Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2014, Article ID 640406, 8 pages
http://dx.doi.org/10.1155/2014/640406

Research Article

The Application of Pattern Recognition in Electrofacies Analysis

Huan Li,[1] Xiao Yang,[2] and Wenhong Wei[1]

[1] *Dongguan University of Technology, Dongguan 523808, China*
[2] *School of Information Science and Technology, Tsinghua University, Beijing 100084, China*

Correspondence should be addressed to Wenhong Wei; weiwh@dgut.edu.cn

Received 24 February 2014; Accepted 26 April 2014; Published 4 June 2014

Academic Editor: Guiming Luo

Pattern recognition is an important analytical tool in electrofacies analysis. In this paper, we study several commonly used clustering and classification algorithms. On the basis of advantages and disadvantages of existing algorithms, we introduce the KMRIC algorithm, which improves initial centers of K-means. Also, we propose the AKM algorithm which automatically determines the number of clusters and apply support vector machine to classification. Finally, we apply these algorithms to electrofacies analysis, where the experiments on the real-world datasets are carried out to compare the merits of various algorithms.

1. Introduction

The basic principle of electrofacies analysis is to determine the lithological types corresponding to electrofacies according to the known lithological and underlying parameters in the key well. Then we conduct clustering and discriminant analysis of key well and noncoring wells to automatically judge the automatica.

Clustering means the process of partitioning an unlabeled dataset into groups of similar objects. Each group, called a cluster, consists of objects that are similar to each other with respect to a certain similarity measure and which are dissimilar to objects of other groups. The applications of cluster analysis have been used in a wide range of different areas, including artificial intelligence, bioinformatics, biology, computer vision, data compression, image analysis, information retrieval, machine learning, marketing, medicine, pattern recognition, spatial database analysis, statistics, recommendation systems, and web mining.

Dong et al. [1] proposed an improvement method based on K-means, which obtains the optimized initial center from a group of initial clustering centers. The K-means algorithm is one of the most popular and widespread partitioning clustering algorithms because of its superior feasibility and efficiency in dealing with a large amount of data. The main drawback of the KM algorithm is that the cluster result is sensitive to the selection of the initial cluster centers and may converge to the local optima. At present, the development tendency of clustering method is to find a global optimal solution in combination with the global optimization methods like simulated annealing, particle swarm, and other local methods like K-means [2–6]. Pelleg and Moore [7] proposed an algorithm which can automatically determine the optimal number of clusters during clustering. The challenge of clustering high-dimensional data has emerged in recent years. Clustering high-dimensional data is the cluster analysis of data anywhere from a few dozens to many thousands of dimensions. Such high-dimensional data spaces are often encountered in areas such as medicine, biology, bioinformatics, and the clustering of text documents, where, if a word-frequency vector is used, the number of dimensions equals the size of the dictionary. In high-dimensional clustering, generally the original space is transformed by PCA, SVD, K-L transformation, and other dimensionality reduction methods first; then the clustering of low-dimensional space is performed. Bertini et al. [8] introduced a high-dimensional visualization technology, showing multidimensional data on two-dimensional plane.

K-means [9, 10] is a clustering method most widely used in science and engineering nowadays. However, it has the following 5 deficiencies [3, 5].

(1) The results are initial center initiative.

(2) Only local optimal solution can be obtained, rather than global optimal solution.

(3) The number of clustering k should be set in advanced artificially.

(4) The error point imposes serious impacts on the results of clustering.

(5) The algorithm lacks scalability.

The paper introduces an improved algorithm according to the deficiencies of K-means.

2. Improve K-Means Method of Initial Center

Aimed at the disadvantages (1) and (4) in K-means algorithm, we propose a K-means algorithm with refined initial centers (KMRIC for short) based on the works of predecessors [1].

(1) Randomly extract J sample subsets S_i, $i = 1, 2, \ldots, J$.

(2) Conduct K-means clustering of J sample subsets, respectively, on the whole data field to get J sets CM_i, $i = 1, 2, \ldots, J$, $\mathrm{CM} = \bigcup_{i=1}^{J} \mathrm{CM}_i$, in which there are $K \times J$ points for CM at most.

(3) Conduct K-means clustering on CM by taking CM as the initial clustering center to get J clustering center sets FM_i $i = 1, 2, \ldots, J$.

