Bio-Medical Materials and Engineering 26 (2015) S1863–S1869
DOI 10.3233/BME-151489
IOS Press

An effective fuzzy kernel clustering analysis approach for gene expression data

Lin Sun[a,b,*], Jiucheng Xu[a,b] and Jiaojiao Yin[a]

[a] *College of Computer and Information Engineering, Henan Normal University, Xinxiang, China*
[b] *Engineering Technology Research Center for Computing Intelligence and Data Mining, Henan Province, China*

Abstract. Fuzzy clustering is an important tool for analyzing microarray data. A major problem in applying fuzzy clustering method to microarray gene expression data is the choice of parameters with cluster number and centers. This paper proposes a new approach to fuzzy kernel clustering analysis (FKCA) that identifies desired cluster number and obtains more steady results for gene expression data. First of all, to optimize characteristic differences and estimate optimal cluster number, Gaussian kernel function is introduced to improve spectrum analysis method (SAM). By combining subtractive clustering with max-min distance mean, maximum distance method (MDM) is proposed to determine cluster centers. Then, the corresponding steps of improved SAM (ISAM) and MDM are given respectively, whose superiority and stability are illustrated through performing experimental comparisons on gene expression data. Finally, by introducing ISAM and MDM into FKCA, an effective improved FKCA algorithm is proposed. Experimental results from public gene expression data and UCI database show that the proposed algorithms are feasible for cluster analysis, and the clustering accuracy is higher than the other related clustering algorithms.

Keywords: Spectral analysis, maximum distance, fuzzy clustering, gene expression data

1. Introduction

In computational biology, clustering is a useful technique for gene expression data as it groups similar objects together and allows biologist to identify potential relationships between genes [1]. Unsupervised clustering methods have been applied to gene expression data analysis, and the unsupervised ensemble approaches improve accuracy and reliability of clustering results [2]. However, traditional clustering approaches are inadequately flexible when a gene experiences differential coregulation in different samples of the same data set as a result of being involved in differing functional relationships [3].

In recent years, the application of kernels in fuzzy c-means (FCM), fuzzy k-means, and evolution algorithms is effective in terms of improving clustering performance. However, FCM has drawbacks such as the result of clustering process deteriorates while noise and outliers exist in data set, blindness of random prototype initialization leads clustering process as a time consuming task and it works well only on spherical shaped data set not in general shaped data set [4]. To satisfy more general data set,

[*] Address for correspondence: Lin Sun, College of Computer and Information Engineering, Henan Normal University, Xinxiang, China. Tel.: 03733329075; E-mail: linsunok@gmail.com.

classical fuzzy clustering-based kernel methods are adapted to adopt a kernel-induced metric in data space to replace original Euclidean norm metric in FCM [5]. Shen et al. [6] proposed a weighted fuzzy kernel-clustering algorithm in kernel feature space. Liu and Zhang [7] proposed a new kernel function and dynamic weighted kernel FCM clustering method for gene expression data analysis, however, the determination initial cluster integer need to be further studied.

Assumption of cluster number and clustering algorithm employed can improve effectiveness and stability of clustering analysis. Cluster number and distance-based similarity measures for cluster centers are two important assumptions for clustering analysis approaches. In most of automatic clustering algorithms, the cluster number must be first defined, and this is true for most popular algorithms like FCM clustering algorithm [8]. Another important issue in FCM is choosing initial cluster centers, which generally has been done randomly, then a prototype initialization method is designed to assign initial cluster centers without human intervention to reduce the number iterations and use silhouette method to obtain clustering validity and cluster number [9]. Cluster validity indices are proposed to validate clustering results so as to obtain optimal cluster number [10]. A fuzzy point symmetry-based genetic clustering technique is proposed to determine the number of clusters present in a data set [11]. Further research has been done on automatic center initialization methods to reduce the computational complexity of FCM by improper center of actual classes of data set [9]. A method automatically determining number of clusters and locations of cluster centers is proposed, however, it does not provide good generalization capabilities in obtaining appropriate centers [12]. Nowadays, further studies about cluster number and centers of fuzzy clustering are still active. This paper focuses on creating such a solution.

