Crisp Interpretations of Fuzzy and Possibilistic Clustering Algorithms

Olfa Nasraoui and Raghu Krishnapuram
Department of Electrical and Computer Engineering
University of Missouri-Columbia,
Columbia, MO 65211
USA
Phone: (1) 314-882-7766
Fax: (1) 314-882-0397
E-Mail: raghu@ece.missouri.edu

1. Introduction

The Hard *C* Means algorithm (HCM) and the Fuzzy *C* Means algorithm (FCM) are commonly used in many pattern recognition and computer vision applications. When noise is present in the data set, both algorithms can give distorted results or even fail completely. Krishnapuram and Keller [5, 6] presented the possibilistic family of clustering algorithms which differs from the previous algorithms in that the membership of a point in a cluster is independent of all other clusters. Hence the difference between these three families of algorithms lies in the way that a point can belong to the clusters in the data set: hard memberships force a given point to belong exclusively to one cluster, fuzzy memberships allow a point to belong to all clusters with varying degrees, but the memberships in all clusters must sum to one, and possibilistic memberships are rather degrees of typicality of feature points in different clusters, where each cluster is considered to be independent of all other clusters.

In this paper, we derive equivalent crisp objective functions for each of the three families of algorithms. We show that the concept of memberships can be totally bypassed using these crisp reformulations. We also show that the objective functions of the hard and fuzzy *C* Means can in fact be derived from the same crisp objective function. We then show that the Possibilistic *C* Means is equivalent to *C* simultaneous and independent *W* or *M* estimators of the clusters' prototypes. To conclude, we propose a more general objective function which includes many common clustering algorithms as a special case.

2. Equivalent crisp objective functions for hard and fuzzy clustering algorithms

Let $\boldsymbol{X} = \{\boldsymbol{x}_j \mid j = 1...N\}$ be a set of feature vectors in an n-dimensional feature space with coordinate-axis labels $[x_1, x_2, \ldots, x_n]$, where $\boldsymbol{x}_j = [x_{j1}, x_{j2}, \ldots, x_{jn}]^T$. Let $B = (\beta_1,...,\beta_C)$ represent a C-tuple of prototypes each of which characterizes one of the C clusters. Each β_i consists of a set of parameters. In the following, we use β_i to denote both cluster i and its prototype. In the Hard C Means, only the distance from each feature vector to the cluster to which it is assigned (the closest one in the minimum distance classifier sense) is summed in the objective function, i.e., the objective function is defined to be

$$J_H(B; \boldsymbol{X}) = \sum_{i=1}^{C} \sum_{\boldsymbol{x}_j \in \beta_i} d^2(\boldsymbol{x}_j, \beta_i) \quad (1)$$

where $d^2(\boldsymbol{x}_j, \beta_i)$ represents the distance from a feature point $\boldsymbol{x}_j$ to cluster β_i. The above objective function can be written as

$$J_H = \sum_{j=1}^{N} \min_i d^2(\boldsymbol{x}_j, \beta_i). \quad (2)$$

It is well known that the min operator can be written as a limit of the following expression when $p \to -\infty$.

$$\min_i d^2(\boldsymbol{x}_j, \beta_i) = \lim_{p \to -\infty} \left\{ \sum_{i=1}^{C} d^2(\boldsymbol{x}_j, \beta_i)^p \right\}^{1/p} \quad (3)$$

Hence, we can rewrite J_H as

$$J_H = \lim_{p \to -\infty} J(p),$$

where

$$J(p) = \sum_{j=1}^{N} \left\{ \sum_{i=1}^{C} d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)^p \right\}^{1/p} \quad (4)$$

We assume that the prototypes are represented by cluster centers $\boldsymbol{c}_i$ and $d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i) = \|\boldsymbol{x}_j-\boldsymbol{c}_i\|^2$. Differentiating $J(p)$ with respect to the centers gives

$$\frac{\partial J(p)}{\partial \boldsymbol{c}_i} = -2 \sum_{j=1}^{N} d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)^{p-1} \left\{ \sum_{i=1}^{C} d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)^p \right\}^{(1-p)/p} (\boldsymbol{x}_j-\boldsymbol{c}_i) \quad (5)$$

