Technological Applied Sciences
ISSN: 1308 7223 (NWSATAS)
ID: 2018.13.3.2A0154

Status : Original Study
Received: November 2017
Accepted: July 2018

Özer Özdemir
Aslı Kaya
Anadolu University, Eskişehir-Turkey
ozerozdemir@anadolu.edu.tr; asli.k@anadolu.edu.tr

DOI	http://dx.doi.org/10.12739/NWSA.2018.13.3.2A0154	
ORCID ID	0000-0003-2446-5139	0000-0003-2155-9391
CORRESPONDING AUTHOR	Özer Özdemir	

GUSTAFSON-KESSEL AND FUZZY C-MEANS ALGORITHMS BY COLON CANCER DATA IN FUZZY CLUSTERING

ABSTRACT

Microarray technology has made it possible to simultaneously measure the expression levels of large numbers of genes in a short time. For the analysis of microarray data, clustering techniques are frequently used. So in this study, in cases where classical clustering analysis is insufficient to analyze data, fuzzy c-means algorithm and Gustafson-Kessel algorithm, which are improved to supply with advancing alternative statistical methods, are used. Firstly, the number of the optimum cluster was decided since the number of the cluster was not known at the beginning. Then, validity indexes and elbow criterion are applied to find the optimal number of clusters for both algorithms. It is seen that for both algorithms, the elbow was situated in the c=3 position as a result of the experimental result. At the end of the study, it is graphically stated that the fuzzy c-means algorithm is getting better clusters for the colon cancer dataset.

Keywords: Fuzzy Clustering, Fuzzy C-Means Algorithm, Cancer, Gustafson-Kessel Algorithm, Colon Cancer Data

1. INTRODUCTION

High throughput techniques are becoming more and more important in many areas of basic and applied biomedical research [13]. The emergence of microarray technology made it possible to trace expression levels of thousands of genes at the same time. Two statistical operations commonly applied to microarray data are classification and clustering but the most significant area is clustering microarray data analysis [1]. Since 40 years ago, clustering, which is one of the renowned data mining techniques, is being extensively studied and applied in numerous applications. The first step towards this aim is to adopt a mathematical description of the similarity. Clustering techniques use these mathematical descriptions to group genes in a given sample according to their expression profiles. Clustering algorithms allow each gene to locate the group containing its similar profiles. It is expected that genes in the same cluster have similar biological function. However, biological gene activities are very complex structures. It is known that given genes are subject to regulation by many manners of molecule. The general form of expression of a given gene may therefore correspond to the coincidence of different patterns. To determine this complexity and examine tightly related gene groups, fuzzy clustering algorithms are used that are faster in computing than the classical techniques and contain more flexible capabilities. In contrast to classical (hard) clustering algorithms, fuzzy clustering algorithms

How to Cite:
Özdemir, Ö. and Kaya, A., (2018). Gustafson-Kessel and Fuzzy C-Means Algorithms By Colon Cancer Data in Fuzzy Clustering, **Technological Applied Sciences (NWSATAS)**, 13(3): 243-252. DOI:10.12739/NWSA.2018.13.3.2A0154.

allow each gene to be bound to all clusters via a real valued vector. This vector takes values between 0 and 1 [2]. The use of membership values helps to identify genes associated with other genes or linked to more than one cluster, thus its biological complexity is discovered.

2. RESEARCH SIGNIFICANCE

We introduced Fuzzy C-means and Gustafson Kessel algorithms with the validation indices in the literature. In this paper, we aimed to consider separate groups according to similar expression patterns of gene data from colon cancer patients. A variety of methods have been proposed in the literature for colon cancer disease classification. As far as we know, clustering techniques have not been used in colon cancer data set so far. In this study, we reported successfully that Fuzzy C- Means method provides a more sensitive result. The clusters formed by the Fuzzy C- Means algorithm are well separated.

3. EXPERIMENTAL METHODS

3.1. Fuzzy Clustering Algorithms

Since fuzzy clustering algorithms deal with the uncertainty of real numbers, it helps to reveal clustering patterns that are appropriate for daily life experience. Fuzzy clustering algorithms also use mathematical descriptions, i.e. distance measures, to group similar expressions, such as clustering algorithms. However, unlike classical clustering techniques, each member uses membership functions that allow certain aggregates to be entered into a certain degree.