In fact, many of the selection and classification methods can be combined, and combination of the methods may give us better results. The objective of this paper is to propose an adaptive fuzzy clustering analysis method that produces reliable clustering results for gene expression data sets. However, many existing clustering analysis methods randomly select or artificially determine the clustering number and centers, and this leads to decrease stability and accuracy of clustering results. Until now, few studies have all addressed the above issues to demonstrate such a fuzzy kernel clustering method for gene expression data. Then, to overcome the above defects, ISAM and MDM are proposed to efficiently determine cluster number and centers respectively, which are introduced into FKCA, so the cluster number and centers do not need to be initialized. Thus, a feasible improved FKCA algorithm is proposed. Experiments show that the proposed methods perform well in improving stability and accuracy of clustering results.

2. Materials and methods

2.1. Improved spectrum analysis method

When characteristic differences among data are not obvious, the eigenvalues calculated by SAM [13] are very close, and the results have larger deviation. Hence, it is difficult to identify number of remarkable eigenvalues and get more accurate optimal cluster number. Gaussian kernel function $K(x,y) = \exp(-\beta||x-y||^2)$ can map an infinite dimensional feature space, where x, y are given samples and $\beta > 0$ is a self-defined parameter, and samples with limited number must be linearly separable. Then, by combining Gaussian kernel function with SAM, ISAM is proposed to estimate the optimal cluster number. The specific steps of ISAM algorithm can be expressed as follows:

Algorithm 1. ISAM algorithm

Input: Gene expression data set $\Gamma = \{x_1, x_2, \cdots, x_N\} \subset R^L$.

Output: Optimal cluster number C.
Step 1: Standardize $\Gamma = \{x_1, x_2, \cdots, x_N\} \subset R^L$ with $x'_{ij} = (x_{ij} - \bar{x_i})/\sqrt{\frac{1}{L-1}\sum_{j=1}^{L}(x_{ij} - \bar{x_i})^2}$, where $\bar{x_i} = \frac{1}{L}\sum_{j=1}^{L} x_{ij}$ is the average of x_i. Then, the standardized gene expression data set can be written as $\Gamma' = \{x'_1, x'_2, \cdots, x'_N\} \subset R^L$, where $x'_i = x'_{i1}, x'_{i2}, \cdots, x'_{iL}$.
Step-2: Initialize the related parameters of Gaussian kernel function.
Step 3: Map each sample $x'_i \in R^L$ of Γ' to high-dimensional feature space H through Gaussian kernel function, where $i = 1, 2, \cdots, N$, and then get a nonlinear mapping $\Phi(x'_1), \Phi(x'_2), \cdots, \Phi(x'_l)$. The inner product of samples in H can be shown by $K(x'_i, x'_j) = (\Phi(x'_i)\Phi(x'_j))$. All samples can compose kernel function matrix $K_{i,j} = K(x'_i, x'_j)$ in mapped feature space.
Step 4: Calculate a similarity matrix $D = [d_{ij}]_{N\times N}$ of Γ through Euclidean distance used as similarity measures, where d_{ij} is a correlation between two samples in Γ', denoted by $d(x'_i, x'_j) = \sqrt{K(x'_i, x'_i) - 2K(x'_i, x'_j) + K(x'_j, x'_j)}$.
Step 5: Calculate a regular matrix $R = U^{-1}D$, where U is a diagonal matrix $u_{ij} = \delta_{ij}\sum_{k=1}^{N} d_{ik}$, and D is a similarity matrix. Here, if $i = j$, $\delta_{ij} = 1$, otherwise $\delta_{ij} = 0$.
Step-6: Calculate the eigenvalues of R, draw the module of eigenvalues in two-dimensional figure, and then get C of Γ which is equal to that of remarkable eigenvalues.

