustering, is a measure of the similarity between two data partitions. Given a set of *n* elements $S = \{o_1, o_2, \ldots o_n\}$ and two partitions of S to compare, $X = \{X_1, X_2, \ldots, X_k\}$, a partition of S with *k* subsets, and $Y = \{Y_1, Y_2, \ldots, Y_k\}$, a partition of S with *k* subsets, then define the following:

a: the number of elements in S that are in the same set in X and in the same set in Y.

b: the number of elements in S that are in different sets in X and in different sets in Y.

c: the number of elements in S that are in the same set in X and in different sets in Y.

d: the number of elements in S that are in different sets in X and in the same set in Y.

The Rand index is:

$$R = \frac{a + b}{a + b + c + d} \tag{14}$$

Rand index ranges from 0 to 1, where 0 indicates that the two partitions are entirely different, and 1 indicates that the two partitions are identical.

A form of the Rand index may be defined that is adjusted for the chance grouping of elements, this is called the adjusted Rand index. The adjusted Rand index is the corrected-for-chance version of the Rand index. While the Rand index may only range from 0 to 1, the adjusted Rand index can be negative values if the real index is less than the expected index.

The adjusted form of the Rand Index, the Adjusted Rand Index is defined as following:

$$AdRandIndex = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex} \tag{15}$$

more specifically

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}} \tag{16}$$

where n_{ij}, a_i, b_j are the values from the contingency table.

We also can get the value of ARI by $\{a, \mathrm{b}, \mathrm{c}, \mathrm{d}\}$, which is defined above.

$$ARI = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)} \tag{17}$$

3. Experiments and Results

In this part, we do experiments of the IWK-means algorithm on six real datasets from UCI Machine Learning Repository. The experiments of the IWK-means algorithm are conducted by the following four directions:

(1) The pretreatment of original dataset, we use PCA dimension reduction to make the data processing more efficient;

(2) Before the K-means clustering algorithm, we should determine the best value of K;

(3) In the K-means clustering algorithm, we will choose the initial cluster centers first;

(4) We use the contribution degree from PCA dimension reduction as the weight for calculating the Euclidean distance.

To verify the efficiency of our proposed algorithm, we have accomplished many numerical experiments on six well-known real datasets from UCI Machine Learning Repository. All these datasets we used are briefly described in Tab 1.

Table 1. Description of Datasets

Dataset	Records-number	Attributes-number	Clusters-number
Iris	150	4	3
Wine	178	13	3
Glass	214	10	7
WPBC	198	30	2
Ionosphere	351	34	2
Housing	506	13	2

4.2 The IWK-means Algorithm

The K-means clustering algorithm has a ready-made equation in MATLAB, and the equation can carry out clustering analysis by the given initial cluster centers. Therefore, we introduce the points which are calculated by the proposed method as the initial cluster centers, and utilize the contribution degree which is obtained by PCA dimension reduction as the weight of calculating the Euclidean distance in K-means clustering algorithm. The clustering effect drawing of the IWK-means algorithm is shown as

following:

