L_1-norm based fuzzy clustering

Krzysztof Jajuga

Economic Cybernetics Institute, Academy of Economics, ul. Komandorska 118/120, 53-345 Wrocław, Poland

Received May 1988
Revised September 1988

Abstract: The paper presents the L_1 version of the well-known fuzzy clustering method, namely fuzzy ISODATA, proposed by Bezdek and Dunn. Due to their robustness, L_1-norm based methods gained much attention in statistics.

The presented fuzzy clustering problem uses the distance between observations and location parameter vectors, which is based on the L_1-norm, instead of the inner product induced norm used in classical fuzzy ISODATA.

Two alternative methods to solve the L_1 fuzzy clustering problem are derived. In practice both membership grades and location parameter vectors are unknown. The paper presents two iterative algorithms, each the implementation of the derived method. Finally, numerical examples are presented. One of them refers to famous Iris data.

Keywords: Cluster analysis; fuzzy ISODATA; L_1-norm.

1. Introduction

Most classical statistical methods are based on the L_2-norm, which was preferred by statisticians. On the other hand, the results of theoretical studies indicate that L_1-norm based methods are more robust than those based on the L_2-norm. The paper presents the possible use of the L_1-norm in fuzzy clustering. The proposed method is the L_1 version of the most popular fuzzy clustering method, fuzzy ISODATA. In addition, iterative algorithms to solve the L_1-norm based fuzzy clustering problems are given. Finally, the results of two examples are presented.

2. L_1-version of fuzzy ISODATA

The fuzzy ISODATA method was proposed in [2] and [4]. It is a fuzzy version of the ISODATA clustering method, presented in [1]. The fuzzy ISODATA method consists in the minimization of the following objective function:

$$L = \sum_{i=1}^{n} \sum_{k=1}^{K} f_{ik}^{s} \|\boldsymbol{x}_i - \boldsymbol{v}_k\|^2, \tag{1}$$

under the constraints

$$0 \leq f_{ik} \leq 1, \quad i = 1, \ldots, n; k = 1, \ldots, K, \tag{2}$$

$$\sum_{k=1}^{K} f_{ik} = 1, \quad i = 1, \ldots, n, \tag{3}$$

where m is the dimension of space (the number of variables); n the number of observations (individuals, objects, units); K the number of fuzzy classes; f_{ik} the membership grade of the i-th observation to the k-th fuzzy class; $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{im}]^{\mathrm{T}}$ the i-th observation; $\boldsymbol{v}_k = [v_{k1}, v_{k2}, \ldots, v_{km}]^{\mathrm{T}}$ the location parameter vector for the k-th fuzzy class; s a weighting exponent; and $\|\cdot\|$ the inner product induced norm.

The classical version of fuzzy ISODATA uses Euclidean distance, based on the L_2-norm. Then, assuming $s = 2$,

$$L = \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{m} f_{ik}^2 (x_{ij} - v_{kj})^2. \tag{4}$$

In practice, both membership grades and location parameter vectors are unknown. Therefore, the minimization of the objective function L is composed of two problems:

(a) minimization of L with respect to f_{ik}, v_{kj} being fixed,

(b) minimization of L with respect to v_{kj}, f_{ik} being fixed.

We will consider the fuzzy clustering problem, where L_1-norm is used (this is not the inner product induced norm). Slightly modifying (1) and assuming $s = 2$, we get the following problem:

$$\text{minimize} \quad L = \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{m} f_{ik}^2 \, |x_{ij} - v_{kj}| \tag{5}$$

under constraints (2) and (3).

It can be proved that to minimize (5) with respect to f_{ik}, v_{kj} being fixed, the following formula can be used (as in classical fuzzy ISODATA):

$$f_{ik} = \frac{(d_{ik})^{-1}}{\sum_{s=1}^{K} (d_{is})^{-1}}, \quad i = 1, \ldots, n; k = 1, \ldots, K, \tag{6}$$

where

$$d_{ik} = \sum_{j=1}^{m} |x_{ij} - v_{kj}|, \quad i = 1, \ldots, n; k = 1, \ldots, K. \tag{7}$$

Now consider the problem of minimization of (5) with respect to v_{kj} ($k = 1, \ldots, K$; $j = 1, \ldots, m$), f_{ik} being fixed. This problem can be solved through Km separate minimization problems, since

$$\min L = \sum_{k=1}^{K} \sum_{j=1}^{m} \min L_{kj},$$

where

$$L_{kj} = \sum_{i=1}^{n} f_{ik}^2 \, |x_{ij} - v_{kj}|. \tag{8}$$

This separation is possible, because each component L_{kj} contains only one unknown v_{kj} and this does not depend on the other v_{kj}'s.

