Turk J Math
29 (2005) , 431 – 438.
© TÜBİTAK

Maximum Entropy-Based Fuzzy Clustering by Using L_1-norm Space*

M. Ghorbani

Abstract

One of the most important methods in analysis of large data sets is clustering. These methods are not only major tools to uncover the underlying structures of a given data set, but also promising tools to uncover local input-output relations of a complex system. The goal of this paper is to present a new approach to fuzzy clustering by using L_1-norm space by means of a maximum entropy inference method, where, firstly, the resulting formulas have more beautiful form and clearer physical meaning than those obtained by means of FCM method and secondly, the obtained criteria by this new method are very robust. In order to solve the cluster validity problem and choosing the number of clusters in fuzzy clustering, we introduce a structure strength function as clustering criterion. With the proposed structure strength function, we also discuss the global minimum problem in terms of simulating methods.

Key Words: Fuzzy c-means, Maximum Entropy, Structure Strength, Number of Clusters, L_1-norm Space.

1. Introduction

One of the most important problems in the analysis of multivariate data is the identification of one or more relationships among the data and uncovering the underlying structure. One of the most applicable methods to research involving multivariate data,

*This paper has been supported by Research Institute for Fundamental Sciences, Tabriz, Iran.

medical research, genetics and other fields, are clustering methods. These methods are not only major tools that can be used to uncover the underlying structures of a given data set, but also promising tools to uncover the local input-output relations of a complex system. Therefore, the clustering problem is an optimization problem, where N objects with p characteristics are divided into C homogeneous groups, such that the distance between observations within a subgroup is smaller than the distance between observations belonging to different sub-groups [4, 6]. Among the existing clustering methods, the method of fuzzy C-means proposed by Bezdek [1] is one of the most active and applicable data analysis methods in recent years. Now, the first section of this paper briefly introduces the FCM method. In the second section, maximum entropy-based fuzzy clustering by using L_1-norm space is presented. The third section present the structure strength function and the DSR algorithm. The final section summarizes our findings.

2. Fuzzy C-means Clustering Method

Among existing clustering methods, one of the most active and applicable clustering methods is FCM. Let X_{ij} denote the jth random variable for the ith object, and $x_i = \{x_{i1}, \dots, x_{ip}\}$, are p characteristics of the ith observation. One of the criteria to improve initial partitions is the minimization of the relation

$$J_m = \sum_{i=1}^{n} \sum_{k=1}^{c} \sum_{j=1}^{p} u_{ik}^{m} ||x_{ij} - v_{kj}||^2, \qquad 1 \leq m < \infty, \tag{1}$$

where J_m is called the *loss function* (within-group sum-of-squared error); v_{kj} is interpreted as the prototype (or mean) of jth random variable in the kth cluster; m is a fuzzy parameter; and u_{ik} denotes the grade of membership of the ith object in the kth group and satisfies the following conditions:

1) $u_{ik} \in \{0, 1\}, \qquad 1 \leq i \leq n, 1 \leq k \leq c$

2) $\sum_{k=1}^{c} u_{ik} = 1, \qquad 1 \leq i \leq n$

3) $0 < \sum_{i=1}^{n} u_{ik} < n, \quad 1 \leq k \leq c,$

where the second condition is called the normalization constraint. For $m = 1$, FCM converges in theory to the traditional k-means solution [4, 6]. To minimize (1) subject to $\sum_{k=1}^{C} u_{ik} = 1$ by using Lagrangian multiplier methods, a considered point was demonstrated to be a local minimum solution of (1) if and only if

$$v_{kj}^{(r)} = \frac{1}{\sum_{i=1}^{n} \left(u_{ik}^{(r-1)}\right)^m} \sum_{i=1}^{n} \left(u_{ik}^{(r-1)}\right)^m x_{ij}, \qquad k = 1, \ldots, c$$

