A NEW GENERALIZED WEIGHTED CONDITIONAL FUZZY CLUSTERING

Jacek Łęski
Institute of Electronics
Silesian University of Technology
Akademicka 16, 44-101 GLIWICE, POLAND

Abstract

This paper proposes a family of generalized weighted conditional fuzzy c-means (GWCFCM) clustering algorithms. These algorithms include as a special case the well-known fuzzy c-means method (FCM) and the conditional fuzzy c-means method (CFCM). New clustering algorithm are compared experimentally with the fuzzy c-means using the Anderson's iris database.

1. Introduction

Clustering methods divide a set of N observations (input vectors) $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N$ into c groups denoted $\Omega_1, \Omega_2, \ldots, \Omega_c$ so that members of the same group are more similar than members of other groups. The number of clusters may be pre-specified or it may be decided by the clustering method. In Fig. 1 we have an example of seven two-dimensional vectors ($N = 7$). Intuitively, we may choose two clusters ($c = 2$) with a solidly marked curve or three clusters ($c = 3$) bound by a dotted curve. The solution in this case is rather easy, because data are two-dimensional. Generally, practice data vectors are p-dimensional, and can not be visualized like data in Fig. 1. The optimal (in the sense of the used criterion) solution can be theoretically by testing all possible partitions found. However,

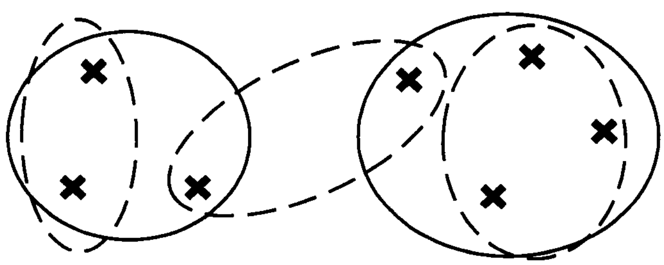

Fig. 1 An example of clustering for 7 data vectors.

in practice such an approach is unrealistic, because there is a very large number of possible combinations of data assignment. There are

$$\frac{1}{c!}\sum_{i=1}^{c}\binom{c}{i}(-1)^{(c-i)}\, i^N \tag{1}$$

partitions of N vectors into c nonempty subsets (see Feller, 1959). For example, there are $2.54\cdot 10^{347}$ different partitions for 500 data vectors grouped in 5 clusters. For data from Fig. 1, where $N = 7$, there are 63 different partitions in 2 clusters and 301 in 3 clusters. Partitions marked in the Figure seem to be the most reasonable. However, in practice clustered vectors are in high-dimensional space, and selection of optimal partitions must be performed automatically. The result of that clustering may be presented as partition matrix U. Dimension of that matrix is $c \times N$, and its elements are:

$$u_{ik} = \begin{cases} 1, & \underline{x}_k \in \Omega_i, \\ 0, & \underline{x}_k \notin \Omega_i. \end{cases} \tag{2}$$

where Ω_i stands for i-th cluster.

If we denote the vector space of all real $(c \times N)$-dimensional matrices over $\boldsymbol{R}$ as V_{cN}, then the set of all possible partition matrixes is defined by:

$$M_c = \left\{ U \in V_{cN} \mid 1°, 2°, 3° \right\}, \tag{3}$$

where the conditions are:

1° each $\underline{x}(k)$ is in or not in i-th cluster:

$$\underset{\substack{1\le i\le c\\ 1\le k\le N}}{\forall} \quad u_{ik} \in \{0, 1\}, \tag{4}$$

2° each $\underline{x}(k)$ is exactly in one of c cluster:

$$\underset{1\le k\le N}{\forall} \quad \sum_{i=1}^{c} u_{ik} = 1, \tag{5}$$

3° no cluster is empty:

$$\underset{1\le i\le c}{\forall} \quad 0 < \sum_{k=1}^{N} u_{ik} < N. \tag{6}$$

Generally, clustering methods can be divided into: hierarchical, graph theoretic, by decomposition of density function, by minimization of criterion function. A very popular way of clustering data is to define a criterion function (scalar index) that measures quality of any partition. The simplest and most frequently used criterion is the sum-of-square-error (Duda and Hart, 1973):

$$J_e(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik} \| \underline{x}_k - \underline{v}_i \|^2, \tag{7}$$

where $U = [u_{ik}]$ is partition matrix (2), $V = [\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_c] \in V_{pc}$ is a matrix consisting of cluster's centers. The most popular algorithm for approximating the minimum of J_e is the iterative process called the hard c-means (HCM) or ISODATA (Iterative Self-Organizing DATA clustering) (see Tou and Gonzalez, 1974).

