Control and Cybernetics

vol. **24** (1995) No. 4

Optimization in fuzzy clustering

by

Krzysztof Jajuga

Academy of Economics in Wrocław,
ul. Komandorska 118/120,
53-345 Wrocław,
Poland

The paper reviews some of the fuzzy clustering methods based on the optimization approach. This approach applied in fuzzy clustering consists in optimization of the objective function reflecting the quality of clustering. The methods presented in the paper are the versions of the generalized fuzzy ISODATA method. The important cases outlined are: classical fuzzy ISODATA, fuzzy ISODATA with fuzzy covariance matrix, fuzzy linear varieties, fuzzy linear elliptotypes. The author refers also to some relatively new proposals, namely: the method based on L_1-norm and the general fuzzy clustering method for Minkowski distances. Finally, application of fuzzy clustering method for linear regression estimation is presented.

1. Fuzzy clustering - an introduction

Clustering methods are among the most common used statistical methods. They aim at partitioning of a set of (multivariate) observations into subsets, called clusters, so that the following conditions are satisfied:

- the observations belonging to the same cluster are as similar as possible;
- the observations belonging to different clusters are as dissimilar as possible.

In clustering methods similarity is usually measured via the distance between multivariate observations. Very often Euclidean distance or more general Minkowski and Mahalanobis distances are used.

Most of the clustering methods perform well in the case of well separated clusters. Such clusters occur for example when so called natural clusters exist. However, in real applications it often happens that the set of observations does not contain well separated clusters. The structure existing in this set is such one that there are no "sharp borders" between clusters. In such cases fuzzy approach proved to be useful. In fuzzy approach instead of clustering understood in usual sense (called crisp clustering or hard clustering), fuzzy clustering is used.

A fuzzy clustering problem can be stated as follows. Suppose that m-variate observations are given:

$$x_i = [x_{i1}, x_{i2}, \dots, x_{im}]^T, \; i = 1, \dots, n,$$

where:
n number of observations.

Fuzzy clustering is a family of K fuzzy sets determined on the set of m-variate observations, in such a way that the following conditions are satisfied:

1. $f_{ij} \geq 0 \quad i = 1, \dots, n; \; j = 1, \dots, K;$
2. $\sum_{j=1}^{K} f_{ij} = 1 \quad i = 1, \dots, n;$
3. $\sum_{i=1}^{n} f_{ij} > 0 \quad j = 1, \dots, K;$

where:
K the number of fuzzy clusters,
f_{ij} the membership grade of the i-th observation in the j-th fuzzy cluster.

Since fuzzy clustering is more general than hard clustering, it is more useful. It can be applied also when the clusters are not fuzzy. Then the membership grades are approximately equal to 0 or 1 and the hard clusters are easily detected.

It should be mentioned that sometimes clustering is used also for fuzzy data, that is when there is uncertainty in data. We will not consider this case. In this paper there is no uncertainty or vagueness in multivariate observations.

The idea of fuzzy clustering was outlined by Bellman, Kalaba and Zadeh (1966), and the first formal proposals of fuzzy clustering methods were given by Ruspini (1969). The most common approach in fuzzy clustering is the optimization approach. In this paper we review some fuzzy clustering methods based on the optimization approach. Due to the scope of this paper, we will limit ourselves to the most important methods.

2. Fuzzy ISODATA method

In optimization approach the objective function reflecting the quality of classification is determined and the fuzzy clustering is obtained through the minimization (or maximization) of this function. Many fuzzy clustering methods based on the optimization approach were proposed (see e.g. Backer, 1978; Ruspini, 1970,1973; Kaufman and Rousseeuw, 1990). One of the very first fuzzy clustering method based on the optimization approach was fuzzy ISODATA proposed by Bezdek (1973) and Dunn (1974). This is by no doubt the most well known fuzzy clustering method.

The objective function of early versions of fuzzy ISODATA can be regarded as a special case of the following function:

$$\sum_{i=1}^{n}\sum_{j=1}^{K} f_{ij}^{2} d_{ij} \tag{1}$$

where:

d_{ij} the distance between the i-th observation and the j-th fuzzy cluster.

The distance between the observation and fuzzy cluster reflects the similarity of this observation to fuzzy cluster. The higher this similarity, the higher should be the membership grade of this observation in a fuzzy cluster. Therefore, in "good" fuzzy clustering large values of d_{ij} correspond to small values of f_{ij} and vice versa. Thus, the objective function (1) has to be minimized.