It can be seen from Figure 1 that the clustering center is obtained from different subsample set, near the real clustering center, and clustering is formed by different subsample set. In (3), selecting the one with the minimum sum of squares of deviations as the improved initial clustering center can reduce the randomness brought by random selection. In (2), to eliminate the influence of error point, the modified K-means algorithm (KmeansMod) is adopted. KmeansMod has the following modification based on the standard K-means: when the standard K-means algorithm is completed, the data point contained in each clustering will be checked. If the data point contained in a clustering is zero, the original center will be replaced by taking the data point furthest to the clustering center as a new center and then the K-means algorithm is reran.

KMRPIC algorithm eliminates the sensitivity of K-means algorithm to data input consequence and initial centers, which is an obvious improvement compared with K-means effect. When applied to large-scale data, KMRIC can reduce the iterations and improve the execution efficiency.

3. Adaptive K-Means

The number of clusters k of K-means algorithm should be set in advance manually. However, actually we do not know the value of k, especially in the case of high dimension of data, so it is more difficult to select the correct value of k.

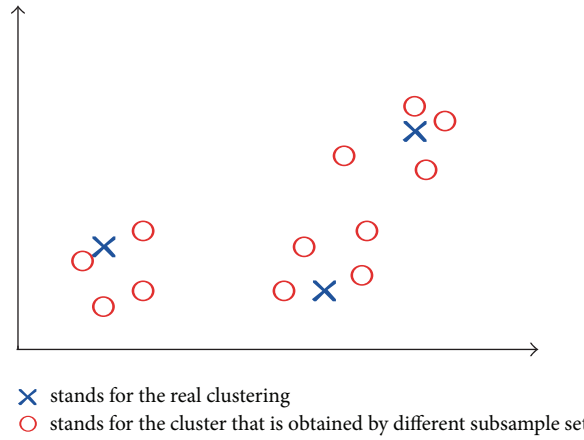


Figure 1: Multicombination clustering center obtained from multiple sample subsets.

X-means put forward by Pelleg and Moore [7] can automatically determine the number of clusters. However, X-means is prone to split data into more clusters than the actual ones, which is particularly obvious when the data is not strictly subject to the normal distribution. Lewis [11] statistics are introduced as the standard of measuring the normal distribution and propose an adaptive K-means (AKM for short).

The AKM algorithm first assumes that all data are in the same cluster; then the number of clustering is gradually increased in the subsequent iterations. In each iteration, whether each cluster satisfied the normal distribution is judged at once; if not, the cluster should be split into two clusters. After each splitting, K-means clustering is carried out in the whole data field to improve the clustering results. The iteration ends until there is no splitting and then the final clustering results will be obtained. The schematic diagram of AKM algorithm is shown in Figure 2. In Figure 2, clustering is divided into three categories firstly; then each category is split into two subclasses. At last, the results are got after one splitting to judge whether each subclass follows Gaussian distribution.