Membership is defined by:

$$u_{ij} : \forall i, \forall j \quad \text{for} \quad i = 1, 2, ..., n, \; j = 1, 2, ..., c$$

$$u_{ij} \succ 0$$

$$\sum_{j=1}^{c} u_{ij} = 1 \tag{1}$$

If any data is closer to the cluster center, the membership value of that cluster becomes the largest.

The sum of the membership grades of the given word is equal to 1. Fuzzy clustering algorithms usually use the objective function. Objective function based algorithms aim to solve clustering problem by turning it into optimization problem [3 and 4].

3.2. The Fuzzy C-Means Clustering Algorithm (FCM)

The most widely used algorithm, based on the least reduction of the objective function, was developed by Bezdek in 1974 [5].

This method is based on the fuzzy logic (1965) proposed by Zadeh [6].

The objective function used in the algorithm is as follows:

$$J(u, v) = \sum_{j=1}^{n} \sum_{i=1}^{c} u_{ij}^{m} \left\| x_j - v_i \right\|^2 \tag{2}$$

where n is the total number of patterns in a given data set and c is the number of cluster. X={x_1, x_2, …,x_n}$\subset R^s$ and V={v_1, …, v_c}$\subset R^s$ are the feature data and cluster centroids; and U=[uij]c×n is a fuzzy partition matrix composed of the membership grade of pattern x_j to each cluster i. $\left\| x_j - \upsilon_i \right\|^2$ is the Euclidean norm between x_j and v_i.

The weighting exponent m is called the being effective on the clustering performance of FCM [7].

The cluster centroids and the respective membership functions that solve the constrained optimization problem in Equation (2) are,

$$v_i = \frac{\sum_{j=1}^{n}(u_{ij})^m xj}{\sum_{j=1}^{n}(u_{ij})^m}, \quad 1 \le i \le c \qquad (3)$$

$$u_{ij} = \frac{1}{\left[\sum_{k=1}^{c}\left(\frac{\|x_j - v_i\|}{\|x_j - v_k\|}\right)^{1/(m-1)}\right]} \quad 1 \le i \le c, 1 \le j \le n \qquad (4)$$

These equations are obtained from iterative optimization process. The FCM algorithm is executed in the following steps:

Step 1: Given a pre-selected number of cluster c, a chosen value of m, initialize memberships u_{ij} of x_j belonging to cluster i such that $\sum_{i=1}^{c} u_{ij} = 1$.

Step 2: Calculate the fuzzy cluster centroid v_i for i = 1, 2,….. c using Equation (2).

Step 3: Update the membership u_{ij} using Equation (3).

Step 4: If the improvement in J(U, V) is less than a certain threshold (ε), then halt; otherwise go to step 2.

Here, ε is the stop criterion between 0 and 1, and t is the number of repetitions. Through this process J converges to a local minimum. The FCM algorithm depends on the randomly initialized values at startup and updates it iteratively using these values. Better performance can be achieved by using an algorithm to identify all centers or by repeatedly running the FCM with different start centers [1 and 8].

3.3. Gustafson Kessel Algorithm (GK)

Gustafson and Kessel, who developed the algorithm for fuzzy c-means, aimed to discover clusters in the form of ellipses. In Gustafson and Kessel's algorithm, Mahalanobis distance is used instead of Euclidean distance. Because Mahalonobis distance, forms clusters in the shape of ellipse while Euclidean distance is used to detect cluster-shaped clusters.

In this algorithm; objective function:

$$J(X;U,V,A) = \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}{}^{m} d^2 \qquad (5)$$

Mahalanobis distance equation is calculated as

$$d^2 = (x_i - \upsilon_j)^T A_i (x_i - \upsilon_j) \qquad (6)$$

The size of each annulus is defined for a local Ai norm reduction matrix that is used as one of the optimization variables in Equation (5).

This distance allows the norm to conform to the local topological structure of the data. The objective function is minimized by the GK algorithm using an alternative optimization method proposed by Gustafson Kessel (1979) [9].

A norm matrix;

$$\|A_i\| = \rho_i, \rho > 0$$
$$\|A_i\| = \left[\rho_i \det(F_i)\right]^{1/n} F_i^{-1} \qquad (7)$$

Algorithm steps:

Step 1: Adjust the initial value of c, m, the termination criterion ε and membership matrix U.

