Determining Number of Cluster in Fuzzy Clustering

Ozer OZDEMIR and Asli KAYA

Abstract— Clustering analysis is one of the multivariate statistical techniques that help to divide data groups of which are not exactly known to subgroups according to similarities and explore different correlation and structures in large data sets. In particular, fuzzy clustering analysis has recently been researched and used in various fields. Determining the number of cluster is an important task in fuzzy cluster analysis. In this study, notions relating to cluster validity were introduced and cluster validity indices in literature were reviewed. These indices were used in common genetic data set with fuzzy c-means algorithms and changeable fuzzifier parameter. The result was simply analyzed.

Keywords— Clustering, Fuzzy clustering, Fuzzy c-means, Validity index

1. INTRODUCTION

Clustering is an unsupervised classification method which is aimed to separate similar data to the same classes. A data in a multi-dimensional space is edited by coherent groups. While the data in same class are homogeneous, the different ones are heterogeneous which are not resembled to each other. The purpose of any clustering technique is find out clustering number *(c)* by changing *U(X)* partition matrices of $X = \left\{x_1, x_2, ..., x_n\right\}$,unlabeled. The size of *U* partition matrix is $c \;x\; n$ and represented as $U = \left[u_{ij}\right]$ $i = 1,...,c$ and $j = 1,...,n$ where u_{ij} is the membership of pattern x_j to clusters X_i. In hard clustering the following condition holds: this value is equal to 1 if $x_j \in X_i$ else is 0. The purpose is to classify data sets *X* such

$$X_i \neq 0 \quad \text{for } i = 1, 2, \dots, c, \tag{1}$$

$$X_i \cap X_j = 0 \quad \text{for } i = 1, 2, \dots, c,\; j=1,2,\dots, c \;\; i \neq j \tag{2}$$

and

$$\bigcup_{i=1}^{c} X_i = X. \tag{3}$$

When the fuzzy clustering is main subject, the conditions below is validated:

$$0 < \sum_{j=1}^{n} u_{ij} < n \;\text{ for }\; i = 1, 2, \dots, c, \tag{4}$$

$$\sum_{i=1}^{c} u_{ij} = 1 \quad \text{for } j = 1, 2, \dots, n, \tag{5}$$

$$\sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} = n. \tag{6}$$

2. THE FUZZY *C*-MEANS CLUSTERING ALGORITHM

Fuzzy *C*-Means (FCM) algorithm is developed from Hard *C*-Means algorithm in terms of partially belong a data to more than one cluster. FCM unsupervised classification algorithm defined through 1973. FCM attempts to find the most characteristic point in each cluster, which can be considered as the “centroid” of the cluster and, then, the grade of membership for each object in the clusters. Such aim is achieved by minimizing the objective function. Object function:

$$J(u, \upsilon) = \sum_{j=1}^{n} \sum_{i=1}^{c} u_{ij}{}^{m} \left\| x_j - \upsilon_i \right\|^2 \tag{7}$$

where *n* is the total number of patterns in a given data set and *c* is the number of clusters; $X = \left\{x_1, x_2, ..., x_n\right\} \subset R^s$ and *V* = {*v*1, . . . , *vc*} $\subset R^s$ are the feature data and cluster centroids; and $U = \left[u_{ij}\right]$ $c{\times}n$ is a fuzzy partition matrix composed of the membership grade of pattern x_j to each cluster *i*. $J(u, \upsilon)$ value is the total of pattern measurement of all weighted least square errors. The weighting exponent *m* is called the being effective on the clustering performance of FCM. The cluster centroids and the respective membership functions that solve the constrained optimization problem in (7) are

$$\upsilon_i = \frac{\sum_{j=1}^{n} (u_{ij})^m x_j}{\sum_{j=1}^{n} (u_{ij})^m} ,1 \leq i \leq c, \tag{8}$$

Ozer OZDEMIR is with the Department of Statistics, Anadolu University Eskisehir 26470 TURKEY (corresponding author to provide phone: 902223350580-4668; e-mail: ozerozdemir@anadolu.edu.tr).

Aslı KAYA is with the Department of Statistics, Anadolu University Eskisehir 26470 TURKEY (e-mail: asli.kaya532@gmail.com).