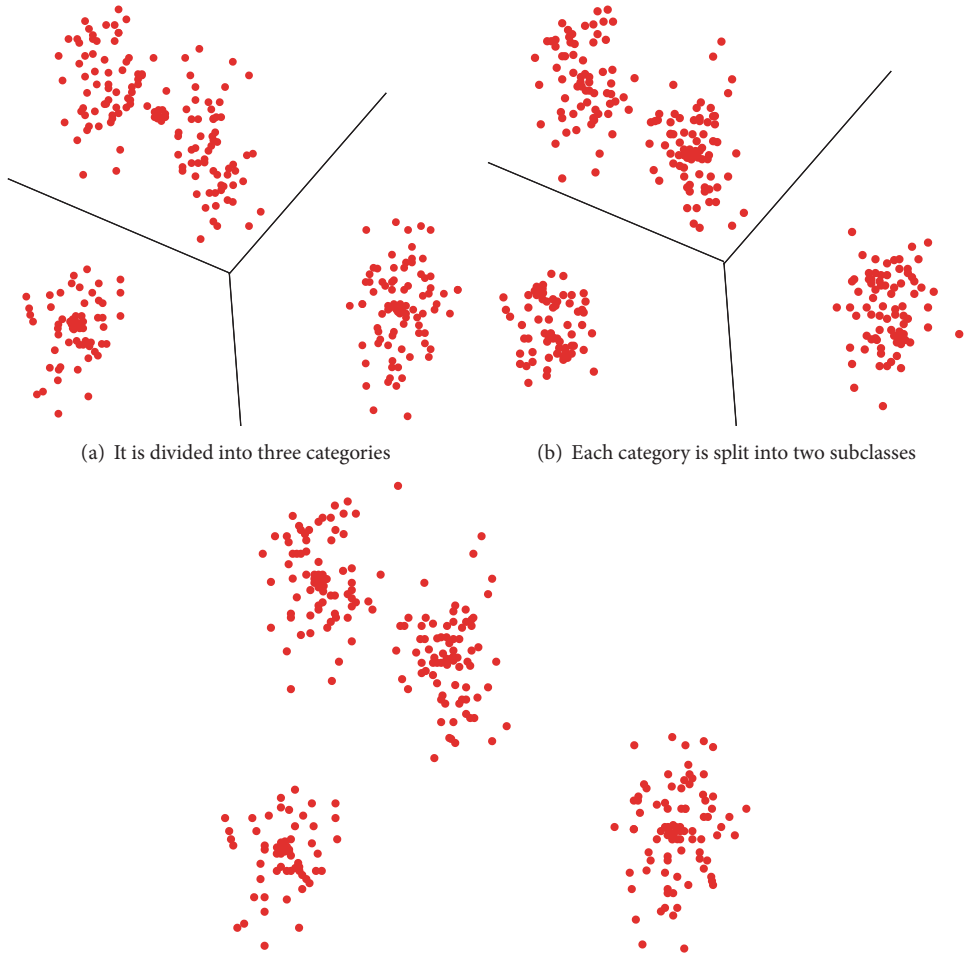
The judgment of splitting is as follows.

(1) Select the confidence level α.

(2) Run KMRIC program and split X into two to get two clustering centers c_1, c_2.

(3) Let $v = c_1 - c_2$ be an N-dimensional vector connecting the two centers, which is the main direction of judging the normal distribution. X is projected on v: $x_i' = (\langle x_i, v\rangle / \|v\|^2)X'$ is transformed to make its mean as 0 and variance as 1.

(4) Suppose that $z_i = F(x'_{(i)})$. The results $A_*^2(Z)$ with respect to confidence level α are not significant, so accept H_0; reserve the original clustering center c and abandon c_1 and c_2. Otherwise, reject H_0, and replace the original clustering center c by c_1 and c_2.



(a) It is divided into three categories

(b) Each category is split into two subclasses

(c) Get the results after one splitting to judge whether each subclass follows Gaussian distribution

FIGURE 2: Schematic diagram of AKM algorithm.

$A^2_*(Z)$ is the statistics of Anderson Darling:

$$A^2(Z) = -\frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\log(z_i) + \log(1 - z_{n+1-i})\right] - n. \tag{1}$$

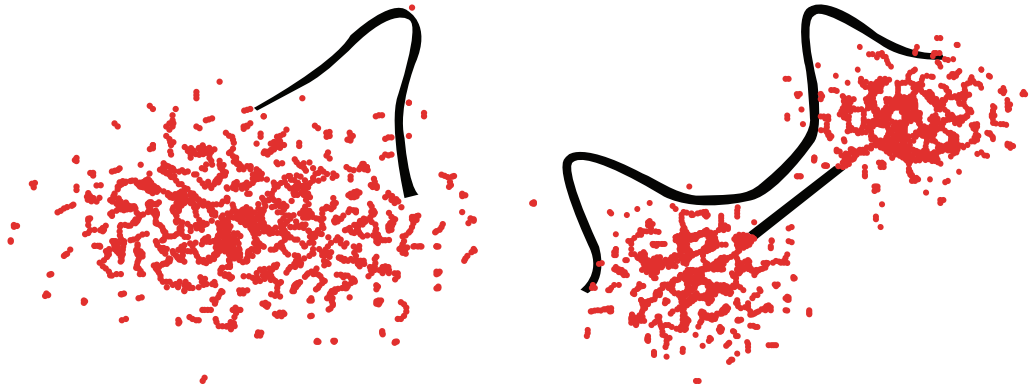
Figure 3 shows two distribution circumstances. In Figure 3(a), the subclass follows Gaussian distribution, but in Figure 3(b), the subclass does not follow Gaussian distribution. AKM algorithm can judge whether each subclass follows Gaussian distribution.