Setting the above equation to 0 yields the following implicit equation which can be solved iteratively

$$\boldsymbol{c}_i(p) = \frac{\sum_{j=1}^{N} \left(\frac{d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)^p}{\sum_{k=1}^{C} d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_k)^p} \right)^{(p-1)/p} \boldsymbol{x}_j}{\sum_{j=1}^{N} \left(\frac{d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)^p}{\sum_{k=1}^{C} d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_k)^p} \right)^{(p-1)/p}} \quad (6)$$

Note that $\boldsymbol{c}_i(p)$ occurs on both sides of the above equation since $d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)$ involves $\boldsymbol{c}_i(p)$. Therefore one way to solve for $\boldsymbol{c}_i$ is to use the following equations in an alternating fashion.

$$\boldsymbol{c}_i(p) = \frac{1}{N_i} \sum_{j=1}^{N} w_{ij} \boldsymbol{x}_j \text{ , where } N_i = \sum_{j=1}^{N} w_{ij} \text{ and} \quad (7)$$

$$w_{ij} = \left(\frac{d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)^{\mathrm{p}}}{\sum_{k=1}^{C} d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_k)^{\mathrm{p}}} \right)^{(p-1)/p} \quad (8)$$

In FCM, the centers are updated using the following equation [1]

$$\boldsymbol{c}_{i\mathrm{F}} = \frac{\sum_{j=1}^{N} u_{ij}^m \boldsymbol{x}_j}{\sum_{j=1}^{N} u_{ij}^m} \quad (9)$$

The expression of $\boldsymbol{c}_i(p)$ is identical to the center update equation in (9) with the weights $w_{ij} = u_{ij}^m$, and $m = \frac{(p-1)}{p}$ (1,). Hence, the objective function of the FCM can be reformulated as the sum of generalized harmonic means

$$J_{\mathrm{F}} = \sum_{j=1}^{N} \left\{ \sum_{i=1}^{C} d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)^p \right\}^{1/p} \quad (10)$$

where p (- ,0) controls the amount of crispness or equivalently fuzziness in the resulting partition. At the lower end of the spectrum as $p \varnothing -$ or equivalently $m \varnothing 1$, minimizing the objective function becomes stricter, requiring prototypes that are reasonably close to every point This results in a hard partition with hard centers, i.e.,

$$J_{\mathrm{H}} = \lim_{p \varnothing -} J_{\mathrm{F}},$$

and the centers that minimize J_{H} are given by

$$\boldsymbol{c}_{i\mathrm{H}} = \lim_{p \varnothing -} \boldsymbol{c}_i(p). \quad (11)$$

It can be seen from the expression for the weights or memberships in the centers update equations in (8) that as $p \varnothing -$, the values of memberships become 1 if the point belongs to the given cluster (total commitment), or 0 otherwise (no commitment). As p or m increases, the objective function is minimized by choosing prototypes such that the generalized harmonic mean of the distances of every point to all prototypes is minimized, thus the partition

becomes fuzzier. For instance, when p = -1 (corresponding to m = 2 in the FCM), the objective function in (10) becomes

$$J_F = \sum_{j=1}^{N} \left(\frac{1}{\sum_{i=1}^{C} \frac{1}{d^2(\boldsymbol{x}_j, \beta_i)}} \right) \quad (12)$$

This objective function is minimized by choosing prototypes such that the harmonic mean distance of every point to all prototypes is minimized. At the higher end of the spectrum, as $p \varnothing 0$, it is easy to show that all cluster centers converge to a unique point, i.e.,

$$\lim_{p \varnothing 0} \boldsymbol{c}_i(p) = \frac{1}{N} \sum_{j=1}^{N} \boldsymbol{x}_j ,$$

Therefore all centers become identical, that is the partition is extremely fuzzy, yielding absolutely no information about the structure of the feature space.

From the above discussion, we see that the HCM and FCM are derived from the same objective function, and p (- ,0) or equivalently $m = \frac{(p\text{-}1)}{p}$ (1,) controls the amount of fuzziness present in the resulting partition. The partition gets less fuzzy for lower values of m and p, eventually reaching a hard partition when $m \varnothing 1$ or equivalently $p \varnothing$ - . It can be seen that the objective function in (10) does not really involve any fuzzy memberships, and the memberships (or weights) occur as temporary variables if we choose to minimize the objective function using the alternating optimization technique.