.

$$u_{ij} = \left[\sum_{k=1}^{c} \left(\frac{\|x_j - \upsilon_i\|^2}{\|x_j - \upsilon_k\|^2} \right)^{1/(m-1)} \right]^{-1}, 1 \le i \le c, 1 \le j \le n. \qquad (9)$$

These equations form the iterative optimization process. The aim is to iteratively improve a sequence of sets of fuzzy clusters until no further improvement in $J(u,\upsilon)$ is possible.

The FCM algorithm is executed in the following steps:

1) Initialize membership u_{ij} of *xj* belonging to cluster *i*., given a pre-selected number of cluster *c*, a chosen value of *m*.

2) Calculate the fuzzy cluster centroid v_i for i = 1, 2, . . . , c using Eq. (8).

3) Update the membership u_{ij} using Eq.(9).

4) Repeat *2)* and *3)* until the value of $J(u,\upsilon)$ is no longer decreasing.

3. Comparing of the Performances of Validity Indices

To test validity indices in a common used genetics data set, comparing of defined indices above are done with fuzzy c-means algorithm. In all experiments, different values used for fuzzier parameter m in algorithm and observed impact of cluster number of changing values. The test for convergence in the FCM algorithm was performed using $\varepsilon = 10^{-5}$, and the distance function $\|.\|$ was defined as Euclidean distance.

4.1. Yeast data

In this data set, the expression profiles of 6200 yeast genes were measured 0-160 min times period during two cell cycles in 17 hybridization experiments (Cho *et al.*, 1998). We used the same selection of 2845 genes made by Tavazoie *et al.* (1999).

4. Results

The main purpose of this section is to compare the performance of some of the above mentioned indices at the changing fuzzifier parameter level in determining the number of clusters. Test results for real data set have been reported.

c *Yeast data* *FCM*

m=1.15

c	PC	CE	S	MPC	Xie	Kwon
2	0.975	0.042	0.142	0.949	0.147	425.09
3	**0.968**	**0.053**	**0.166**	**0.952**	**0.176**	**511.17**
4	0.958	0.069	0.197	0.944	0.211	620.77
5	0.953	0.078	0.255	0.941	0.255	818.53
6	0.942	0.096	0.30	0.930	0.304	1001.8

m=1.5

PC	CE	S	MPC	Xie	Kwon
0.909	**0.155**	**0.124**	**0.818**	**0.134**	**388.86**
0.876	0.219	0.142	0.814	0.159	462.44
0.846	0.279	0.167	0.795	0.192	565.48
0.813	0.344	0.219	0.766	0.259	774.18
0.778	0.413	0.246	0.733	0.302	915.21

m=2

c	PC	CE	S	MPC	Xie	Kwon
2	**0.793**	**0.340**	**0.113**	**0.587**	**0.113**	**327.7**
3	0.703	0.535	0.127	0.554	0.127	370.8
4	**0.627**	0.706	0.155	0.502	0.155	459.2
5	0.559	0.863	0.210	0.448	0.210	636.6
6	0.505	0.996	0.220	0.407	0.220	681.2

m=2.5

PC	CE	S	MPC	Xie	Kwon
0.699	**0.469**	**0.124**	**0.398**	0.092	**266.031**
0.564	0.759	0.150	0.346	**0.090**	275.2
0.467	0.995	0.198	0.289	0.108	325.453
0.393	1.197	0.286	0.241	0.142	438.839
0.336	1.371	0.300	0.204	0.136	437.929