$$u_{ik}^{(r)} = \frac{1}{\sum_{s=1}^{c} \left(\frac{d_{ik}^{(r-1)}}{d_{sk}^{(r-1)}}\right)^{\frac{2}{m-1}}} \qquad i = 1, \ldots, n, \;\; k = 1, \ldots, c$$

where $d_{ik} = \sum_{j=1}^{p} (x_{ij} - v_{kj})^2$.
The iterative FCM algorithm is stopped if

$$\max |u_{ik}^{(r+1)} - u_{ik}^{(r)}| < \epsilon,$$

where ϵ is a small positive integer and r denotes number of iterations [1, 7, 11]. Though, FCM is a more applicable method, there are some objections against FCM:

1) Due to using Euclidean distance based on L_2-norm space, the presence of outliers in the data set degrades the quality of the computed clustering centers [7, 8, 9].

2) The physical meaning of the fuzziness parameter m and a way of choosing its optimal value is not well understood.

3) FCM deals with local minimum solutions only: it does not possess a mechanism with which one can get global minimum solutions.

4) There is no easy way to determine the number of clusters.

To solve these problems, in the next section, a new fuzzy clustering method by means of maximum entropy inference and using L_1-norm space is proposed.

3. Maximum Entropy-Based Fuzzy Clustering by Using L_1-norm Space

To express maximum entropy-based fuzzy clustering, we first introduce the entropy criterion. Let X be a random variable with probability mass function $P(x) = P(X = x)$, and set of possible values $\{x_i, \; i = 1, \ldots, n\}$. The entropy of random variable X is denoted by $H(X)$ and is defined by

$$H(X) = -\sum P(x) \log P(x).$$

Therefore, the structure of maximum entropy inference based on grades of membership $\{u_{ik}\}$ maximizes

$$\{-\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}\log u_{ik}\}. \tag{2}$$

As mentioned before, FCM tries to find set of prototypes and u_{ik}, that minimize the loss function with respect to normalization constraint. The goal of maximum entropy-based fuzzy clustering by using L_1-norm space is to maximize (2) with respect to

$$L = \sum_{i=1}^{n}\sum_{k=1}^{c}\sum_{j=1}^{p} u_{ik}||x_{ij} - v_{kj}||$$

and the normalization constraint. Here, L is a loss function (the within-group sum-of-absolute error), n is number of data pairs, c is the number of clusters, and v_{kj} and u_{ik} are the same as (1). To maximize (2) subject to the above conditions, we use the Lagrange multiplier rule. The Lagrange function is

$$\begin{aligned} L(u_{ik}, v_{kj}, \sigma, \lambda) &= -\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}\log u_{ik} \\ &+ \sigma(\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}d_{ik} - L) + \lambda(\sum_{k=1}^{c} u_{ik} - 1), \end{aligned} \tag{3}$$

where $d_{ik} = \sum_{j=1}^{p} ||x_{ij} - v_{kj}||$ [9]. Therefore, we see that with maximum-entropy based inference, the fuzzy clustering problem becomes one of finding a set of prototypes which minimize the loss function and membership assignment that satisfy the normalization constraint and maximize the Lagrange function.
Since v_{kj} is fixed, the Lagrange function is maximized via

$$u_{ik} = \frac{e^{\frac{-d_{ik}}{2\sigma^2}}}{\sum_{s=1}^{c} e^{\frac{-d_{sk}}{2\sigma^2}}}.$$

But maximizing (3) with respect to v_{kj}, $k = 1, \ldots, c$, $j = 1, \ldots, p$, u_{ik} being fixed, is equivalent to minimizing

$$L = \sum_{k=1}^{c}\sum_{j=1}^{p} L_{kj}, \qquad \perp \qquad L_{kj} = \sum_{i=1}^{n} u_{ik}^2|x_{ij} - v_{kj}|. \tag{4}$$