Step 2: Using the formula of the fuzzy set centers obtained when the objective function is minimized Equation (3), compute center

Step 3: Using Equation (6), the fuzzy covariance matrix is calculated for each set.

Step 4: Distances are calculated using Mahalanobis distance.

Step 5: The matrix of new membership values is calculated.

Step 6: Compare the new membership values with the old membership values. Depending on the termination criterion, the algorithm either stops or restarts the algorithm from the cluster center's account.

4. RESULTS AND DISCUSSION

Classification of microarray data which are used in diagnosis of cancer studies is one of the important topics in bioinformatics field [14]. So, in this study, microarray gene expressions of colon cancer patients were obtained from the National Center for Biotechnology Information (NCBI) database, which is open to access [11]. The colon cancer data set contains 62 samples and 7249 genes. Firstly, for both Fuzzy c-means algorithm and the Gustafson Kessel algorithm the optimal number of clusters has to be defined. The partition coefficient (PC), classification entropy (CE), partition index (SC), separation index (S), Xie and Beni's Index (XB), Dunn Index (DI) and Alternative Dunn’s Index (ADI) are used for determining optimum number of clusters. To find the optimal number of clusters, elbow criterion is used.

Elbow criterion is a common rule for determining which cluster number to select. Elbow criterion says that by graphing a validation measure explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph (the elbow). Unfortunately, this elbow is not always definitively identified [10]. Due to the large size of the data, long-term analysis and not being able to be drawn the graph of the cluster, MATLAB program was studied with [12]. For each study (between 2 and 10) a number of clusters must be determined, with a different number of clusters being determined; thus creating the optimum number of clusters. The results of values of validity indexes in the range of c=2,…,10 (when m=2) using FCM algorithm are shown in Table 1.

Table 1. Validation measures for FCM and m=2.00

Indexes C	PC	CE	SC	S	XB	DI	ADI
2	0.9839	0.0274	0.6193	0.0000828	5.0348	0.0343	0.0237
3	0.9517	0.0826	0.3000	0.0000722	4.3752	0.0077	0.0033
4	0.9134	0.1499	0.2908	0.0000738	4.3097	0.0072	0.0021
5	0.8915	0.1901	0.2483	0.0000526	3.9471	0.0052	0.0006
6	0.8592	0.2505	0.2244	0.0000535	3.5289	0.0064	0.0000252
7	0.8229	0.3202	0.2192	0.0000529	3.2661	0.0056	0.0000774
8	0.8076	0.3502	0.2214	0.00005	3.1788	0.0052	0.0000813
9	0.7743	0.4169	0.2195	0.0000516	3.3512	0.0050	0.0000087
10	0.7515	0.4651	0.2305	0.0000536	3.1027	0.0043	0.0000372

The Fuzzy c-mean algorithm ran for m=2. The validation indices’ results concern clustering of 2-10 clusters are depicted in Table 1. The values of the validation methods depending on the number of clusters will be plotted in Figure 1, Figure 2 and Figure 3.

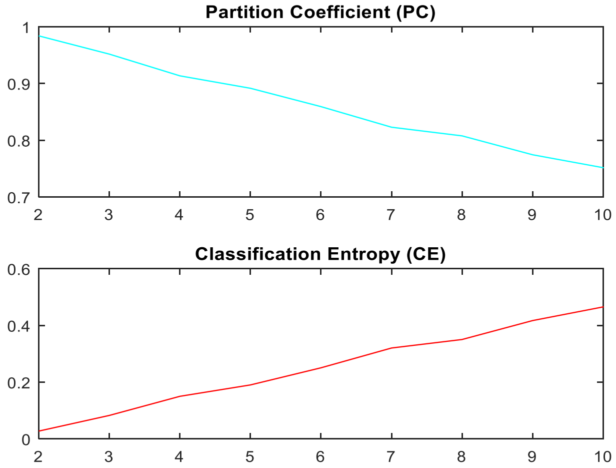

Figure 1. Partition coefficient (PC) and classification entropy (CE) where m=2.0 for FCM validation

In Figure 1 the results of the partition coefficient and the classification entropy are plotted. The biggest disadvantage of the PC is the monotonic decreasing with c, which makes it hardly to detect the ideal number of clusters. The same problem holds for CE: monotonic increasing caused by the lack of direct connection to the data. The optimal number of cluster cannot be rated based on those two validation methods. On the score of Figure 1, the number of clusters can be only rated to 4.