Through using Gaussian kernel function to map Γ' to H and Euclidean distance to calculate similarity matrix $D = [d_{ij}]_{N\times N}$ in H, ISAM not only optimizes the features of Γ' and highlights the feature differences, but also identifies the differences and estimates the more precise cluster number.

2.2. *Maximum distance method*

In traditional kernel clustering algorithms, the cluster centers are usually randomly selected or artificially, and this leads to volatility of clustering results. Then, it is necessary that the cluster centers can be determined to eliminate volatility of clustering analysis and get more stable clustering results. In order to solve the problem, by combining subtractive clustering [14] with max-min distance mean [15], MDM is proposed to determine cluster centers. The specific steps of MDM can be expressed as follows:

Algorithm 2. MDM algorithm
Input: Gene expression data set $\Gamma = \{x_1, x_2, \cdots, x_N\} \subset R^L$, and cluster number C.
Output: Cluster centers $w_1, w_2, \cdots, w_C$.
Step 1: Obtain the distance matrix $D = [d_{ij}]_{N\times N}$ of Γ' with Algorithm 1.
Step 2: Calculate an average distance $\overline{TD} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}$ of all samples, and get the sum of distance $D_i = \sum_{j=1}^{N} d_{ij}$ between sample x'_i and other samples, where $i = 1, 2, \cdots, n$.
Step 3: If $D_i > \overline{TD}$, remove x'_i as an isolated point from Γ, and then delete all of the sample points satisfying $D_i > \overline{TD}$ to get a high density area Γ'_1.
Step 4: Calculate a density index $\rho_i = \sum_{j=1}^{N} \exp(\frac{-\|x'_i - x_j^{prime}\|}{(0.5\gamma_a)^2})$ of each sample point in Γ'_1, where $\gamma_a > 0$ stipulates some field scope of x'_i. If it is within the scope of γ_a, there are multiple adjacent samples around x'_i. Thus, it has a high density index.
Step 5: Select the sample point with biggest ρ_i in Γ'_1 as the first cluster center w_1.
Step 6: Select the second cluster center w_2 in Γ'_1, which satisfies $d(w_1, w_2) = \max(d(x'_i, w_2))$, $i = 1, 2, \cdots, n$. Namely, the selected w_2 is a sample point which is furthest away from w_1 in Γ'_1.

Step 7: Select the third $w_3 \in \Gamma_1'$ satisfying $\min(d(w_3, w_1), d(w_3, w_2)) = \max(\min(d(x_i', w_1), d(x_i', w_2)))$, and then obtain $w_c \in \Gamma_1'$ satisfying $\min(d(w_C, w_1), d(w_C, w_2), \cdots d(w_C, w_{C-1})) = \max(\min(d(x_i', w_1), d(x_i', w_2), \cdots, d(x_i', x_{C-1})))$, where $i = 1, 2, \cdots, n$.

Through removing the isolated points and the noise points, and dividing up the high density area from gene expression data set, MDM not only selects the cluster centers in the high density area, but also reduces computational time and obtains the steady cluster centers.

2.3. Improved fuzzy kernel clustering analysis algorithm

On the basis of ISAM and MDM, an improved FKCA algorithm is designed. Firstly, the cluster number and centers can be obtained by ISAM and MDM, respectively. Then, the cluster number and centers are used as the initial assumptions of improved FKCA algorithm to perform clustering analysis for gene expression data set. The process of improved FKCA algorithm can be summarized as follows:

Algorithm 3. Improved FKCA algorithm

Input: Gene expression data set $\Gamma = \{x_1, x_2, \cdots, x_N\} \subset R^L$.

Output: Cluster number C, cluster centers $w_1, w_2, \cdots, w_C$, and C clusters.

Step 1: Estimate an optimal cluster number C with ISAM algorithm.

Step 2: Determine the initial cluster centers w_c, where $c = 1, 2, \cdots, C$ with MDM algorithm.