To minimize (4), at least two different methods may be used.
Method 1: Note that

$$L_{kj} = \sum_{i=1}^{n} |u_{ik}^2 x_{ij} - u_{ik}^2 v_{kj}|.$$

Suppose that $y_i = u_{ik}^2 x_{ij}$, $a = v_{kj}$ and $z_i = u_{ik}^2$, we have,

$$L_{kj} = \sum_{i=1}^{n} |y_i - a z_i|. \tag{5}$$

Let $b_i = \frac{y_i}{z_i}$. To minimize (5) with respect to a, the following algorithm can be used. Rearrange b_i in an ascending order. Then,

$$|b_{(1)}| + |b_{(2)}| + \ldots + |b_{(k-1)}| < \frac{1}{2}T$$
$$|b_{(1)}| + |b_{(2)}| + \ldots + |b_{(k-1)}| + |b_{K)}| > \frac{1}{2}T,$$

where $T = \sum |b_i|$ and $b_{(1)}, b_{(2)}, \ldots, b_{(n)}$ are set as order statistics for b_i. The value of a minimizing (5) is $\frac{y_{(k)}}{z_{(k)}}$ [2, 8].
Method 2: By noting (4), we learn that

$$L_{kj} = \sum_{i=1}^{n} u_{ik}^2 |x_{ij} - v_{kj}| = \sum_{i=1}^{n} w_{ik}(x_{ij} - v_{kj})^2,$$

where

$$w_{ik} = \frac{u_{ik}^2}{|x_{ij} - v_{kj}|}.$$

In this case, minimizing (3) is changed to the minimization of the weighted sum of squares. Clearly, the optimum solution is given as

$$v_{kj} = \frac{\sum_{i=1}^{n} w_{ik} x_{ij}}{\sum_{i=1}^{N} w_{ik}}.$$

Because of the dependence of w_{ik} and herewith u_{ik}, on v_{kj}, numerical methods must be used. Thus with the above in mind, we can formulate an algorithm to find the maximum

entropy-based fuzzy clustering as follows:
1) If the membership assignment $\{u_{ik}\}$ are fixed, then the centroid vectors $\{v_{kj}\}$ are a solution which locally minimize (3) if and only if

$$v_k^{(r)} = \frac{\sum_{i=1}^{n} u_{ik}^{(r-1)} x_{ij}}{\sum_{i=1}^{n} u_{ik}^{(r-1)}}. \tag{6}$$

2) If the centroid vectors $\{v_{kj}\}$ are fixed, then the grades of membership that minimize (3) by satisfying $\sum_{k=1}^{c} u_{ik} = 1$ are

$$u_{ik}^{(r)} = \frac{e^{\frac{-(d_{ik}^{(r-1)})^2}{2\sigma^2}}}{\sum_{s=1}^{c} e^{\frac{-(d_{sk}^{(r-1)})^2}{2\sigma^2}}} \quad \forall i, k, \tag{7}$$

where r denotes the number of iteration. The procedure stops when $|u_{ik}^{(r+1)} - u_{ik}^{(r)}| < \epsilon$.

3.1. Structure Strength and Data Structure Recognition Algorithm

One of the most important subjects in clustering is *determining the number of clusters* a problem known as *cluster validation.* This subject is widely discussed in [1], [5] and [11]. Some proposed criterion, such as partition coefficient, partition entropy and others strictly depend on the parameter m. That is, all proposed criteria are invalid when m is very large or small. In this paper, a new concept for determining the number of clusters is introduced and is called *structure strength.* The existence of data structure means that knowledge of a part allows us to guess the rest of the whole; therefore the process of structure strength is knowledge discovering. The structure strength of a system will be expressed by the following formula:

$$\begin{aligned} S &= \text{structure strength} \\ &= \text{(the effectiveness of the classification)} \\ &+ \text{(the accuracy of the classification).} \end{aligned}$$