AKM integrates the determination process of the number of clusters and the clustering process, which can automatically determine the optimal number of clusters, thus avoiding the subjectivity in the selection of number of clusters and the blindness of initialization, and can also distinguish the errors.

4. Discriminant Method

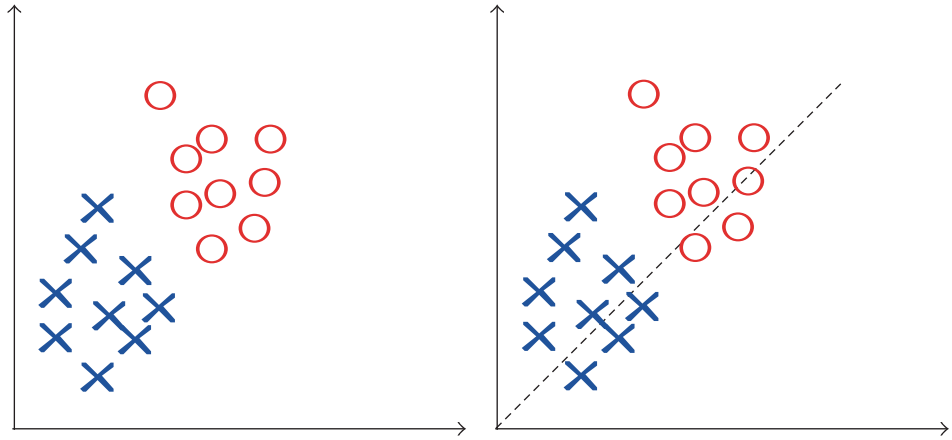
4.1. Fisher Classification. Fisher method actually is about the dimension compression. Projecting the samples which can be easily separated in higher space on a straight line arbitrarily may be difficult to be identified for different types mixed together. Generally, the best direction can always be found to separate the samples when projected on that direction. But how to find out the best direction and how to realize the transformations of projection toward the best direction are the very two problems to be solved by Fisher algorithm. Figure 4 shows analysis schematic diagram of Fisher algorithm using linear discriminant. In Figure 4(a), the sample cannot be identified when being projected on coordinate axis, and in Figure 4(b), the projection samples can be identified by looking for a direction.

4.2. Potential Function Classification. Potential function, a common method used in nonlinear classifier, is a way to solve the classification problems of pattern via the conception of electric field. In the potential function classification, the samples belonging to one category are treated as positive charge while the samples belonging to another category are treated as the negative charge, thus turning the classification problems of pattern to the matter of transferring the positive



(a) The subclass follows Gaussian distribution (b) The subclass does not follow Gaussian distribution

Figure 3: Judge whether each subclass follows Gaussian distribution.



(a) The sample cannot be identified when being projected on coordinate axis (b) The projection samples can be identified by looking for a direction

Figure 4: Schematic diagram of Fisher linear discriminant analysis.

charge and negative charge, and the equipotential line where its electric potential is zero is the decision boundary. The training course of potential function algorithm is a process of accumulating electric potential when the samples are input one after another by exploiting the potential function.

4.3. Least Squares Support Vector Machine (LS-SVM). Based on the VC dimension theory of statistical learning theory and the structural risk minimization principle, support vector machines method [12] converts the practical problem to high-dimensional feature space through nonlinear transformation and realizes the nonlinear discriminant function in the original space by constructing linear discriminant function in higher space. By means of introducing the least squares linear system into support vector machine to replace the traditional one, quadratic programming method, which is adopted to settle the problems of classification and estimation, is a kind of extension of traditional support vector machine.

5. Procedures of Electrofacies Analysis

The procedure of electrofacies analysis is shown in Figure 5.

5.1. Feature Extraction of Log Data. The primary step to establish electrofacies is to extract a set of log data features that can reflect the lithologic character of sedimentary rock. Generally, there are 9 types of well-logging items or more and those logging items are interrelated. There are two ways to eliminate gibberish, simplify control methods, and reduce calculated amount: (1) principal component analysis. (2) Select logging items manually. The extracted logging items will be recorded in Table stdlogdata as the data source for clustering analysis.