To minimize (8), at least two different methods may be proposed. They are based on two alternative presentations of (8).

Method 1. Note that

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

Clearly, this brings our problem of minimization of (8) to the minimization of

$$\sum_{i=1}^{n} |y_i - az_i| \tag{9}$$

where $y_i = f_{ik}^2 x_{ij}$, $z_i = f_{ik}^2$, $a = v_{kj}$ and the subscripts j and k were omitted for simplicity.

On the other hand, the problem of minimization of (9) is well-known and solved (the discussion is presented, for example in [3]). To solve this problem, the following algorithm can be used:

1. Determine $b_i = y_i / z_i$, $i = 1, \ldots, n$, assuming $z_i \neq 0$ $(i = 1, \ldots, n)$.

2. Rearrange the z_i's according to ascending order of b_i's and get $\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n$.

3. Minimize

$$\sum_{j=1}^{r} |\tilde{z}_j| - \sum_{j=r+1}^{n} |\tilde{z}_j|$$

with respect to r.

4. If the minimum is negative, take $a = b_r$. If the minimum is positive, take $a = b_{r+1}$. If the minimum is equal to zero, take $b_r \leqslant a \leqslant b_{r+1}$.

Method 2. Note that

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

where

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

This brings our problem to the minimization of the weighted sum of squares. Clearly, the optimal solution is given as

$$v_{kj} = \frac{\sum_{i=1}^{n} w_{ik} x_{ij}}{\sum_{i=1}^{n} w_{ik}}. \tag{10}$$

However, w_{ik} depends on v_{kj}, and thus the solution (10) cannot be used directly. Instead, it can be applied in iterative algorithms.

3. L_1 fuzzy clustering algorithms

In practice, both membership grades f_{ik} $(i = 1, \ldots, n;\ k = 1, \ldots, K)$ and location parameters v_{kj} $(k = 1, \ldots, K;\ j = 1, \ldots, m)$ are unknown. To get approximate solutions, iterative algorithms, similar to one proposed for classical fuzzy ISODATA, may be applied.

We present two algorithms. Each of them uses one of the methods described in Section 2. In both algorithms we are given:

(a) the m-variate observations:

$$\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{im}]^{\mathrm{T}}, \quad i = 1, \ldots, n;$$

(b) the number of fuzzy classes, K;

(c) the initial membership grades of all observations to fuzzy classes: f^0_{ik} ($i = 1, \ldots, n$; $k = 1, \ldots, K$), satisfying the following relations:

1. $0 \leqslant f^0_{ik} \leqslant 1, \quad i = 1, \ldots, n; k = 1, \ldots, K;$
2. $\sum_{k=1}^{K} f^0_{ik} = 1, \quad i = 1, \ldots, n;$
3. $\sum_{i=1}^{n} f^0_{ik} > 0, \quad k = 1, \ldots, K.$

In addition, in Algorithm 2, the initial location parameters are given, for example by the formula

$$v^0_{kj} = \frac{\sum_{i=1}^{n} f^0_{ik} x_{ij}}{\sum_{i=1}^{n} f^0_{ik}}, \quad k = 1, \ldots, K; j = 1, \ldots, m.$$

In the l-th iteration of both algorithms ($l = 1, 2, \ldots$):

1. Location parameters for each fuzzy class, v^l_{kj} ($k = 1, \ldots, K$; $j = 1, \ldots, m$) are determined.

1.1. In Algorithm 1 they are obtained through the minimization of the function

$$\sum_{i=1}^{n} |(f^{l-1}_{ik})^2 x_{ij} - v^l_{kj}(f^{l-1}_{ik})^2|. \tag{11}$$

Clearly, (11) is equivalent to (9) and the same method may be used.