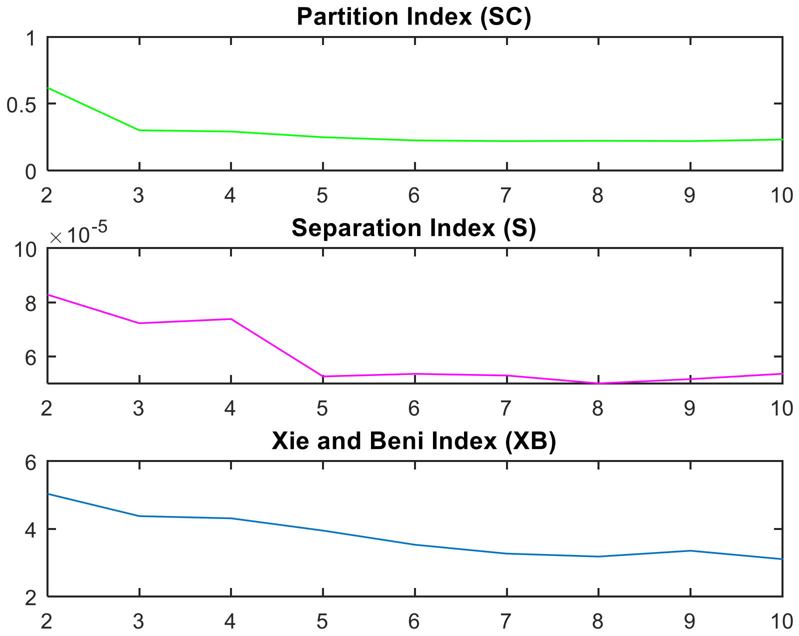

Figure 2. Partition index (SC), separation index (S) and XP index (m=2.0) for FCM validation

Figure 2 gives more information about the optimal number of clusters. While for XB has elbow is reached at c=3, SC index the is reached local minimum at c=3.

For the S index, it is difficult to find the optimal number of clusters. The points at c=3, c=5 and c=10, can be seen as an elbow.

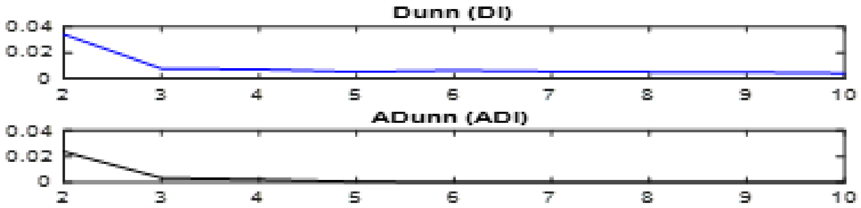

Figure 3. Dunn’s index (DI) and alternative Dunn’s index (m=2.0) for FCM validation

Besides, in Figure 3 it can be seen that DI and ADI index has an elbow is reached at c=3. The optimal number of clusters for the FCM algorithm is chosen to be 3. The results of the seven validation indices for each run of GK are shown in Table 2.

Table 2. Validation measures for GK

Indexes C	PC	CE	SC	S	XB	DI	ADI
2	0.5065	0.6861	1257.66	0.1682	3.7927	0.0008	0.0069
3	0.3439	1.0830	178.55	0.0257	2.5867	0.0009	0.0000158
4	0.2595	1.369	34.512	0.0062	1.9506	0.0007	0.0000067
5	0.2116	1.5853	30.92	0.0061	1.6042	0.0009	0.0000055
6	0.1788	1.7633	18.49	0.0032	1.3411	0.0011	0.0000189
7	0.1591	1.9000	1.29	0.0002	1.1859	0.0008	0.0000113
8	0.1428	2.0232	1.94	0.0004	1.0651	0.0009	0.0000005
9	0.1276	2.1454	2.11	0.0005	0.9375	0.0009	0.0000001
10	0.1187	2.2348	0.95	0.0002	0.864	0.0008	0.000017

The results of values of validity indexes in the range of c=2,…,10 (when m=2) using GK algorithm are shown in Table 2.

The values of the validation methods depending on the number of clusters will be plotted in Figure 4, Figure 5 and Figure 6.