Step 3: Initialize a membership matrix μ_{ci} and an iteration index $t = 0$.

Step 4: Calculate an objective function $J_H = \sum_{c=1}^{C} \sum_{i=1}^{N} \mu_{ci}^q (\Phi(x_i') - \Phi(w_c))(\Phi(x_i') - \Phi(w_c))^T$, where $\Phi(x_i')$ is a form of sample i in H, $\Phi(w_c)$ is a cluster center of c in H, C is a cluster number, $q \in [1, +\infty)$ is a weighted index, and $\mu_{ci} \in [0, 1]$ is a membership matrix satisfying $\sum_{c=1}^{C} \mu_{ci} = 1$.

Step 5: While $|J_H(t) - J_H(t-1)| > \varepsilon$ do

Step 6: Set $t = t + 1$.

Step 7: Convert the above objective function to $J_H = \sum_{c=1}^{C} \sum_{i=1}^{N} \mu_{ci}^q Q_{ci}$ because J_H is composed of gene vectors in H, where Q_{ci} is Euclidean distance between sample i and class c in H. Due to $\mu_{ci}' = \frac{1}{\sum_{m=1}^{C} (Q_{ci}/Q_{mi})^{\frac{1}{m-1}}}$, then $Q_{ci} = K_{ii} - \frac{2}{S} \sum_{m=1}^{N} \mu_{cm} K_{im} + \frac{1}{S_c^2} \sum_{m=1}^{N} \sum_{l=1}^{N} \mu_{cm} \mu_{cl} K_{cl}$, where $S_c = \sum_{i=1}^{N} \mu_{ci}$ and $K_{ij} = K(x_i', x_j')$, and get Q_{ci} between each sample point and the cluster centers.

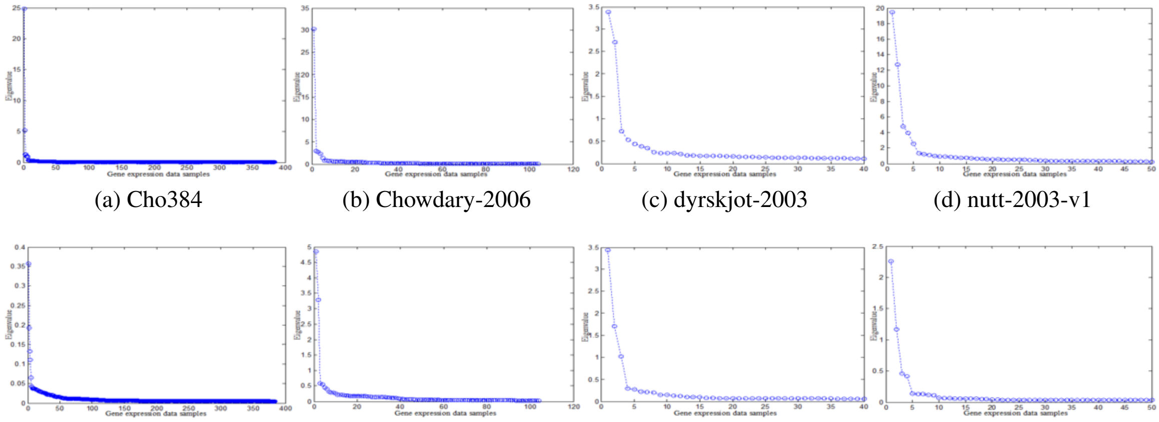
Step 8: Recalculate the membership function μ_{ci}' of every sample points.

Step 9: End while

Using ISAM and MDM to obtain cluster number and centers, improved FKCA can eliminate sensitivity of traditional clustering analysis from initial cluster centers, and get more steady clustering results.