3. Equivalent crisp objective functions for the Possibilistic C Means

The Possibilistic C Means family of clustering algorithms is designed to alleviate the noise problem by relaxing the constraint on memberships used in the FCM [5, 6]. It uses the following objective function

$$\boldsymbol{J}_P(\boldsymbol{B}, \boldsymbol{U}; \boldsymbol{X}) = \sum_{i=1}^{C} \sum_{j=1}^{N} (u_{ij})^m \, d^2(\boldsymbol{x}_j, \beta_i) + \sum_{i=1}^{C} \eta_i \sum_{j=1}^{N} (1 - u_{ij})^m , \quad (13)$$

where η_i are suitable positive numbers. It is easy to show [6] that $\boldsymbol{U}$ may be a global minimum of $\boldsymbol{J}_m(\boldsymbol{B}, \boldsymbol{U}; \boldsymbol{X})$ only if the memberships are updated by

$$u_{ij} = \frac{1}{1 + \left(\frac{d^2(\boldsymbol{x}_j, \beta_i)}{\eta_i} \right)^{\frac{1}{m-1}}} . \quad (14)$$

We now derive the connection between the PCM and the W estimator [2] of robust statistics. Suppose we are interested in estimating the prototypes of C clusters simultaneously using the W estimator. We can consider minimizing the sum of C independent W estimator merit functions as follows

$$J_W = \sum_{i=1}^{C} \sum_{j=1}^{N} W(d^2(\boldsymbol{x}_j, \beta_i), \boldsymbol{a}_i) \; d^2(\boldsymbol{x}_j, \beta_i) \quad (15)$$

where W is a function of the distances and a parameter vector $\boldsymbol{a}_i$. Treating the weights as constants, the centers that minimize J_W are given by

$$\boldsymbol{c}_{iW} = \frac{\sum_{j=1}^{N} W(d^2(\boldsymbol{x}_j, \beta_i), \boldsymbol{a}_i) \, d^2(\boldsymbol{x}_j, \beta_i)}{\sum_{j=1}^{N} W(d^2(\boldsymbol{x}_j, \beta_i), \boldsymbol{a}_i)} . \quad (16)$$

Hence the centers and the weights are iteratively updated in an alternative manner. If the Cauchy estimator [3] is used, then the weights become

$$W(d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)\ ;\ \eta_i\ ) = \left[\frac{1}{1+\left(\frac{d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)}{\eta_i}\right)}\right]^{-1} \quad ,$$

where η_i are scale parameters that relate to the variance of each cluster. We see that the center update equation is identical to that of the PCM for the special case of $m = 2$, where each weight is seen to be a possibilistic membership or typicality of a feature point in a given cluster. The more general possibilistic weight function uses a fuzzifier m that affects its shape as will be discussed below,

$$W(d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)\ ;\ \eta_i\ ,\ m) = \left[\frac{1}{1+\left(\frac{d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)}{\eta_i}\right)^{1/(m-1)}}\right]^{m-1} \quad . \tag{17}$$

Given the equivalent weights between the possibilistic C Means and the W-estimator, the possibilistic objective function can now be rewritten in its equivalent form as

$$J_{\mathrm{P}} = \sum_{i=1}^{C}\sum_{j=1}^{N}\left[\frac{1}{1+\left(\frac{d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)}{\eta_i}\right)^{1/(m-1)}}\right]^{m-1} d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i) \quad .(18)$$

It is also interesting to note that J_{P} is of the form

$$J_{\mathrm{P}} = \sum_{i=1}^{C}\sum_{j=1}^{N} \rho(d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)\ ;\eta_i\ ,m) \tag{19}$$

where the loss function

$$\rho(d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)\ ;\ \eta_i\ ,\ m)\ = \left[\frac{1}{1+\left(\frac{d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)}{\eta_i}\right)^{1/(m-1)}}\right]^{m-1} d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)$$

$$= \left[\frac{d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)^{\ 1/(m-1)}}{1+\left(\frac{d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)}{\eta_i}\right)^{1/(m-1)}}\right]^{m-1}$$

Hence the PCM can also be considered as a robust estimator representing C independent simultaneous M-estimators [4], with η_i being the scale parameter related to the spread of the corresponding cluster, and m being a parameter that determines the shape of the figure of merit function.