It is worth to indicate that sometimes the more general objective function is considered, where the exponent 2 in (1) is replaced by s ($s > 1$). This parameter controls the fuzziness of the clustering. As s tends to 1, the clustering becomes less fuzzy. For $s = 1$, we obtain hard clustering. As s tends to infinity, the clustering becomes more fuzzy. In infinity all membership grades are equal to $1/K$.

In the objective function (1) both membership grades and distances between the observations and fuzzy clusters are not known. Therefore, minimization of the objective function is performed through two tasks:

- minimization with respect to values of f_{ij}, given values of d_{ij};
- minimization with respect to values of d_{ij}, given values of f_{ij}.

The solution of the first task is well-known. It can be proved (see e.g. Bezdek, 1981) that given values of d_{ij}, the optimal membership grades are given according to the following rule:

- if for some i, there exists k such that $d_{ik} = 0$, then:
$$f_{ij} = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}$$
- if for i, all $d_{ij} > 0$, then:
$$f_{ij} = (d_{ij})^{-1} / \left(\sum_{l=1}^{K} (d_{il})^{-1} \right) \tag{2}$$

In (2) the membership grade of the observation in the fuzzy cluster is given as a fraction of the similarity (a reciprocal of a distance) of this observation to the fuzzy cluster in the sum of the similarities of this observation to all fuzzy clusters.

The solution of the second task, namely the minimization with respect to values of d_{ij}, given values of f_{ij}, depends on the particular form of the distance.

The most simple case is the version, where squared Euclidean distance is used, so that the objective function is defined as:

$$\sum_{i=1}^{n}\sum_{j=1}^{K} f_{ij}^2 (x_i - v_j)^T (x_i - v_j) \tag{3}$$

where:
v_j the location vector of j-th fuzzy cluster.

Then the optimal location vectors are given as:

$$v_j = \left(\sum_{i=1}^{n} f_{ij}^2 x_i\right) / \left(\sum_{i=1}^{n} f_{ij}^2\right) \tag{4}$$

To conclude, it is worth to remind that this version of fuzzy ISODATA is useful when the clusters are of equal size and of hyperspherical shape (both conditions have to be met at least approximately).

3. Some other versions of fuzzy ISODATA method

The classical fuzzy ISODATA, where squared Euclidean distance is used, is a special case of the method based on the objective function (1). There are some other cases, useful from the point of view of real applications.

Gustafson and Kessel (1979) in one of the first extensions of classical fuzzy ISODATA method proposed to use squared Mahalanobis distance. Their objective function is therefore given as:

$$\sum_{i=1}^{n}\sum_{j=1}^{K} f_{ij}^2 (x_i - v_j)^T M_j (x_i - v_j) \tag{5}$$

where:
v_j the location vector of j-th fuzzy cluster,
M_j the symmetric and positive definite matrix reflecting the scatter of the j-th fuzzy cluster.

The objective function (5) takes into account the shape of clusters, allowing this shape to differ across the clusters. However this shape should be at least approximately hyperellipsoidal.

To make the minimization of the objective function tractable, each M_j is constrained by requiring its determinant to be fixed, for example by letting it equal to one. Allowing M_j to vary while keeping its determinant fixed corresponds to seeking an optimal cluster shape fitting the observations to a fixed volume.

This gives the following optimization problem:

minimize

$$\sum_{i=1}^{n}\sum_{j=1}^{K} f_{ij}^2 (x_i - v_j)^T M_j (x_i - v_j)$$

with respect to:

$$|M_j| = 1,\ j = 1, \dots, K.$$

It can be proved that given values of f_{ij}, the optimal parameters are given as:

$$v_j = \left(\sum_{i=1}^{n} f_{ij}^2 x_i\right) / \left(\sum_{i=1}^{n} f_{ij}^2\right)$$

$$M_j = |P_j|^{-1/m} P_j^{-1}$$

$$P_j = \left(\sum_{i=1}^{n} f_{ij}^2 (x_i - v_j)(x_i - v_j)^T\right) / \left(\sum_{i=1}^{n} f_{ij}^2\right)$$

So the location vector is the weighted mean vector and the scatter matrix is proportional to the inverse of the weighted covariance matrix, scaled in such a way that its determinant is equal to 1. In both cases the weights reflect the membership grades of the observations to fuzzy clusters.