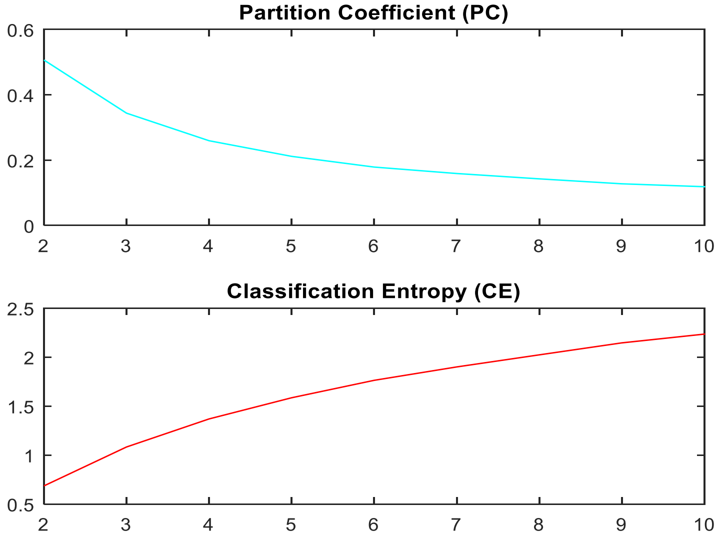

Figure 4. Partition coefficient (PC) and classification entropy (CE) where m=2.0 for GK

In the empirical studies, when the PC and CE indices are close to each other for a cluster numbers, it is seen that the cluster number is considered to be equal to the optimal number of clusters.

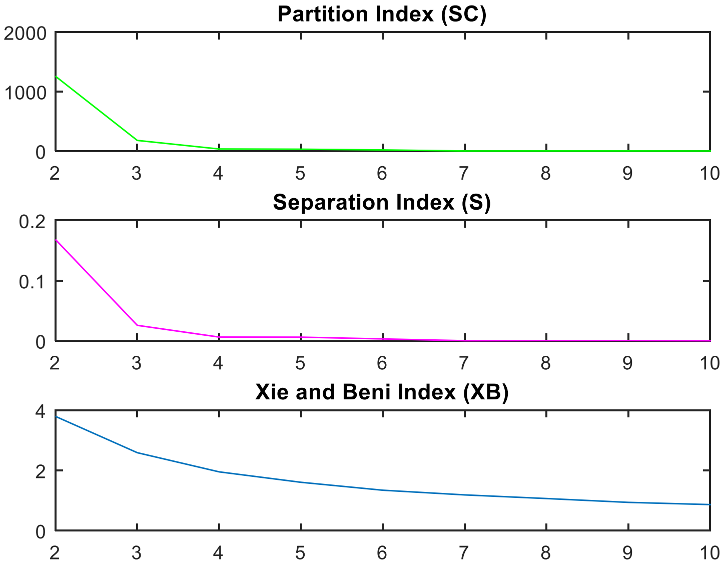

Figure 5. Partition index (SC), separation index (S) and XP index (m=2.0) for GK

Figure 5 displays that situation in c=2 for the PC and CE indices.

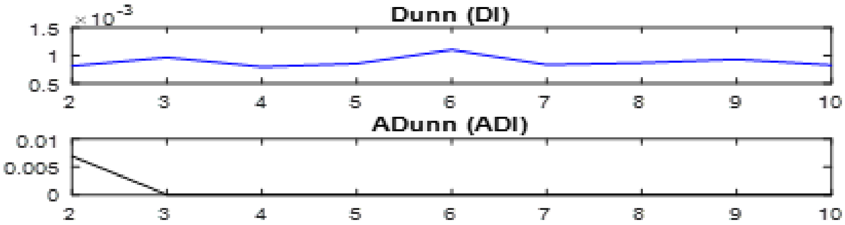

Figure 6. Dunn’s index (DI) and alternative Dunn’s index (m=2.0) for FCM validation

In Figure 6 show that the SC, S and XB indices the local minimum is reached at c=3. In Figure 7, for the DI index, it is difficult to find the optimal number of clusters. The points at c=4 and c=7, can be seen as an elbow.