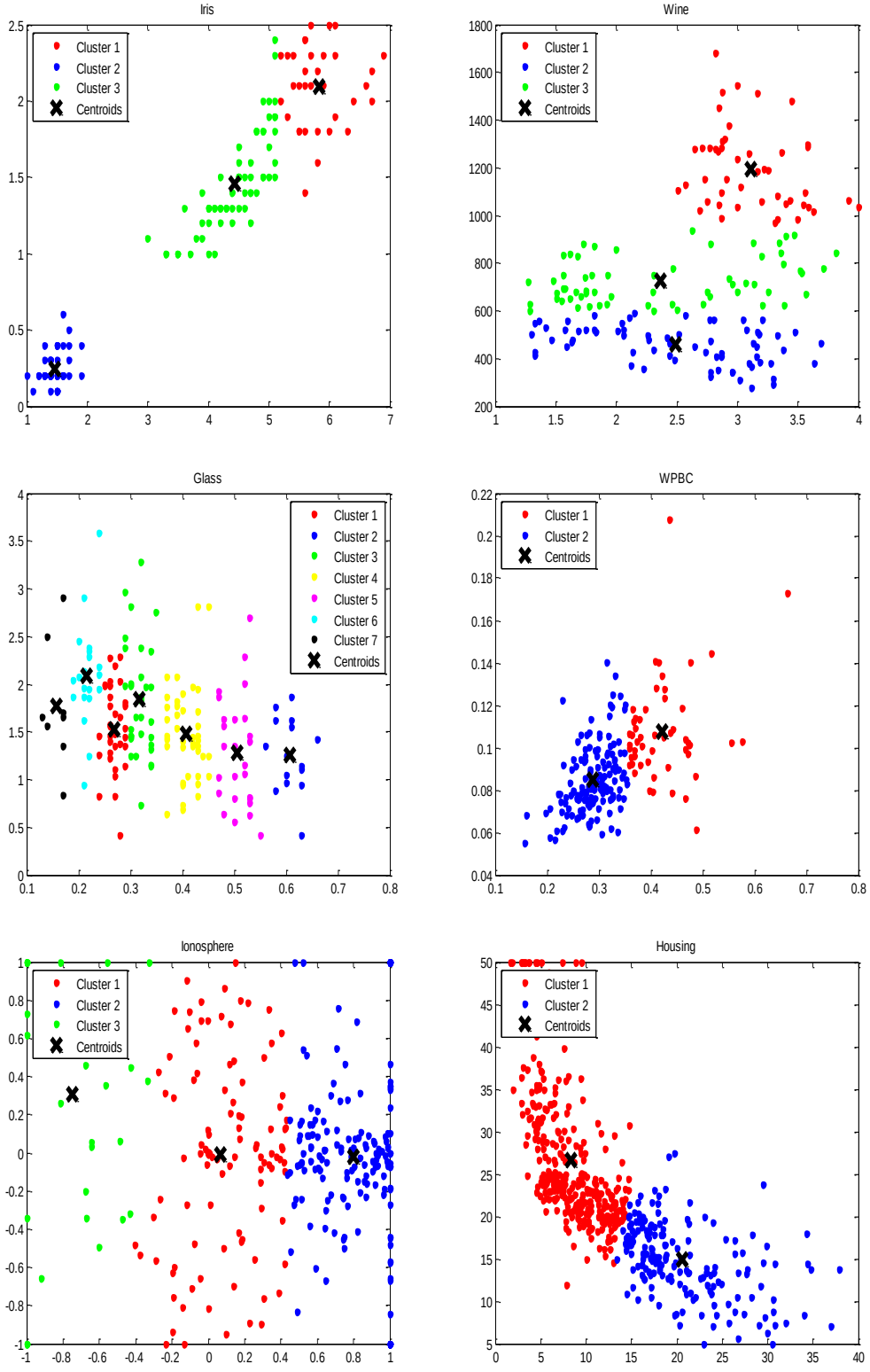

Figure 1. The Clustering Effect of the IWK-means Algorithm

4.3 The Result of the Evaluation Criteria

The effectiveness of the IWK-means algorithm is verified by comparing it with four

other clustering algorithms, which are K-means algorithm, FCM algorithm, FKPC algorithm and MPCK-means algorithm.

1.K-means algorithm. This algorithm is an unsupervised clustering algorithm and suitable for low-dimensional datasets.

2.FCM algorithm. Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters.

3.FKPC algorithm. Fuzzy kPlane Clustering (FKPC) considers the possibility of each sample point belongs to each cluster based on the kPlane Clustering.

4.MPCK-means algorithm. MPCK-means involves both metric learning and constraints satisfaction.

We have got the clustering effect drawings by the IWK-means algorithm above, and we will calculate the evaluation criteria of the Rand Index and the Adjusted Rand Index for comparing the clustering effect of the existing clustering algorithms.

In the Table 2, we show the clustering results of the K-means algorithm in using the initial cluster centers by proposed method for all datasets in comparison to random initialization, that the clustering results of K-means algorithm using proposed method is better.

Table 2. Comparison Results between Proposed Method and Random Initial Centers According To the Rand Index

Dataset	Randomly	Proposed method
Iris	0.8418	**0.8797**
Wine	0.8945	**0.9147**
Glass	0.6693	**0.8543**
WPBC	0.5167	**0.5679**
Ionosphere	0.5691	**0.6213**
Housing	0.5217	**0.7442**

Table 3. The Results of Comparing Clustering Algorithm by Rand Index (The Datasets We Use In the Iwk-Means Algorithm Have Passed the Pretreatment)

Dataset	K-means	FCM	FKPC	MPCK-means	IWK-means
Iris	0.8418	0.8426	0.8797	0.8831	**0.9826**
Wine	0.8945	0.6877	0.5954	0.9059	**0.9912**
Glass	0.6693	0.7177	0.5897	0.5443	**0.9887**
WPBC	0.5167	0.8254	0.8345	0.5121	**0.9987**
Ionosphere	0.5691	0.5937	0.5842	0.5753	**0.9843**
Housing	0.5217	0.5466	0.5013	0.8679	**0.9958**

Table 4. The Results of Comparing Clustering Algorithm by Adjusted Rand Index (The Datasets We Use In the Iwk-Means Algorithm Have Passed the Pretreatment)