The next version of fuzzy ISODATA method was proposed by Bezdek et al. (1981) and is called fuzzy linear varieties method. In this method each fuzzy cluster is represented by r-dimensional linear variety. Here a linear variety of dimension r $(0 \leq r \leq m)$ through an m-dimensional point v, spanned by the linearly independent m-dimensional vectors $s_1, s_2, \dots, s_r$, is the following set:

$$V_r = \left\{ w \in R^m | w = v + \sum_{j=1}^{r} t_j s_j,\ t_j \in R \right\}$$

Similarly as before, the objective function is based on the distances between the observations and the representations of fuzzy clusters (in this case the representations are linear varieties). This distance between observation and any linear variety is given as:

$$d(x_i, V_r) = [(x_i - v)^T (x_i - v) - \sum_{k=1}^{r} (x_i - v)^T s_k]^{0.5} \tag{6}$$

It is worth to mention that for $r = 0$ the distance (6) becomes Euclidean distance and we obtain the classical fuzzy ISODATA method.

The distance (6) is obtained by projecting $x_i - v$ onto the span obtained by $\{s_k\}$ and then calculating the length of $x_i - v$ minus its best least squares approximation. In the case of $r = 1$ it is simply the shortest distance from the point to a line. In general case it is the shortest Euclidean distance between the observation and any point belonging to the linear variety.

If we use this distance in the objective function (1), we get the following optimization problem: minimize

$$\sum_{i=1}^{n}\sum_{j=1}^{K} f_{ij}^2[(x_i - v_j)^T(x_i - v_j) - \sum_{k=1}^{r}(x_i - v_j)^T s_{kj}] \tag{7}$$

where:

$v_j, s_{1j}, s_{2j}, \ldots, s_{rj}$ unknown vectors defining linear r-dimensional variety for the j-th fuzzy cluster.

It can be proved that given the values of f_{ij}, the solution is given as:

$$v_j = \left(\sum_{i=1}^{n} f_{ij}^2 x_i\right) / \left(\sum_{i=1}^{n} f_{ij}^2\right)$$

s_{kj} being the eigenvector corresponding to the k-th largest eigenvalue of fuzzy scatter matrix S_j for the j-th fuzzy cluster, where:

$$S_j = \sum_{i=1}^{n} f_{ij}^2 (x_i - v_j)(x_i - v_j)^T$$

It is worth to mention that this method is useful for the detection of the clusters of linear shape. Some weakness of this method is the assumption that the linear varieties for different clusters are of the same dimension, which is not necessarily true for the data sets occurring in practice.

The next version of fuzzy ISODATA given by the objective function was also proposed by Bezdek et al. (1981). The method, being a slight modification of the fuzzy linear varieties method, is called fuzzy linear elliptotypes method. Here the objective function is defined as:

$$\sum_{i=1}^{n}\sum_{j=1}^{K} f_{ij}^2 \{\alpha[(x_i - v_j)^T(x_i - v_j)] + \\ +[(1-\alpha)((x_i - v_j)^T(x_i - v_j) - (x_i - v_j)^T s_j)]\} \tag{8}$$

The function (8) is a convex combination of two objective functions. The first one is function (3) of classical fuzzy ISODATA and the second one is function (7) of fuzzy linear varieties method. Here the distance between the observation and fuzzy cluster is a convex combination of the Euclidean distance between observation and fuzzy cluster center and the orthogonal distance from observation to the line passing through fuzzy cluster center. This idea can be useful if one looks for clusters of linear shape but also such clusters that contain a center near or in convex hull. The disadvantage of this method is the arbitrary choice of the coefficients of the convex combination.

The optimal solution in fuzzy linear elliptotypes method is exactly the same as in fuzzy linear varieties method. The only exception is of course the distance used in the objective function.

4. Minkowski distances in fuzzy clustering

In the fuzzy clustering methods presented so far, the distances of observations from fuzzy clusters were somehow derived from Euclidean distance. It is well known that Euclidean distance is a special case of Minkowski distance, given as:

$$d_{ij} = \left(\sum_{l=1}^{m} (x_{il} - v_{jl})^p \right)^{1/p} \tag{9}$$

where:
p parameter in Minkowski distance, in case of Euclidean distance $p = 2$.

Jajuga (1991) gives L_1-norm based fuzzy clustering method. This is the version of fuzzy ISODATA for the case of Minkowski distance where $p = 1$, corresponding to the known Manhattan distance, called also city block metric. By introducing this distance to the function (1), we get the following objective function:

$$\sum_{i=1}^{n} \sum_{j=1}^{K} f_{ij}^2 \sum_{l=1}^{m} |x_{il} - v_{jl}| \tag{10}$$

where:
v_{jl} the l-th component of the location vector of the j-th fuzzy cluster.