The above method assume that each data vector can belong to one and only one class. This method can be natural for clustering compact and well-separated clusters. However, in practice clusters overlap, and some data vectors can belong partially to several clusters. A natural way to describe this situation results in the fuzzy set theory (Zadeh 1965), and belonging or membership of vector $\underline{x}_k$ to i-th cluster (u_{ik}) is a value from [0, 1] interval. This idea was first introduced by Ruspini (1969). The so-called fuzzy c-partition as a set of all possible fuzzy partitions to c clusters is defined by:

$$M_{fc} = \left\{ U \in V_{cN} \;\middle|\; \underset{\substack{1 \le i \le c \\ 1 \le k \le N}}{\forall} u_{ik} \in [0, 1], \underset{1 \le k \le N}{\forall} \sum_{i=1}^{c} u_{ik} = 1, \underset{1 \le i \le c}{\forall} 0 < \sum_{k=1}^{N} u_{ik} < N \right\}. \tag{8}$$

Fuzzy c-means criteria function has the form:

$$J_m(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} \left(u_{ik}\right)^m d_{ik}^2, \tag{9}$$

where $U \in M_{fc}$, $V \in V_{pc}$, d_{ik} is any inner product induced norm:

$$d_{ik}^2 = \| \underline{x}_k - \underline{v}_i \|_A^2 = (\underline{x}_k - \underline{v}_i)^T A (\underline{x}_k - \underline{v}_i), \tag{10}$$

where A is positive definite matrix, m is weighting exponent $m \in [1, \infty)$. Criteria (9) were introduced by Dunn (1973) for $m = 2$. Bezdek generalized (9) to the infinite family of fuzzy c-means criterion where $m \in [1, \infty)$. Using Lagrange multipliers the following theorem can be proved (see Bezdek 1982): If we fix parameters m and c, and define sets:

$$\underset{1 \le k \le N}{\forall} \begin{cases} I_k = \{ i \mid 1 \le i \le c;\, d_{ik} = 0 \}, \\ \tilde{I}_k = \{1, 2, \ldots, c\} \setminus I_k, \end{cases} \tag{11}$$

then $(U_{opt}, V_{opt}) \in (M_{fc} \times V_{pc})$ may be globally minimal for $J_m(U, V)$ only if:

$$\underset{\substack{1 \le i \le c \\ 1 \le k \le N}}{\forall} \quad u_{ik} = \begin{cases} \left(\dfrac{1}{d_{ik}} \right)^{\frac{2}{m-1}} \Bigg/ \left[\displaystyle\sum_{j=1}^{c} \left(\dfrac{1}{d_{jk}} \right)^{\frac{2}{m-1}} \right], & I_k = \varnothing, \\ 0, \quad \displaystyle\sum_{i \in I_k} u_{ik} = 1, & I_k \neq \varnothing, \end{cases} \tag{12}$$

and

$$\underset{1 \le i \le c}{\forall} \quad \underline{v}_i = \left[\sum_{k=1}^{N} (u_{ik})^m \underline{x}_k \right] \Bigg/ \left[\sum_{k=1}^{N} (u_{ik})^m \right]. \tag{13}$$

Optimal partition matrix U_{opt} is a fixed point of (12) and (13), and the solution is obtained from Pickard algorithm. This solution is called fuzzy ISODATA or fuzzy c-means (FCM), and can be described as:

1° fix c $(1 < c < N)$, $m \in [1, \infty)$. Initialize $U^{(0)} \in M_{fc}$,

2° calculate fuzzy centers $V^{(j)} = [\underline{v}_1^{(j)}, \underline{v}_2^{(j)}, \ldots, \underline{v}_c^{(j)}]$ using (13) and $U^{(j)}$,

3° update fuzzy partition matrix $U^{(j+1)}$ for $(j+1)$-th iteration using (12),

4° if $\| U^{(j+1)} - U^{(j)} \| > \varepsilon$, then, $j \leftarrow j + 1$, goto 2°.