For instance, if the number of observations and the number of clusters is n and k, respectively, and the loss function is $L(k)$, then the structure strength will be

$$S(k) = \alpha E + (1-\alpha) A = \alpha \log\left(\frac{N}{k}\right) + (1-\alpha)\log\left(\frac{L(1)}{L(k)}\right). \tag{8}$$

The first term decreases when k increases; but the second term increases with decreasing $L(k)$. On the other hand, clustering is inherently a structure recognition approach; its purpose is to find the strongest structure. Therefore, the search for the strongest structure is a process that maximizes $S(k)$ as the clustering criterion. Also, Geman and Geman [3] have proved that if σ^2 is inversely proportional to the logarithmic function of iteration, the global minimization of the loss function can be achieved. Therefore the data structure recognition algorithm is:
1) calculate $L(1)$.
2) for $k = 1, \ldots, C$ initialize u_{ik} for each i, k and let $r = 1, 2, \ldots$.
3) calculate $\{v_{kj}\}$ using (6) and $\{u_{ik}\}$ using (7).
4) if $\max_{\{i,k\}} |u_{ik}^{(r)} - u_{ik}^{(r-1)}| < \epsilon$, then calculate $S(k)$ using (8), ELSE r=r+1.
5) if $S(k) < S(k-1)$ stop, ELSE go to 2.

4. Conclusion

In this paper, a maximum entropy-based fuzzy clustering algorithm by using L_1-norm space has presented. Here, resulting formulas and criteria have very beautiful forms and a clearer physical meaning than those obtained by means of FCM. Unlike FCM, the criteria obtained in the present method were based on global minimum. Also, unlike traditional methods, where number of clusters is already known, in this method, data structure recognition by using the cluster strength function has been used for choosing the number of clusters.

Acknowledgement

I would like to express my gratefulness to the referees for their valuable suggestions and contributions.

References

[1] C. J. Bezdek, *Fuzzy Models for Pattern Recognition.* The Institute of Electrical and Electronics Engineers, Inc., New York (1992).

[2] D. Birkes, and Y. Dodge, *Alternative Methods of Regression*, Wiley (1991).

[3] S. Geman, and D. Geman, Stochastic relaxation, Gibbs distributions, and the bayesian restoration of image, *IEEE Trans. Pattern Anal. Machine Intell.* vol Pami-6 (1984), 721-741.

[4] M. Ghorbani, *Model Selection for Probabilistic Clustering Using Cross-Validated Likelihood*, Thesis, Tarbiat Modarres University, Tehran, Iran (2002).

[5] D. E. Gustafson, and C. W. Kessel, Fuzzy clustering with a fuzzy covariance matrix, in: K. S. Fu, Ed., *Proc. IEEE-CDC,* 2 (1979), 761-766.

[6] J. A. Hartigan, *Clustering Algorithms,* Wiley, New York (1975).

[7] A. Isazadeh, and M. Ghorbani, Fuzzy c-means and its generalization by L_p-norm space, *WSEAS Transaction on Mathematics*, 2 (2003), 168-170.

[8] K. Jajuga, L_1-norm based fuzzy clustering, *fuzzy sets and systems*, 39 (1991), 43-50.

[9] S. Miyamoto, and Y. Agusta, Efficient algorithms for L_p fuzzy c-means and their termination properties, *in Proc. 5th Conf. Int. Federation Classification Sco., Kobe, Japan* (1996), 255-258.

[10] S. Miyamoto, and M. Mukkaidono, Fuzzy c-means as a regularization and maximum entropy approach, *IFSA World Congress Proceedings*, 11 (1997), 86-92.

[11] H. J. Zimmermann, *Fuzzy Set Theory and Its Applications,* Second Edition. Kluwer Academic Publisher, Boston, London (1991).

M. GHORBANI
Faculty of Mathematical Science
Tabriz University
Tabriz-IRAN
e-mail: m.ghorhani@tabrizu.ac.ir

Received 06.09.2004