1.2. In Algorithm 2, they are obtained by the formula

$$v^l_{kj} = \frac{\sum_{i=1}^{n} w^l_{ik} x_{ij}}{\sum_{i=1}^{n} w^l_{ik}}, \quad k = 1, \ldots, K; j = 1, \ldots, m,$$

where

$$w^l_{ik} = \frac{(f^{l-1}_{ik})^2}{|x_{ij} - v^{l-1}_{kj}|}, \quad i = 1, \ldots, n; k = 1, \ldots, K.$$

2. The distances of observations from location parameters vectors of each fuzzy class are determined, using the formula

$$d^l_{ik} = \sum_{j=1}^{m} |x_{ij} - v^l_{kj}|, \quad i = 1, \ldots, n; k = 1, \ldots, K.$$

3. The membership grades are updated, for each i, according to the following rule:

(a) if for each k, $d^l_{ik} \neq 0$, then

$$f^l_{ik} = \frac{(d^l_{ik})^{-1}}{\sum_{s=1}^{K} (d^l_{is})^{-1}}, \quad i = 1, \ldots, n; k = 1, \ldots, K;$$

(b) if exists s such that $d^l_{is} = 0$, then

$$f^l_{ik} = \begin{cases} 1, & k = s, \\ 0, & k \neq s, \end{cases} \quad k = 1, \ldots, K.$$

The iterative procedure given in 1, 2 and 3 is continued until the membership grades obtained in two consecutive iterations do not differ much, for example, if

$$\max_{i,k} |f^l_{ik} - f^{l-1}_{ik}| < \varepsilon, \tag{12}$$

ε being small positive constant (e.g. 0.001).

Finally, some remarks should be made on the determination of initial membership grades, f^0_{ik}. The natural proposal would be to start from the initial partition of observations, given by the function:

$$t : R^m \to \{1, 2, \ldots, K\}.$$

This function gives for each m-variate observation the initial number of the class it belongs to. To obtain this function different methods may be used, for example:

1. Systematic method, where

$$t(\boldsymbol{x}_i) = i \bmod K + 1, \quad i = 1, \ldots, n.$$

2. Random method, where

$$t(\boldsymbol{x}_i) = \text{entier}(\text{rnd}) + 1, \quad i = 1, \ldots, n,$$

rnd being a random number from the interval $(0, K)$.

3. Any hard (non-fuzzy) clustering method.

Then the initial partition is fuzzified, according to the formula:

$$f^0_{ik} = \begin{cases} \delta, & k \neq t(\boldsymbol{x}_i), \\ 1 - \delta(K-1), & k = t(\boldsymbol{x}_i), \end{cases} \quad k = 1, \ldots, K.$$

where δ is a small positive constant ($\delta \leqslant 0.01$).

4. Examples

To illustrate the performance of the proposed L_1 fuzzy clustering methods, we present the results of two examples. In both examples three methods were used. Two of them are L_1 methods, presented in the paper, the third one is classical fuzzy ISODATA, based on L_2-norm. In each method two alternative initial partitions, that is systematic partition and random partition, were used. However, in all cases, the results do not depend on the initial partition. Thus we present only the results obtained for random initial partition.

In all examples the criterion (12) as a termination rule for algorithms was used, where $\varepsilon = 0.001$. It is worth to note that method 1 is more time-consuming in terms of computer time.

Since in the examples the assignment of observations was known a priori (although the methods were applied as if it was unknown), the results allow us to make a brief comparison of methods. First of all, fuzzy classification generates 'hard' classification by assigning each observation to the class with highest membership grade. This classification is compared with the assignment known a priori, yielding the number of misclassified observations.

Example 1. In this example famous Iris data (presented in [5]) were used. Thus $n = 150$, $m = 4$, $K = 3$ (since there are three classes: Iris setosa, Iris virginica and Iris versicolour).