In this algorithm, parameter m influences the fuzziness of the clusters; the larger is m the fuzzier are the clusters. For $m \to 1^+$, fuzzy c-means solution becomes the hard one, and for $m \to \infty$ the solution is as fuzzy as possible: $u_{ik} = 1/c$, for all i,k. There is no theoretical basis for the selection of m, and usually $m = 2$ is chosen.

In the paper by Pedrycz (1998) and (1998a) the method called conditional fuzzy c-means (CFCM) or context-dependent clustering was proposed. In this case data vectors $\underline{x}_k$ are clustered under conditions based on some linguistic terms defined in corresponding data vectors $\underline{y}_k$. These linguistic terms are treated as fuzzy relations, defined by membership functions. Finally we have, a corresponding value $f_k \in [0, 1]$ for each data vector $\underline{x}_k$. In this case c-partitions are defined as:

$$M_{cfc} = \left\{ U \in V_{cN} \;\middle|\; \underset{\substack{1 \le i \le c \\ 1 \le k \le N}}{\forall} \; u_{ik} \in [0, 1], \underset{1 \le k \le N}{\forall} \; \sum_{i=1}^{c} u_{ik} = f_k, \underset{1 \le i \le c}{\forall} \; 0 < \sum_{k=1}^{N} u_{ik} < N \right\}. \tag{14}$$

The necessary conditions for minimization criterion (9) under constraints (14) are (13) for cluster center, and:

$$\underset{\substack{1 \le i \le c \\ 1 \le k \le N}}{\forall} \quad u_{ik} = f_k \Bigg/ \left[\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{1-m}} \right]. \qquad \textbf{(15)}$$

This method develops clusters using similarity of data vectors $\underline{x}_k$ as well as conditional information from dependent variables.

2. New generalized weighted conditional fuzzy c-means

Incorporating additional information to the clustering process by (14) is one of the many possibilities. We propose following method based on the weighted generalized mean, described as conditions:

$$\underset{1 \le k \le N}{\forall} \quad c \biguplus_{i=1}^{c}{}_{(\alpha,\beta)} \, u_{ik} = f_k. \qquad \textbf{(16)}$$

where:

$$\biguplus_{i=1}^{c}{}_{(\alpha,\beta)} \, u_{ik} = \left(\sum_{i=1}^{c} \beta_i \left(u_{ik}\right)^{\alpha} \right)^{\frac{1}{\alpha}}, \qquad \textbf{(17)}$$

$\alpha \in (-\infty, 0) \cup (0, +\infty)$ and $\sum_{i=1}^{c} \beta_i = 1$.

For $\alpha = 1$, (17) gives the weighted arithmetic mean:

$$\biguplus_{i=1}^{c}{}_{(1,\beta)} \, u_{ik} = \sum_{i=1}^{c} \beta_i \, u_{ik}. \qquad \textbf{(18)}$$

For $\alpha \to 0$, we obtain the weighted geometric mean:

$$\biguplus_{i=1}^{c}{}_{(0,\beta)} \, u_{ik} = \prod_{i=1}^{c} \left(u_{ik}\right)^{\beta_i}. \qquad \textbf{(19)}$$

As $\alpha = -1$ and $u_{ik} \neq 0$, (17) approaches the weighted harmonic mean:

$$\biguplus_{i=1}^{c}{}_{(-1,\beta)} \, u_{ik} = \left(\sum_{i=1}^{c} \frac{\beta_i}{u_{ik}} \right)^{-1}. \qquad \textbf{(20)}$$

In this case we have a new c-partition type:

$$M_{gcfc} = \left\{ U \in V_{cN} \;\middle|\; \underset{\substack{1 \le i \le c \\ 1 \le k \le N}}{\forall} \; u_{ik} > 0,\; \underset{1 \le k \le N}{\forall} \; c \biguplus_{i=1}^{c}{}_{(\alpha,\beta)} \, u_{ik} = f_k,\; \underset{1 \le i \le c}{\forall} \; 0 < \sum_{k=1}^{N} u_{ik} < \infty \right\}, \tag{21}$$