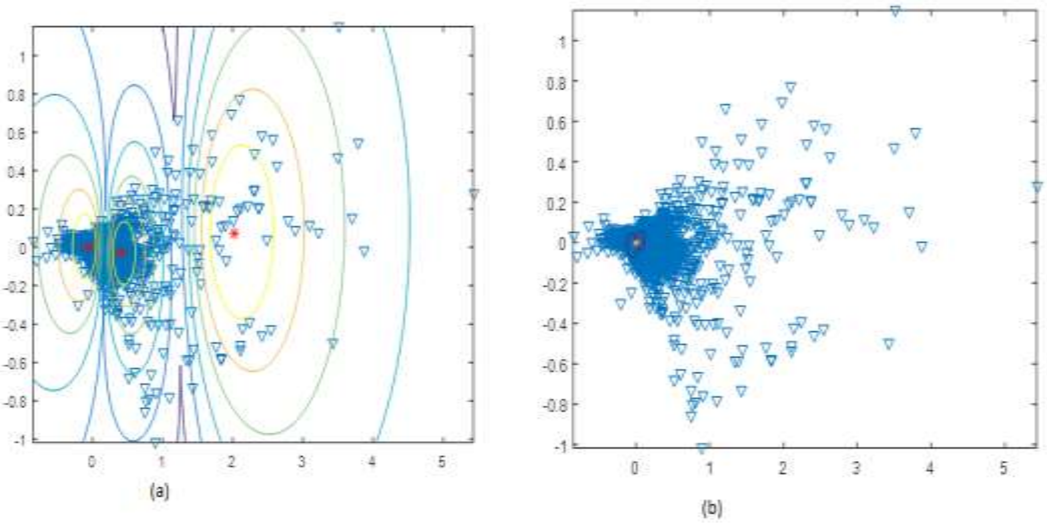

Figure 7. (a) Clustered Data Set with FCM algorithm and (b) Result of Gustafson-Kessel (GK) algorithm

According to these indices and considering that XB, SC and S are more useful, when comparing different clustering methods with the same c, the best partitioning of the data for Gustafson-Kessel algorithm is achieved with 3 clusters. After determining the optimal cluster number, the other initial values (the fuzzifier parameter m=2, and the stop criterion epsilon 1e-6) are randomly selected and the FCM algorithm and the GK algorithm are applied. As a result of the analyzes, clusters were obtained as in Figure 7(a) – 7(b). Figure 7(a) shows that Fuzzy c-means performed better for the colon cancer dataset creating better-separated and meaningful clusters with high compactness. However, the Gustafson-Kessel algorithm could not obtain well-separated and dense clusters (Figure 7(b)). The reason for this is that the Gustafson-Kessel algorithm constructs ellipsoid clusters by adopting the distance norm into the topological structure, the data set contains too noisy data and analyzing complex data sets more efficiently.

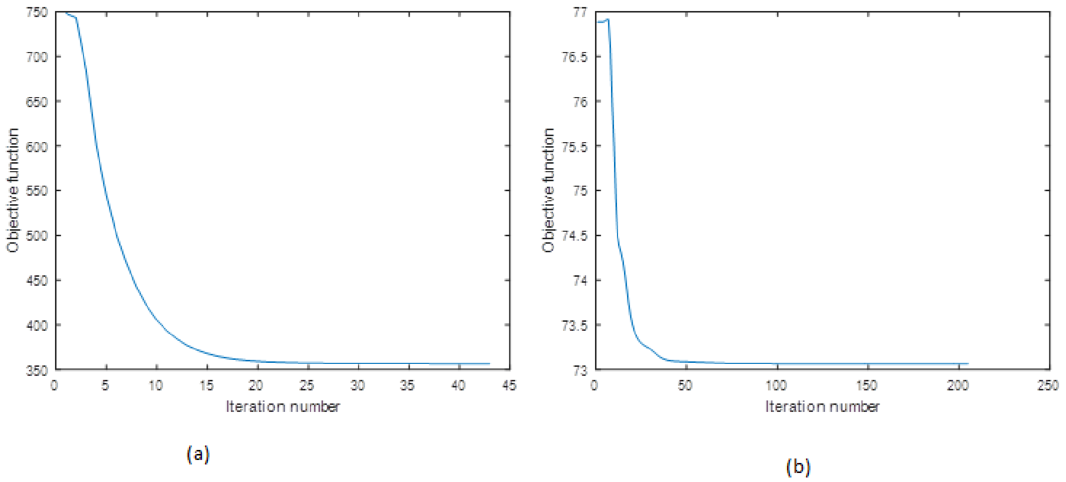

Figure 8. (a) Iteration number of FCM algorithm and (b) Iteration number of Gustafson-Kessel (GK) algorithm

As can be seen from Figure 8(a), no significant change in the objective function can be obtained after the 25th iteration and seen from Figure 8(b), no significant change in the objective function can be obtained after the 50th iteration.