It can be proved that location parameters in the optimal solutions are obtained through the formula:

$$v_{jl} = \left(\sum_{i=1}^{n} a_{ijl} x_{il} \right) / \left(\sum_{i=1}^{n} a_{ijl} \right)$$

where:

$$a_{ijl} = f_{ij}^2 / |x_{il} - v_{jl}|$$

Recently a general fuzzy clustering method where the Minkowski distance is used, was proposed by Groenen and Jajuga (1994). Here the objective function is given as:

$$\sum_{i=1}^{n} \sum_{j=1}^{K} f_{ij}^2 \left(\sum_{l=1}^{m} |x_{il} - v_{jl}|^p \right)^{2h/p} \tag{11}$$

where:
$1 \le p \le 2$, $0 \le h \le 1$.

Here, in addition to varying parameter p, the second parameter, h, is introduced which allows for different powers of distance function. The function (11) contains as special cases the classical fuzzy ISODATA ($p = 2$, $h = 1$) and the L_1-norm based fuzzy clustering methods ($p = 1$, $h = 0.5$).

It should be added that this method was derived for the general case of fuzziness parameter s. For the sake of uniform presentation here we will restrict ourselves to the case of $s = 2$.

In Groenen and Jajuga (1994) the optimal solution is derived by using so called iterative majorization algorithm (see e.g. De Leeuw, 1988). It can be proved that the optimal parameter vectors for fuzzy clusters are given through the formula:

$$v_{jl} = \left(\sum_{i=1}^{n} a_{ijl} x_{il}\right) / \left(\sum_{i=1}^{n} a_{ijl}\right)$$

where:

$$a_{ijl} = f_{ij}^2 \left(\sum_{l=1}^{m} |x_{il} - v_{jl}|^p\right)^{(2h-p)/p} \qquad (x_{il} - v_{jl})^{p-2}$$

As we have already mentioned in all discussed methods the objective functions derived from general function (1) can be extended so that they are the special cases of an even more general objective function:

$$\sum_{i=1}^{n} \sum_{j=1}^{K} f_{ij}^s d_{ij},$$

where $s > 1$

The exponent s controls the fuzziness of the classification. In practice, however, the choice of $s = 2$ works quite well.

5. Fuzzy clustering in linear regression

In each of the presented methods the parameter vector of each fuzzy cluster can be considered as a location vector, like a "generalized cluster center". This allows to consider fuzzy clustering problem as the one of estimation of the location vectors.

However, in some methods other parameters of fuzzy clusters were also determined. For example – in fuzzy linear varieties method the hyperplane was another representation of fuzzy cluster. In fact fuzzy linear varieties method can be regarded as a generalization of two well known statistical methods.

First, if $r = 1$, that is if the linear varieties are lines, we get the generalization of principal component analysis. Strictly speaking, for each fuzzy cluster a kind of first principal component is determined using weighted scatter matrix.

Secondly, if $r = m - 1$, the linear varieties are $(m - 1)$-dimensional hyperplanes, we get the generalization of orthogonal regression. Here for each fuzzy cluster the best least squares fit is obtained and the observations are given weights reflecting their membership grades to fuzzy clusters.

In orthogonal regression the residuals are measured through the shortest distances of points from hyperplane. However more often classical regression is used, where the residuals are measured along the axis of the dependent variable. Therefore it may be useful to consider the fuzzy clustering, where the parameters of fuzzy clusters are regression hyperplanes. Here the regression of dependent variable with respect to the other variables, regressors, is taken. As a rule these hyperplanes are obtained via classical least squares method. The fuzzy clustering method for this case were proposed by Bezdek and Hathaway (1990) and Hathaway and Bezdek (1993) and independently by Jajuga (1993).

The main rationale behind the use of fuzzy clustering approach in linear regression lies in the fact that the requirement for the use of least squares method in linear regression is the homogeneity of the set of observations. However the observations very often form a heterogeneous set, which consists of several homogeneous clusters. Then it is advised to fit regression separately for each cluster. If the clusters are not well separated the detection of clusters by the classical clustering methods may fail. In addition, in order to fit the regression, a sufficient number of observations for each cluster should be given. This may cause problems for small data sets. This problem can be avoided by using fuzzy clustering ideas, since each fuzzy cluster contains all observations (with different membership grades) and there is no problem of small data set.