The solution of minimization criterion function J_m given by (9) under conditions (16) is yielded by theorem: The necessary conditions for solution $(U_{opt}, V_{opt}) \in (M_{gfcf} \times V_{pc})$ of (9) with conditions (16) are:

$$\underset{\substack{1 \le i \le c \\ 1 \le k \le N}}{\forall} \; u_{ik} = \begin{cases} f_k \left(d_{ik}\right)^{\frac{2}{1-m}} \Big/ \left[c \biguplus_{j=1}^{c}{}_{(\alpha,\beta)} \left(d_{jk}\right)^{\frac{2}{1-m}} \right], & I_k = \varnothing, \\ \underset{i \in \tilde{I}_k}{\forall} \; 0,\; c \biguplus_{i \in I_k}{}_{(\alpha,\beta)} \, u_{ik} = f_k, & I_k \ne \varnothing, \end{cases} \tag{22}$$

and:

$$\underset{1 \le i \le c}{\forall} \; \underline{v}_i = \left[\sum_{k=1}^{N} \left(u_{ik}\right)^m \underline{x}_k \right] \Big/ \left[\sum_{k=1}^{N} \left(u_{ik}\right)^m \right]. \tag{23}$$

Proof: If we fix $V \in V_{cp}$, then columns of U are independent, and minimization of (9) can be done term by term:

$$\underset{1 \le k \le N}{\forall} \; g_k(U) = \sum_{i=1}^{c} \left(u_{ik}\right)^m d_{ik}^2. \tag{24}$$

LaGrangian of (9) is:

$$\underset{1 \le k \le N}{\forall} \; G_k(U, \lambda) = \sum_{i=1}^{c} \left(u_{ik}\right)^m d_{ik}^2 - \lambda \left(c\, F_i(u_{ik}) - f_k \right). \tag{25}$$

where $F_i(u_{ik})$ is abbreviation for function (17). Setting LaGrangian's gradient to zero yields:

$$\underset{1 \le k \le N}{\forall} \; \frac{\partial G_k(U, \lambda)}{\partial \lambda} = \left(c\, F_i(u_{ik}) - f_k \right) = 0, \tag{26}$$

and:

$$\underset{\substack{1 \le t \le N \\ 1 \le s \le c}}{\forall} \; \frac{\partial G_t(U, \lambda)}{\partial u_{st}} = m \left(u_{st}\right)^{m-1} d_{st}^2 - \lambda c \frac{\partial F_j(u_{jt})}{\partial u_{st}} = 0. \tag{27}$$

From (27) we get:

$$u_{st} = \left[\frac{c\lambda}{m}\right]^{\frac{1}{m-1}} \left[\frac{1}{d_{st}^2} \frac{\partial F_j(u_{jt})}{\partial u_{st}}\right]^{\frac{1}{m-1}}. \qquad \textbf{(28)}$$

From (26), (16) and using condition:

$$\underset{\substack{\varkappa>0 \\ 1\le k\le N}}{\forall} \quad \biguplus_{i=1}^{c}{}_{(\alpha,\beta)} \varkappa u_{ik} = \varkappa \biguplus_{i=1}^{c}{}_{(\alpha,\beta)} u_{ik}, \qquad \textbf{(29)}$$

we get:

$$c\left[\frac{c\,\lambda}{m}\right]^{\frac{1}{m-1}} F_j\left\{\left[\frac{1}{d_{jt}^2} \frac{\partial F_j(u_{jt})}{\partial u_{st}}\right]^{\frac{1}{m-1}}\right\} = f_t. \qquad \textbf{(30)}$$

Combining (28) and (30), and using condition (29) yields:

$$u_{st} = \left(d_{st}\right)^{\frac{2}{1-m}} f_t \Bigg/ c \biguplus_{j=1}^{c}{}_{(\alpha,\beta)} \left(d_{jt}\right)^{\frac{2}{1-m}}. \qquad \textbf{(31)}$$

If $I_k \neq \varnothing$, then choosing u_{ik} as in (22) results in minimal value of criterion (9), because partition matrix elements are zeros for non-zero distance, and non-zero for zero distances.