3. Experimental results

For the experimental purposes, specific software has been developed to implement the proposed methods using MATLAB. The performances of our proposed algorithms are demonstrated and the experiments are divided into four parts. The first part is to verify the performance of ISAM on four public gene expression data sets, including Yeast cell cycle and three cancer data sets, which can be downloaded at http://faculty.washington.edu and http://bioinformatics.rutgers.edu, respectively. SAM and ISAM are used respectively to estimate optimal cluster number on the gene data sets. The experimental results are summarized in Figure 1. It can be seen from Figures 1(a)-1(d) that the numbers of remarkable eigenvalues on four gene data sets calculated with SAM are 2,1,3,5 respectively, and then their optimal cluster numbers are 2,1,3,5, however, the results of Cho384, Chowdary-2006, and nutt-2003-v1 are not conform



(a) Cho384 (b) Chowdary-2006 (c) dyrskjot-2003 (d) nutt-2003-v1

(e) Cho384 (f) Chowdary-2006 (g) dyrskjot-2003 (h) nutt-2003-v1

Fig. 1. (a)-(d) Cluster number of gene data set with SAM , (e)-(h) Cluster number of gene data set with ISAM.

to their own natures. The three corresponding estimated results are wrong, while the optimal cluster number of dyrskjot-2003 is 3, which is right. But the third remarkable eigenvalue is not obvious, and may be ignored. From Figures 1(e)-1(h), the numbers of remarkable eigenvalues with ISAM on four gene data sets are 5,2,3,4 respectively, and then the optimal cluster numbers are 5,2,3,4, which all conform to their own nature of all data sets. Thus, the performance of ISAM compared with SAM is greatly improved, and ISAM can distinguish, amplify and extract useful features of gene data sets through nonlinear mapping of Gaussian kernel function. Hence, the results of ISAM are more accurate than those of SAM.

The following part of our experiments is to investigate feasibility and validity of MDM. For the above four gene data sets in Figure 1, the performance of MDM is compared with that of random method (RM) [15]. The experimental results are shown in Table 1. Here, one performs 3 experiments as example, and the results of MDM are found the same each time, but the ones of RM are never the same. It follows that the cluster centers selected by MDM are more stable than those of RM for clustering analysis.

The third part of our experiments is to test the performances of our proposed FKCA algorithm (Algorithm 3). For the above four gene data sets, Algorithm 3 is compared with the other state-of-the-art clustering analysis methods to estimate clustering results, which are FKCA algorithm [16], FCM algorithm [17], and k-means algorithm [18]. The experimental results are illustrated in Table 2, where the cluster number of each class and the wrong cluster samples are simplified as n and s, respectively.

Table 1

Comparison of cluster centers for RM and MDM on gene expression data sets

Data sets	Cluster number	RM					MDM				
Cho384	5	119	95	170	329	50	12	150	275	286	357
		221	64	60	384	358	12	150	275	286	357
		10	136	282	199	260	12	150	275	286	357
Chowdary-2006	2	60	22				37	97			
		87	11				37	97			
		99	9				37	97			
dyrskjot-2003	3	6	35	2			16	26	32		
		9	7	31			16	26	32		
		34	12	21			16	26	32		
nutt-2003-v1	4	17	23	27	37		14	18	28	41	
		5	40	11	37		14	18	28	41	
		35	36	16	48		14	18	28	41	