Table 1: Yeast Data Fuzzy Validity Indices

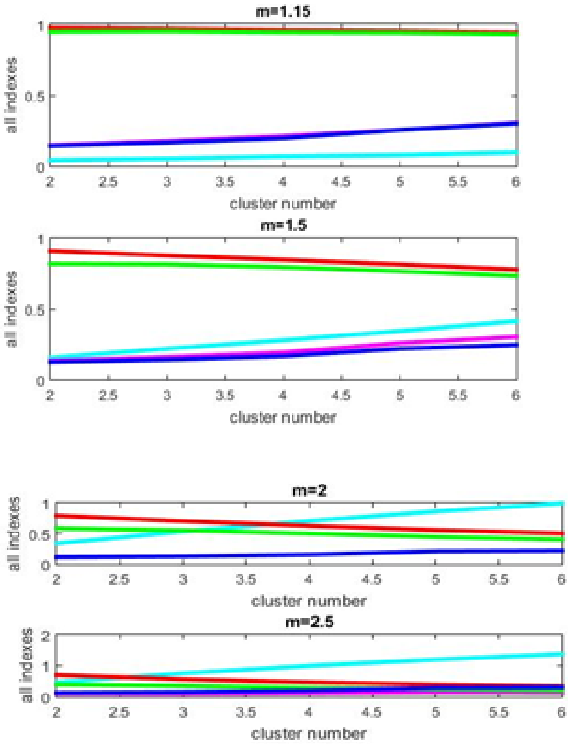

Figure 1: Validity Indices Graphs

The m and the number of clusters c values used for the FCM clusters of the yeast data set is given in Table 1. Except for the MPC and PC validity indices using the maximum value, the best cluster result can be determined by the minimum index value in other indices. The fuzzifier parameter selection should be m=1.15 as appropriate with this constraint. However, when determining the number of clusters, the value of the objective function should also be considered in the yeast data set, as the index values are very close to each other and the desired criteria.

c	*Object Function Value*
2	252953443.013685
3	144436156.901297
4	104203596.123227
5	85582718.0455856
6	75036793

Table 2: Object Function Values

5. Conclusion

This paper introduces the fundamental concepts of cluster validity, while a review of a number of fuzzy cluster validity indices available in the literature is presented. In addition, we conducted extensive comparisons of the mentioned indices in conjunction with the FCM algorithm on widely used data set.

It has been shown that both the number of clusters and the choice of the fuzzifier parameter are significant effects on the algorithm results in the experimental results and that each index does not always give the correct result at the level of the different fuzzifier parameters and that the use of the weighting parameter m= 2 in the general fuzzy clustering algorithm is not suitable for some data sets. For real data, it is clearly it is clearly more difficult to estimate the number of clusters. As a result, mostly, there is no available parameter set and an optimal solution should be applied by intuitively choosing the best candidate.

References

1] Bezdek J. C., "Cluster Validity with fuzzy sets", J. Cybernetics, Vol.3,1974, pp. 58-73

[2] Bezdek J.C., Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, NewYork, 1981.

[3] Gath I., Geva A. B., 1989. Unsupervised Optimal Fuzzy Clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, v.11 n.7, p.773-780.

[4] Güler N., 2006. Bulanık Kümeleme Analizi ve Bulanık Modellemeye Uygulamaları, Muğla Üniversitesi, Yayınlanmamış Yüksek Lisans Tezi, Muğla.

[5] L.A. Zadeh, Fuzzy sets, Inform. and Control 8 (1965).

[6] L. Xie and G. Beni, "A validity measure for fuzzy clustering, IEEE Trans PAMI, Vol. 13, No 8, 1991, pp. 841-847.

[7] Saad M.F, and ALIMI M. A., Validity Index and number of clusters, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3.

[8] Pal N.R, J. C. Bezdek, 1995, On Cluster Validity for the FCM Model,IEEE Transections On Fuzzy System.Vol.3, No.3,August

[9] Rezaee M.R., Lelieveldt B.P.F., Reiber J.H.C., 1998. A New Cluster Validity Index for the FCM, Pattern Recognition Lett., 19 p. 237-246

[10] Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. Nat. Genet., 22, 281–285.

[11] Wang W., and Zhang Y., (2007) On fuzzy cluster validity indices Fuzzy Sets and Systems 158 (2007) 2095 – 2117

[12] https://archive.ics.uci.edu/ml/datasets.html

Ozer Ozdemir was born in Turkey in 1982. He received his B.Sc., M.Sc. and Ph.D. degrees in statistics in the Department of Statistics at Anadolu University, Turkey, respectively in 2005, in 2008 and in 2013. He has worked as a Research Assistant from 2006-2008, as a Lecturer from 2008-2014 and as an Assistant Professor from 2014 in the Department of Statistics at Anadolu University, Turkey.
He has published over 50 international conference papers and journals in his research areas. His research interests include Applied Statistics, Simulation, Artificial Neural Networks, Fuzzy Logic, Fuzzy Modeling, Time Series,

Computer Programming, Statistical Software and Computer Applications in Statistics.

Aslı Kaya was born in Turkey in 1991. She received her B.Sc. degree in Statistics in the Department of Statistics at Anadolu University, Turkey, in 2014. She received her B.Sc. degree in the Department of Business in Faculty of Management at Anadolu University, Turkey, in 2015. She still continue her education as a master student in the Department of Statistics at Anadolu University, Turkey from 2015.