5.2. Clustering Analysis. In order to find out the electrofacies of the same type and establish a standard library in electrofacies analysis, clustering analysis must be conducted to stratum. Finally, the classification results acquired by clustering

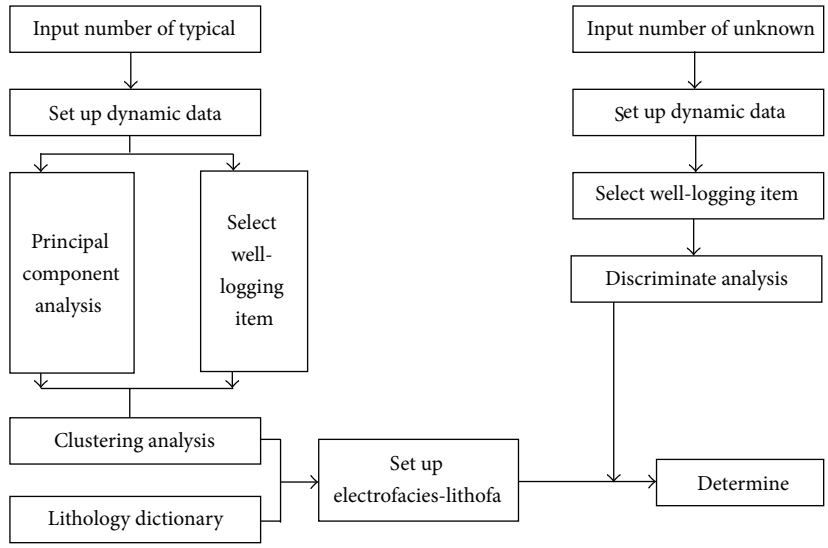


Figure 5: Flow diagram of electrofacies analysis.

should be recorded in the column of "Category" in Table stdlogdata and the lithology be recorded in the column of "Lithology" according to the lithology dictionary.

5.3. Discriminant Analysis. After establishing lithofacies database, namely, the electrofacies of type well, it is possible to discriminate the lithofacies of other wells. After discrimination, the data and discriminant results will be written in Table anylogdata and the logging items be written in the Table anylogitem.

6. Comparison and Analysis of Results of Algorithm

6.1. Experimental Data. The Iris dataset [13] usually serves as the testing dataset for benchmark function, in which each record contains 4 attributes of Iris, totaling 150 samples. The correct classification result is that each type of data has 50 samples. Eight attributes are included in each set of data of electrofacies, totaling 177 samples. As for the real data in electrofacies, there is no strictly accurate number of categories and standard classification. Judging by experience, 8 classifications may be rational.

Table 1: Clustering method comparison under Iris dataset.

	K-means	ISODATA	KMRIC	AKM	Matlab
Type I	30	50	50	53	50
Type II	24	39	39	97	38
Type III	96	61	61	0	62
Accuracy	69.3%	92.6%	92.6%	66.7%	92%

Table 2: Clustering method comparisons under Iris dataset.

	K-means	ISODATA	KMRIC	AKM	Matlab
Type I	56	35	47	46	47
Type II	38	30	46	40	45
Type III	36	26	26	26	23
Type IV	18	23	14	23	23
Type V	11	17	13	13	13
Type VI	9	13	12	12	12
Type VII	8	12	10	10	11
Type VIII	1	10	9	7	3
Type IX	0	8	0	0	0
Type X	0	3	0	0	0

6.2. Analysis of Experimental Results of Cluster

6.2.1. Iris Dataset. It can be easily seen from Figures 6–9 that the cluster obtained by standard K-means algorithm is pretty different from the standard results, while the clustering results obtained by ISODATA and KMRIC come near to the standard ones and are the same as the results obtained by built-in K-means algorithm of Matlab. AKM has only two categories. The second and the third categories are deemed as belonging to the same normal distribution that are never apart for they are approximate to each other and have some parts overlapped (see Table 1 and 2).

6.2.2. Electrofacies Dataset. It can be seen from Figures 10–13 that the clustering results obtained by K-means have large error, while the cluster obtained by KMRIC and AKM is relatively rational and can basically reflect the right classification, and AKM can also identify the accurate number of clustering automatically. Compared with ISODATA, AKM is more accurate in determining the number of clustering and its clustering results are more rational as well. Besides, it proves that the hypothesis testing way to judge the number of clustering of AKM is more universal than that by judging it based on the between-class distance of ISODATA.