5. CONCLUSION

Fuzzy clustering is an appropriate method for selecting genes that exhibit a tight association with given clusters. Conventional fractional clustering methods force clusters that do not match all genes to clusters or even variations in expression. Fuzzy set algorithms have been developed as an alternative to hard clustering techniques in high-dimensional data sets such as microarray gene expression data sets. In this study, Fuzzy c-means and Gustafson-Kessel algorithms were used for aiming to separate groups according to similar expression patterns of gene data from colon cancer dataset. Firstly, for both of algorithm, different scalar validity indexes as partition coefficient (PC), classification entropy (CE), partition index (SC), separation index (S), Xie and Beni's index (XB), Dunn's index (DI) and alternative Dunn's index (ADI) are used in validity index analysis. To find the optimal number of clusters, the so-called elbow criterion is applied. For FCM and GK algorithms the experimental results revealed that, the elbow was located at c=3. After determining the optimal number of cluster, both of algorithm were used on colon data. For both algorithms, all initial parameters are taken as the same. It can be stated that the Fuzzy c- means method provides a more sensitive result. The clusters formed by the FCM algorithm are well separated and compactness as shown in Figure 7(a). However, the Gustafson Kessel algorithm explores clusters in a spherical form, some clusters overlapping in this study (Figure 7(b)). The FCM algorithm may try to produce better results for this data than the GK algorithm. However, it should not be forgotten that the selected initial values and noisy data in dataset are very important affect for this result. A way of improving this research later on is to extent the number of cluster algorithms like hierarchical clustering or Gath Geva algorithm.

EXPLANATION

This study was supported by Anadolu University Graduate School of Sciences as BAP project (1703F081).

NOTICE

This study was presented as an oral presentation at the I. International Scientific and Vocational Studies Congress (BILMES 2017) in Nevşehir/Ürgüp between 5-8 October 2017.

REFERENCES

1. Avcı, U., (2006). Bulanık Kümeleme Algoritmalarının Karşılaştırmalı Analizi ve Bilgisayar Uygulamaları. Yüksek Lisans Tezi, Ege Üniversitesi Fen Bilimleri Enstitüsü, 78s.
2. Babuska, R., (1996). Fuzzy Systems, Modeling and Identification. Delft University of Technology Department of Electrical Engineering.
3. Höppner, F., Klawonn, F., Rudolf, K., and Runkler, T., (1999). Fuzzy Cluster Analysis: Methods for Classification Data Analysis and Image Recognition. John Wiley & Sons, p:5-75.
4. Wang, W. and Zhang, Y., (2007). On Fuzzy Cluster Validity Indices. Fuzzy Sets and Systems, 158, pp:2095-2117.
5. Bezdek, J.C., (1974). Cluster Validity with fuzzy sets. J. Cybernetics, Vol:3, pp:58-73.

6. Zadeh, L.A., (1965). Fuzzy Sets. Information and Control, 8, 338-353.
7. Bezdek, J.C., (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.
8. Bezdek, J.C., Ehrlich, R., and Full, W., (1984). FCM: Fuzzy C-Means Algorithm. Computers and Geoscience, 10(2-3), 191-203.
9. Gustafson, D.E. and Kessel, W.C., (1979). Fuzzy Clustering with a Fuzzy Covariance Matrix. IEEE CDC San Diego, 761-766.
10. Madhulatha, T.S., (2012). An Overview on Clustering Methods, IOSR Journal of Engineering, Vol:2(4) pp:719-725.
11. www.ncbi.nlm.nih.gov/geo.
12. www.mathworks.com/access/helpdesk/toolbox/fuzzy/.
13. Sturn, A., Quackenbush, J., and Trajanoski, Z., (2002). Genesis: Cluster Analysis of Microarray Data. Bioinformatics Applications Note, 18(1), 207-208.
14. Bilen, M., Işık, A.H., and Yiğit, T., (2015). A Hybrid Artificial Neural Network-Genetic Algorithm Approach for Classification of Microarray Data. 23th Signal Processing and Communications Applications Conference (SIU). IEEE.