When applying fuzzy clustering to linear regression, we can still stay in the framework of general objective function (1), by using the function:

$$\sum_{i=1}^{n} \sum_{j=1}^{K} f_{ij}^2 (x_{im} - a_j^T z_i)^2 \tag{12}$$

where:

x_{im} value of the dependent variable for i-th observation,

$z_i = (x_{i1} x_{i2} \ldots x_{i,m-1} 1)^T$ the vector of the values of regressors for i-th observation,

$a_j = (a_{j1} a_{j2} \ldots a_{j,m-1} b_j)^T$ the vector of the regression coefficients for the j-th fuzzy cluster, b_j being the intercept.

In function (12) the distance of the observation from fuzzy cluster is defined as squared residual calculated for this observation basing on the regression of j-th fuzzy cluster. The objective function itself is a sum of the weighted squared residuals over all fuzzy clusters, where the weights reflect the membership grades in fuzzy clusters. This function may be also written as:

$$\sum_{j=1}^{K}(y - Za_j)^T F_j (y - Za_j) \tag{13}$$

where:

y a vector, whose components are values of dependent variable, x_{im},

Z a $n \times m$ matrix vector, whose columns are vectors z_i,

F_j an diagonal $n \times n$ matrix, where the elements on the main diagonal are $f_{1j}^2, f_{2j}^2, \ldots, f_{nj}^2$.

Given values of f_{ij}, the objective function (13) is minimized for:

$$a_j = (Z^T F_j Z)^{-1} Z^T F_j y$$

So the vectors of regression coefficients are obtained as weighted least squares coefficients, where the weights are squared membership grades.

To implement this method the iterative algorithm has to be used, where in the consecutive iterations the regressions for fuzzy clusters are updated and then squared residuals are used to update the membership grades.

References

Backer E. (1978) *Cluster Analysis by Optimal Decomposition of Induced Fuzzy Sets*. Delft University Press, Delft.

Bellman R.E., Kalaba R.E., Zadeh L.A. (1966) Abstraction and pattern classification, *Journal of Mathematical Analysis and Applications*, **13**, 1-7.

Bezdek J.C. (1973) Fuzzy mathematics in pattern classification, PhD thesis, Cornell University, Ithaca.

Bezdek J.C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York.

Bezdek J.C., Coray C., Gunderson R., Watson J. (1981) Detection and characterization of cluster substructure, I. Linear structure: fuzzy c-lines, II. Fuzzy c-varieties and convex combinations thereof, *SIAM Journal of Applied Mathematics*, **40**, 339-372.

Bezdek J.C., Hathaway R.J. (1990) Generalized regression and clustering, *Proceedings of the International Conference on Fuzzy Logic and Neural Networks*, Iizuka, 575-578.

De Leeuw J. (1988) Convergence of the majorization method for multidimensional scaling, *Journal of Classification*, **5**, 163-180.

Dunn J.C. (1974) A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters, *Journal of Cybernetics*, **3/3**, 32-57.

Groenen P., Jajuga K. (1994) Fuzzy clustering with squared Minkowski distances, submitted to Fuzzy Sets and Systems.

GUSTAFSON D.E., KESSEL W.C. (1979) Fuzzy clustering with a fuzzy covariance matrix, in: Gupta M.M., Ragade R.K., Yager R.R. (ed.), *Advances in Fuzzy Set Theory and Applications*, 605-620. North-Holland, New York.

HATHAWAY R.J., BEZDEK J.C. (1993) Switching regression models and fuzzy clustering, *IEEE Transactions on Fuzzy Systems*, **1,3**, 195-204.

JAJUGA K. (1991) L_1-norm based fuzzy clustering, *Fuzzy Sets and Systems*, **39**, s. 43-50.

JAJUGA K. (1993) *Statystyczna analiza wielowymiarowa*)Multivariate Statistical Analysis; in Polish). Wydawnictwo Naukowe PWN, Warszawa.

KAUFMAN L., ROUSSEEUW P.J. (1990) *Finding Groups in Data.* Wiley, New York.

RUSPINI E.H. (1969) A new approach in clustering, *Information and Control*, **15**, 22-32.

RUSPINI E.H. (1970) Numerical methods for fuzzy clustering, *Information Sciences*, **2**, 319-350.

RUSPINI E.H. (1973) New experimental results in fuzzy clustering, *Information Sciences*, **6**, 273-284.