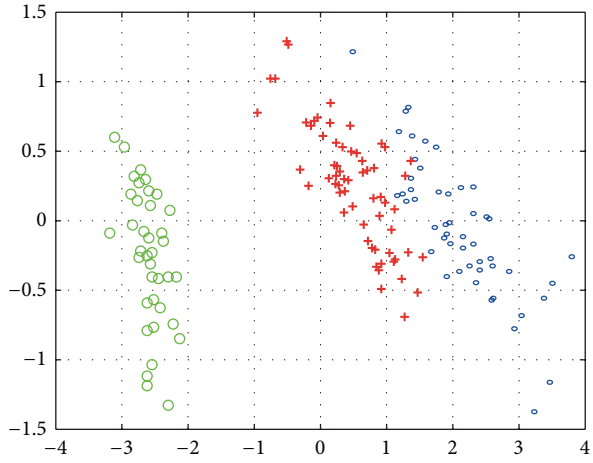


Figure 6: Clustering results of dataset by Matlab figure.

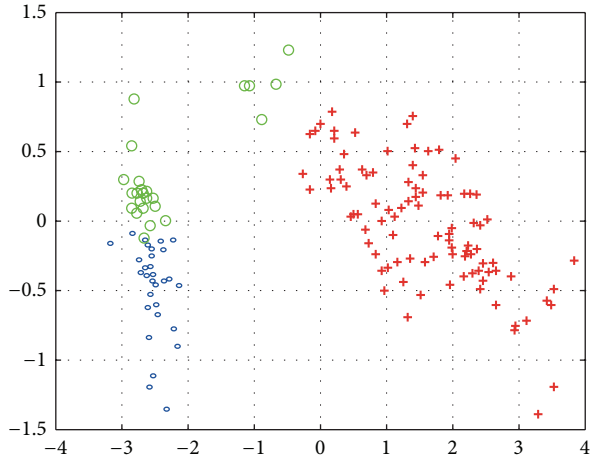


Figure 7: Clustering results obtained by standard K-means.

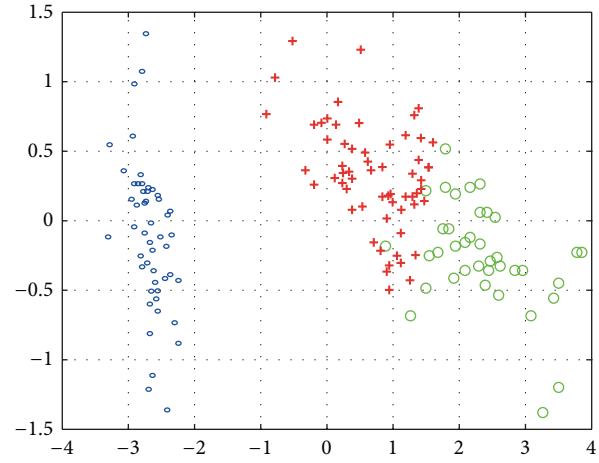


Figure 8: Clustering results obtained by ISODATA and KMRIC.

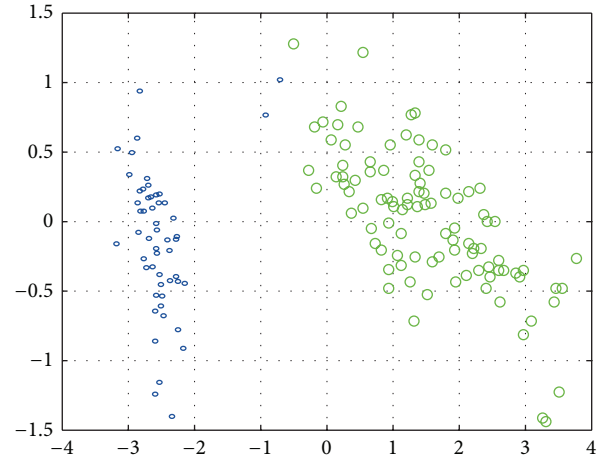


Figure 9: Clustering results obtained by AKM.

Table 3: Number of misclassification and accuracy of various discriminant methods under Iris dataset.

	Fisher	Potential function	LS-SVM
Type I	0	0	0
Type II	1	0	0
Type III	0	0	0
Total	1	0	0
Accuracy	96.7%	100%	100%

Table 4: Number of misclassification of various discriminant methods under electrofacies dataset.