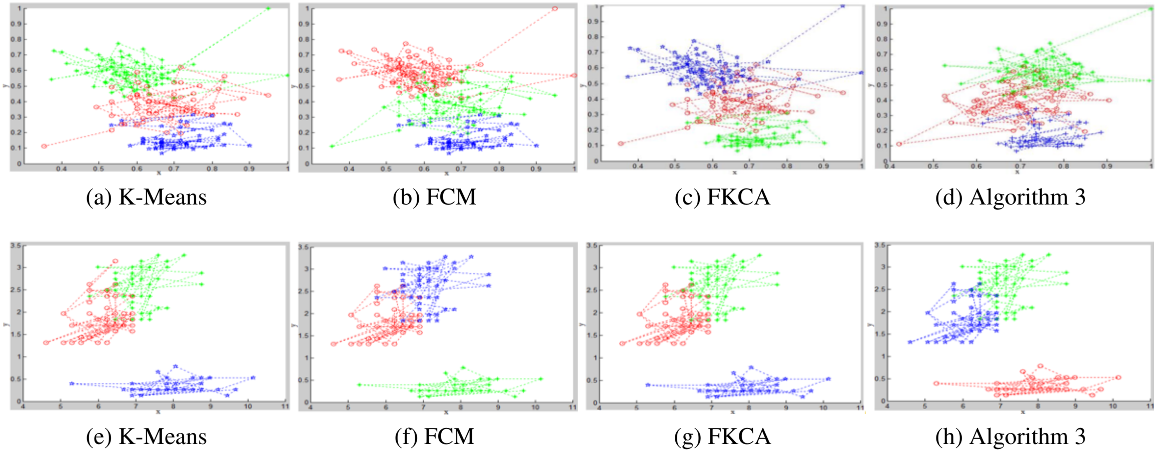
From Table 2, when the cluster numbers of four gene data sets are initialized to 5,2,3,4 classes respectively and the cluster centers are selected randomly or artificially, then FKCA algorithm [16] gives 14,4,5,10 wrong cluster sample points respectively, FCM algorithm [17] gives 56,5,5,7 wrong ones respectively, and k-means algorithm [18] gives 41,0,3,7 wrong ones respectively. However, when the cluster number and centers are not initialized, Algorithm 3 divides the four gene data sets into 5,2,3,4 classes, and gives 8,1,3,4 wrong ones respectively. Thus, the numbers of wrong cluster sample points from Algorithm 3 are significantly less than those of the other three algorithms. Hence, Algorithm 3 need not the initialized cluster number and centers, eliminates the effects of initialized cluster number and centers artificially, and can obtain the more stable and accurate clustering results for gene expression data sets.

The last part of our experiments is to compare our proposed Algorithm 3 with the above three algorithms (FKCA [16], FCM [17], and k-means [18]) on Iris and Wine from UCI databases. The experimental results are shown in Figure 2, from which when the cluster number and centers of Iris and Wine are not initialized artificially, Algorithm 3 can divide them into three classes. Then, using the membership matrixes of cluster centers in Algorithm 3, the sample numbers included in three classes of Iris and Wine are 52,48,50 and 70,58,50 respectively, and the wrong cluster sample points are all 2. However, the wrong cluster sample points of other three algorithms on Iris and Wine are 5,4,6 and 7,2,3 respectively. It is clearly shown that the accuracy of Algorithm 3 is better than the other three algorithms. Therefore, Algorithm 3 can also be used in clustering analysis of other types of data sets.

Table 2

Comparison results of four clustering analysis methods on gene expression data sets

Methods	Cho384						Chowdary-2006			dyrskjot-2003				nutt-2003-v1				
	n					s	n		s	n			s	n				s
K-Means	22	48	63	94	157	41	62	42	0	7	14	19	3	11	4	13	22	7
FCM	22	38	58	119	147	56	67	37	5	10	5	25	5	20	16	10	4	7
FKCA	50	58	72	81	123	14	66	38	4	25	6	9	5	8	3	15	24	10
Algorithm 3	52	56	71	78	127	8	63	41	1	23	7	10	3	9	10	14	17	4



(a) K-Means (b) FCM (c) FKCA (d) Algorithm 3

(e) K-Means (f) FCM (g) FKCA (h) Algorithm 3

Fig. 2. (a)-(d) Clustering analysis results of four methods on Wine, (e)-(h) Clustering analysis results of four methods on Iris.

4. Conclusion

In this paper, our major contribution is that ISAM and MDM are introduced into FKCA to determine cluster number and centers for tumor clustering, and then an effective improved FKCA algorithm is developed to cope with gene expression data. The corresponding steps of ISAM, MDM and improved FKCA are given in detail respectively, whose superiority and stability are demonstrated through performing some experimental comparisons on gene expression data sets and UCI database. Experiments show that the proposed hybrid methods compared with the other related clustering algorithms are efficient and feasible in improving the stability and accuracy of clustering analysis results for public data sets.