■

If we use weighted generalized mean (17) then for $I_k = \varnothing$:

$$\underset{\substack{1\le i\le c \\ 1\le k\le N}}{\forall} \quad u_{ik} = \frac{\frac{1}{c} f_k \left(d_{ik}\right)^{\frac{2}{1-m}}}{\left(\sum_{j=1}^{c} \beta_j \left(d_{jk}\right)^{\frac{2\alpha}{1-m}}\right)^{\frac{1}{\alpha}}}. \qquad \textbf{(32)}$$

For $f_k = 1$, we obtain weighted generalized c-means method, and for β_is equal to $1/c$, and α=1 we obtain conditional c-means.

There is no theoretical basis for the selection of β_j, and we use these weight, which provides a measure of the average distance between cluster center $\underline{v}_i$ and data vectors:

$$\underset{1 \le i \le c}{\forall} \quad \beta_i = \frac{\sum_{k=1}^{N} d_{ik}^2}{\sum_{j=1}^{c} \sum_{k=1}^{N} d_{jk}^2}. \tag{33}$$

3. Numerical example

The iris database is perhaps the best known database to be found in the pattern recognition literature. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. The data were collected by Anderson (1935). The vector of features consists of: 1).sepal length in cm, 2).sepal width in cm, 3).petal length in cm, 4).petal width in cm. We consider three classes of patterns: Iris Setosa, Iris Versicolour i Iris Virginica. For calculations parameter α equal to two was applied. A confusion matrix for 500 iterations and three classes has been shown in Table 1. In this Table the following abbreviations are utilized: FCM - the fuzzy c-means, GWCFCM (without context) - the generalized weighted conditional fuzzy c-means with f_ks equal to one, GWCFCM (with context) - the generalized weighted conditional fuzzy c-means with context based on linguistic term A defined on petal length with the following membership function:

$$f_k = A(x_{3,k}) = \exp\left\{-\frac{(x_{3,k} - 4)^2}{50}\right\}, \tag{34}$$

where $x_{3,k}$ denotes petal length for k-th data vector.

Table 1. Results for clustering of the iris problem.

Confusion Matrix								
FCM			GWCFCM (without context)			GWCFCM (with context)		
50	0	0	50	0	0	50	0	0
0	47	3	0	45	5	0	46	6
0	13	37	0	9	41	0	5	45

4. Conclusion

This paper proposes the infinite family of generalized weighted conditional fuzzy clustering method. The clustering problem is formulated, the well-known fuzzy c-means and conditional c-means are recalled. The generalized weighted conditional fuzzy c-means is introduced as a constrained minimization problem of criterion function. The necessary conditions (with proof) for obtain local minimum of the criterion function are shown. Simple numerical example on Anderson's iris database is also included. The existing fuzzy c-means and the conditional c-means methods can be obtained as a special case of the method proposed in this paper.

References

Anderson, E. (1935): The irises of the gaspe peninsula. Bull. Amer. IRIS Soc. 59, 2-5

Bezdek, J.C. (1982): Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York London

Feller, W. (1959): An introduction to probability theory and its applications. John Wiley, New York

Duda, R.O., Hart, P.E. (1973): Pattern classification and scene analysis. John Wiley & Sons, New York

Dunn, J.C. (1973): A fuzzy relative of the ISODATA process and its use in detecting compact well-separated cluster. Journal Cybernetics 3(3), 32-57

Pedrycz, W. (1998): Conditional fuzzy clustering in the design of radial basis function neural network. IEEE Trans. Neural Networks 9(4), 601-612

Pedrycz, W. (1998a): Fuzzy set technology in knowledge discovery. Fuzzy Sets and Systems 98, 279-290

Ruspini, E.H. (1969): A new approach to clustering. Inform. Control 15(1), 22-32

Tou, J.T., Gonzalez, R.C. (1974): Pattern recognition principles. Adison-Wesley, London

Zadeh, L.A. (1965): Fuzzy sets. Information and Control 8, 338-353