	Fisher	Potential function	LS-SVM
Type I	—	0	0
Type II	—	0	2
Type III	—	0	0
Type IV	—	1	2
Type V	—	0	0
Type VI	—	0	0
Type VII	—	0	3
Type VIII	—	0	2
Total	—	1	9
Accuracy	—	94.9%	76.9%

6.3. Experimental Results and Analysis of Classification

6.3.1. Iris Dataset. See Table 3.

6.3.2. Electrofacies Dataset. It can be seen from Tables 3 and 4 that these three classification methods all work well when processing the Iris data for the data structure of Iris is quite simple and low in dimension. As for electrofacies data, Fisher discriminant analysis is not applicable due to the singular within-class scatter S_w matrix, while the potential function and LS-SVM still have better accuracy to classification. The multiclassification of LS-SVM application remains for further study.

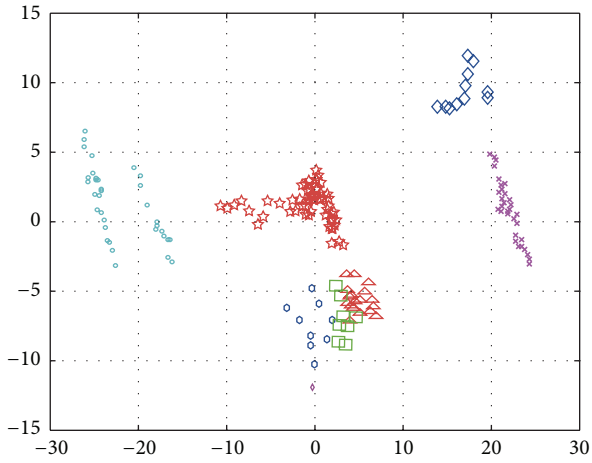


Figure 10: Clustering results obtained by standard K-means.

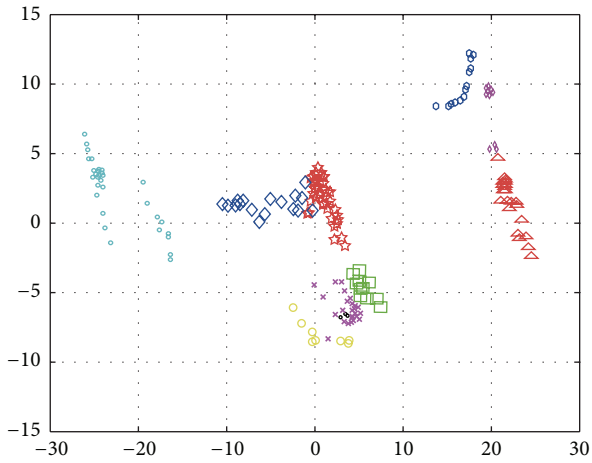


Figure 11: Clustering results obtained by ISODATA.

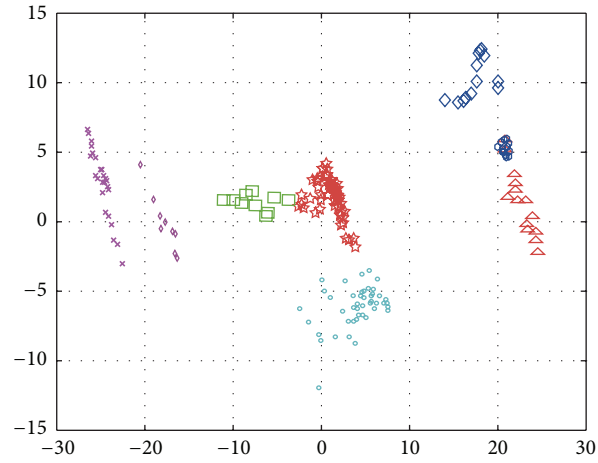


Figure 12: Clustering results obtained by KMRIC.

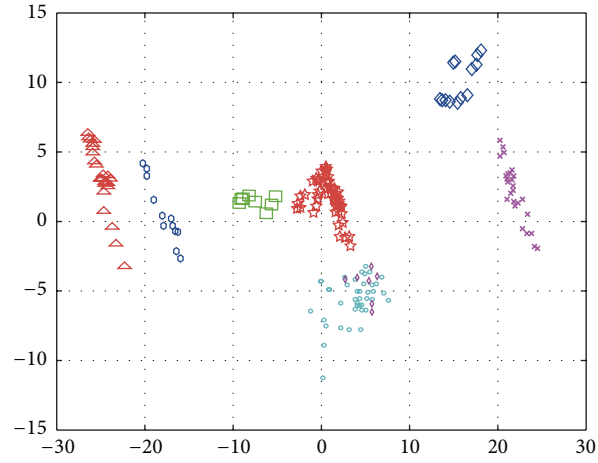


Figure 13: Clustering results obtained by AKM.