Acknowledgements

This work is supported by National Natural Science Foundation of China (61370169, 61402153), Project of Henan Sci. Technol. Dep. (142102210056), and Project of Henan Educ. Dep. (13A520529).

References

[1] S. Ghosha, S. Mitraa and R. Dattagupta, Fuzzy clustering with biological knowledge for gene selection, Applied Soft Computing **16** (2014), 102–111.

[2] A. Roberto and V. Giorgio, Fuzzy ensemble clustering based on random projections for DNA microarray data analysis, Artificial Intelligence in Medicine **45** (2009), 173–183.

[3] L. Tari, C. Baral and S. Kim, Fuzzy c-means clustering with prior biological knowledge, Journal of Biomedical Informatics **42** (2009), 74–81.

[4] S. Ramathilagam, R. Devi and S.R. Kannan, Extended fuzzy c-means: an analyzing data clustering problems, Cluster Computing **16** (2013), 389–406.

[5] H.Y. Zhang, Q.T. Wu and J.X. Pu, A novel fuzzy kernel clustering algorithm for outlier detection, IEEE International Conference on Mechatronics and Automation, Harbin, China, Aug. 5-8, 2007, pp. 2378–2382.

[6] H.B. Shen, J. Yang, S.T. Wang, et al., Attribute weighted mercer kernel based fuzzy clustering algorithm for general non-spherical datasets, Soft Computing **10** (2006), 1061–1073.

[7] W.Y. Liu and B. Zhang, Fuzzy clustering algorithm of kernel for gene expression data analysis, Proceeding of the International Symposium on Intelligent Information Systems and Applications, Qingdao, China, Oct. 28-30, 2009, pp. 553–556.

[8] P.Y. Mok, H.Q. Huang, Y.L. Kwok, et al., A robust adaptive clustering analysis method for automatic identification of clusters, Pattern Recognition **45** (2012), 3017–3033.

[9] S.R. Kannan, S. Ramathilagam and P.C. Chung, Effective fuzzy c-means clustering algorithms for data clustering problems, Expert Systems with Applications **39** (2012), 6292–6300.

[10] I. Berget, B. Mevik and T. Naes, New modifications and applications of fuzzy c-means methodology, Computational Statistics and Data Analysis **52** (2008), 2403–2418.

[11] S. Saha and S. Bandyopadhyay, A new point symmetry based fuzzy genetic clustering technique for automatic evolution of clusters, Information Sciences **179** (2009), 3230–3246.

[12] S.S. Khan and A. Ahmad, Cluster center initialization algorithm for K-means clustering, Pattern Recognition Letters **25** (2004), 1293–1302.

[13] A. Pothen, H.D. Simon and K.P. Liou, Partitioning sparse matrices with eigenvectors of graphs, SIAM Journal on Matrix Analysis and Applications **11** (1990), 430–452.

[14] C.J. Xiao and M. Zhang, Research on fuzzy clustering based on subtractive clustering and fuzzy c-means, Computer Engineering **31** (2005), 135–137.

[15] F. Yuan, Z.Y. Zhou and X. Song, K-means clustering algorithm with meliorated initial center, Computer Engineering **33** (2007), 65–66.

[16] P.P. Lin and S.Z. Ye, The tumor extraction algorithm of liver MRI using fuzzy kernel clustering, Journal of Fuzhou University(Natural Science Edition) **40** (2012), 181–187.

[17] X.X. Sun, X.X. Liu and Q.R. Xie, The Implementation of the fuzzy C-means clustering algorithm, Computer Applications and Software **25** (2008), 48–50.

[18] S.B. Zhou, Z.Y. Xu and X.Q. Tang, Method for determining optimal number of clusters in K-means clustering algorithm, Journal of Computer Applications **30** (2010), 1995–1998.