Dataset	K-means	FCM	FKPC	MPCK-means	IWK-means
Iris	0.3895	0.3681	0.4515	0.4754	**0.6161**
Wine	0.3064	0.4742	**0.5142**	0.4946	0.1438
Glass	0.2167	0.0358	0.3775	**0.4268**	-0.0146
WPBC	0.3528	0.4589	0.5640	0.5157	**0.7668**
Ionosphere	0.1576	0.2154	0.2854	0.3429	**0.4572**
Housing	0.2541	0.1958	0.2674	0.3865	**0.4704**

Table 3 and Table 4 show that the clustering effect of the IWK-means algorithm is better than the algorithm of FCM, FKPC and MPCK-means by the evaluation criteria of the Rand Index and the Adjusted Rand Index. However, from the Table 5, we know that the Adjusted Rand Index of the IWK-means algorithm does not always have the highest value of every dataset in comparison to the existing clustering algorithms. We need to do a further research for the reason.

4. Conclusions

In this paper we have proposed a modified K-means clustering algorithm, which is called the IWK-means algorithm. Firstly, we carry out a pretreatment for the datasets from UCI Machine Learning Repository, namely the PCA dimension reduction, and then we obtain the low-dimensional datasets which would represent the original datasets as much as possible. Secondly, we get the optimal K value by distance evaluation function, and we use the two attributes which have the two largest contribution degrees as the horizontal and vertical coordinate for calculating the distance between the initial clustering centers, thus we obtain the initial clustering centers. Finally, we will introduce the contribution degree of each attribute which is used as the weight into the formula to calculate the distance in K-means algorithm. Through the three steps we can get the final clustering effect and it is the whole process of the IWK-means algorithm.

The proposed algorithm is very effective, which is shown by the evaluation criteria of the Rand Index and the Adjusted Rand Index. Experimental results show the cluster structures by the IWK-means clustering algorithm as compared to the existing clustering algorithm.

However, the proposed algorithm has three deficiencies. Firstly, for some datasets, the sum of two largest contribution degrees by PCA dimension reduction is less than 95%, we can introduce more attributes to reduce the deviation of experiment, but this experiment needs to determine the initial cluster centers by calculating the two-dimensional distance between each center. Secondly, the value of the Adjusted Rand Index is instable; the values of Wine dataset and Glass dataset are small. Thirdly, the experiment of determining the optimal K value is not applied to the datasets which does not have the specific K value. We should get a further research on these deficiencies.

Acknowledgments

The authors wish to thank the helpful comments and suggestions from the reviewers. We would also like to thank Dr. Jinghao Xue who gave us many help in the research. This research is supported by the National Social Science Foundation of China (Grant no. 13CGL146) and the Science and Technology Commission of Chongqing Province China (Grant no. cstc2012jjA40027) and the Scientific and Technological Research Program of Chongqing Municipal Education Commission (Grant no. KJ130518), and is also supported by the Youth Scientific Research Program of Chongqing University of Posts and Telecommunications (Grant no. A2012-87).

References

[1] J. A. Hartigan, "Clustering algorithms", John Wiley & Sons, Inc., **(1975)**.
[2] J. MacQueen, "Some methods for classification and analysis of multivariate observations", Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, **(1967)**, pp. 281-297. 1967.
[3] K. Pearson, "LIII On lines and planes of closest fit to systems of points in space", The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 2, no. 11, **(1901)**, pp. 559-572.
[4] Y. Xu, *et al.*, "A method for speeding up feature extraction based on KPCA", Neurocomputing, vol. 70, no. 4, **(2007)**, pp. 1056-1061.
[5] Y. Xu, C. Lin and W. Zhao, "Producing computationally efficient KPCA-based feature extraction for classification problems", Electronics letters, vol. 46, no. 6, **(2010)**, pp. 452-453.
[6] Y. Xu and D. Zhang, "Represent and fuse bimodal biometric images at the feature level: complex-matrix-based fusion scheme", Optical Engineering, vol. 49, no. 3, **(2010)**, pp. 037002-037002.
[7] D. Zhang, R. Lukac and X. Wu, "PCA-based spatially adaptive denoising of CFA images for single-sensor digital cameras", Image Processing, IEEE Transactions on, vol. 18, no. 4, **(2009)**, pp. 797-812.
[8] Y. Xu, et al., "An approach for directly extracting features from matrix data and its application in face recognition", Neurocomputing, vol. 71, no. 10, **(2008)**, pp. 1857-1865.
[9] M. Kirby and L. Sirovich, "Application of the KL Procedure for the Characterization of Human Faces", IEEE Trans Pattern Anal Machine Intell, vol. 12, no. 1, **(1990)**, pp. 103-108.
[10] P. N. Belhumeur, J. P. Hespanha and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection", Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 19, no. 7, **(1997)**, pp. 711-720.