7. Conclusion

On the basis of analyzing the strengths and weaknesses of the existing main algorithms for clustering, this paper proposed the KMRIC algorithm for improving initial points and the AKM algorithm for determining the number of clusters. The support vector machine has also been used for classification. Finally, the algorithms are applied to electrofacies analysis. Through the experimental analysis, comparison was made among algorithms. According to the experimental results, the KMRIC algorithm erases the sensibility of K-means algorithm to data input sequence and initial centers, and it achieves an obvious improvement relative to K-means and ISODATA; AKM algorithm mixes the process of determining the number of clusters and the clustering process together to avoid the subjectivity in selecting the number of clusters and the blindness in initial divisions. Under general condition, the number of clusters and rational clusters can be found correctly.

There are some other problems that remain open. The volatility of results, which was caused by the randomness of selecting initial points in KMRIC, existed in KMRIC and AKM. To address this problem, we can lower the randomness by selecting the optimal initial points repeatedly. Hierarchical clustering is a very stable method but its disadvantage is the massive calculation cost. How to combine the hierarchical clustering and the abovementioned methods may be taken as the improvement direction in future.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

Research of the authors was supported by the National Nature Science Foundation of China (no. 61103037), Nature Science Foundation of Guangdong Province (no. S2013010011858), Project of Guangdong University of Outstanding Young Talents Cultivation (no. 2012LYM_0125), and Dongguan Science and Technology Project (no. 2012108102007).

References

[1] S. Dong, D. D. Zhou, and W. Ding, "Flow cluster algorithm based on improved K-means method," *IETE Journal of Research*, vol. 59, no. 4, pp. 326–333, 2013.

[2] J. Q. He, H. Dai, and X. Song, "The combination stretching function technique with simulated annealing algorithm for global optimization," *Optimization Methods and Software*, vol. 29, no. 3, pp. 629–645, 2014.

[3] J. Liu and T. Z. Liu, "Detecting community structure in complex networks using simulated annealing with k-means algorithms," *Physica A: Statistical Mechanics and Its Applications*, vol. 389, no. 11, pp. 2300–2309, 2010.

[4] S. H. Kim and L. Li, "Statistical identifiability and convergence evaluation for nonlinear pharmacokinetic models with particle swarm optimization," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 2, pp. 413–432, 2014.

[5] S. Kalyani and K. S. Swarup, "Particle swarm optimization based K-means clustering approach for security assessment in power systems," *Expert Systems with Applications*, vol. 38, no. 9, pp. 10839–10846, 2011.

[6] D. H. Wang, J. F. Wang, and X. Y. Xu, "A relevance vector machine and bare-bones particle swarm optimization hybrid algorithm for PD pattern recognition of XLPE cable," *Journal of Computational Information Systems*, vol. 8, no. 2, pp. 451–458, 2012.

[7] D. Pelleg and A. W. Moore, "X-means: extending K-means with efficient estimation of the number of clusters," in *Proceedings of the 17th International Conference on Machine Learning*, pp. 727–734, 2000.

[8] E. Bertini, A. Tatu, and D. Keim, "Quality metrics in high-dimensional data visualization: an overview and systematization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2203–2212, 2011.

[9] L. M. Li and Z. S. Wang, "Method of redundant features eliminating based on k-means clustering," *Applied Mechanics and Materials*, vol. 488, pp. 1023–1026, 2014.

[10] C. H. Lin, C. C. Chen, H. L. Lee et al., "Fast K-means algorithm based on a level histogram for image retrieval," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3276–3283, 2014.

[11] P. A. W. Lewis, "Distribution of the Anderson-Darling statistic," *Annals of Mathematical Statistics*, vol. 32, pp. 1118–1124, 1961.

[12] M. Z. Tang and C. H. Yang, "Excellent operational pattern recognition based on simultaneously optimizing cost-sensitive support vector machine," *CIESC Journal*, vol. 64, no. 12, pp. 4509–4514, 2013.

[13] C. L. Blake and C. J. Merz, "UCI repository of machine learning databases," 1998.