Method 1 misclassified 11 observations:

$$51, 53, 57, 71, 73, 77, 78, 84, 87, 107, 120.$$

Method 2 misclassified 15 observations:

$$51, 52, 53, 55, 57, 59, 66, 71, 73, 76, 77, 78, 84, 87, 107.$$

Classical fuzzy ISODATA misclassified 17 observations:

$$51, 53, 78, 100, 102, 107, 114, 120, 122, 124, 127, 128,$$
$$134, 139, 143, 147, 150.$$

It is worth noting that for L_1 methods misclassifications occurred basically for Iris versicolour (51–100) and for fuzzy ISODATA mainly for Iris virginica (101–150). The comparison of membership grades (not stated here) for all methods indicates that L_1 methods yield classifications which are more fuzzy than the classification obtained by L_2 fuzzy ISODATA. On the other hand, the differences in membership grades for method 1 and method 2 are insignificant.

Example 2. Here, 50 observations were drawn from the mixture of bivariate normal distributions. Two different partitions of observations into classes were assumed:

1. $n_1 = 25$, $n_2 = 25$;
2. $n_1 = 45$, $n_2 = 5$.

Also two different sets of mean vectors of mixture components were assumed:

1. $$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$
2. $$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 4 \\ 4 \end{bmatrix}.$$

Finally, three different sets of covariance matrices were assumed:

1. $$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$
2. $$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix},$$
3. $$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}.$$

Table 1. The average number of misclassified observations for different parameters of mixtures

Partitions	Mean vectors	Covariance matrices	The average number of misclassified observations		
			Method 1	Method 2	Fuzzy ISODATA
1	1	1	14.79	14.51	12.24
1	2	1	2.51	0.01	0.02
1	1	2	15.33	15.43	15.31
1	2	2	2.31	0.35	0.27
1	1	3	16.44	15.48	20.80
1	2	3	3.25	0.00	0.00
2	1	1	20.99	21.50	20.51
2	2	1	19.89	18.50	0.05
2	1	2	20.37	21.39	20.17
2	2	2	18.78	18.77	0.76
2	1	3	21.03	22.14	20.96
2	2	3	21.47	21.89	6.47

Thus, 12 experiments (3 × 2 × 2) were performed and in each experiment 100 repetitions were done. Three considered methods (that is, two L_1-norm based methods and fuzzy ISODATA) were compared with respect to the average number of misclassified observations. The results are presented in Table 1.

From the results it can be seen that:

(a) There are no indications of the superiority of any method in cases where the classes have the same size (understood as the number of observations belonging to them) – except for the case of negatively correlated variables, where the L_1-norm based methods performed slightly better.

(b) Method 2 and classical fuzzy ISODATA performed better than method 1 in the case of classes which have the same size and are well-separated.

(c) L_1-norm based methods performed much worse than classical fuzzy ISODATA in cases where the classes differ much in size.

5. Conclusions

As a summarization of all examples, some general conclusions can be stated:

(1) Method 2 should be preferred over method 1, because of its faster convergence and slightly better performance.

(2) All three methods do not depend on initial partitions.

(3) L_1 fuzzy clustering methods yield classifications which are more fuzzy than those obtained via L_2 fuzzy ISODATA.

(4) L_1-norm based methods are not recommended for classifications where the classes are of unequal size.

(5) L_1-norm based methods can be recommended for cases of approximately equal size of classes and departure from normality of the distributions in the classes.

Of course, these conclusions hold only for the examples which were presented here. Further studies ought to be done to find criteria to decide for which cases the L_1-norm based fuzzy clustering methods perform better than L_2 fuzzy clustering methods and vice versa.

References

[1] G.H. Ball and D.J. Hall, A clustering technique for summarizing multivariate data, *Behavioral Sci.* **12** (1967) 153–165.

[2] J.C. Bezdek, Numerical taxonomy with fuzzy sets, *J. Math. Biology* **1** (1974) 57–71.

[3] P. Bloomfield and W.L. Steiger, *Least Absolute Deviations* (Birkhäuser, Boston, MA, 1983).

[4] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters, *J. of Cybernet.* **3** (1973) 32–57.

[5] R.A. Fisher, The use of multiple measurements in taxonomic problems, *Ann. Eugenics* **7** (1936) 179–188.