4. Generalized Objective Function for Clustering

The crisp version of the objective function of the FCM in () can be generalized to the following functional form

$$J_f = \sum_{j=1}^{N} f^{-1}\left[\sum_{i=1}^{C} f(d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i),\boldsymbol{a}_i)\right] \tag{20}$$

where f is a monotonically non-increasing function and $\boldsymbol{a}_i$ is a vector of parameters. For the case of the FCM f is a power function

$$f(d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i),p) = \left\{d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i)\right\}^{p} \quad ,$$

In their deterministic annealing approach to clustering, Rose et. al. [7] used the following objective function

$$J_f = -\frac{1}{\beta}\sum_{j=1}^{N} \log\left[\sum_{i=1}^{C} \exp(-\beta\, d^2(\boldsymbol{x}_j,\boldsymbol{\beta}_i))\right].$$

In this case,

$$f(d^2(\boldsymbol{x}_j,\beta_i),\beta) = \exp(-\beta\, d^2(\boldsymbol{x}_j,\beta_i).$$

It can easily be shown that the centers that mininmize the generalized objective function J_f are given by the implicit equation

$$\boldsymbol{c}_{if} = \frac{\sum_{j=1}^{N}\left(\frac{f(d^2(\boldsymbol{x}_j,\beta_i))}{\sum_{k=1}^{C} f(d^2(\boldsymbol{x}_j,\beta_k))}\right)\boldsymbol{x}_j}{\sum_{j=1}^{N}\left(\frac{f(d^2(\boldsymbol{x}_j,\beta_i))}{\sum_{k=1}^{C} f(d^2(\boldsymbol{x}_j,\beta_k))}\right)}. \quad (21)$$

5. Conclusion

We have showed that the objective functions of the HCM anf FCM can be derived from the same non-fuzzy objective function. The only difference between them is that the surface of the objective function for FCM is smoother, since the summation in the FCM involves the distances from a given feature point to all clusters, inducing an averaging effect, whereas the summation in the HCM involves the distances from feature points assigned to a given cluster. This increases the HCM's chances to get stuck in local minima. From the center update equations, we can also conclude that for a noiseless data set, when both the HCM and FCM converge to the global minima of their respective objective functions, the centers' estimates found using HCM should be more accurate than those resulting from FCM, since in the latter, the center's calculation for a given cluster is affected by points belonging to other clusters. This tends to bring the centers closer to each other, i.e., if the clusters are well separated and noise-free, the HCM solution is better provided the algorithm converges to the global minimum. When the clusters are too close, the HCM centers estimates can be more distorted than those of the FCM due to arbitrary assignments of points that lie beween 2 clusters, whereas FCM assigns 0.5 membership in each cluster to these points, hence not drastically distorting the results. However for a noisy data set, the FCM estimates tend to be always more accurate than those of the HCM, since a noise point contributes only partially to the centers of all clusters in the FCM resulting in a less drastic effect on the centers estimates.

We have also established a connection between the PCM and robust *W*- and *M*-estimators. The PCM is more robust in the presence of noise because its objective function and center update equations involve unconstrained weights that decrease with the distance from the cluster centers. This results in low weights for outliers and decreases their influence. Another feature of the PCM is that it tries to find the *C* best clusters independently of each other, so it is possible that *C* identical clusters minimize the PCM objective function. However this only happens when the algorithm is badly initialized. The PCM's strength does not lie in "partitionning", but rather in finding valid clusters and giving a robust estimate of the centers after a suitable preliminary initialization.

Finally, we have shown that the objective functions for the HCM and FCM are part of a more general family of objective functionals, which opens the gate to considering new objective functions that fall into this category.

6. References

1. J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.
2. C. Goodall, “M-estimator of Location: An outline of the Theory,” in Understanding Robust and Exploratory Data Analysis, D.C. Hoaglin, F. Mosteller, and J. W. Tukey, Eds. New York: 1983, pp. 339-403.
3. P.W. Holland, and R. E. Welsch, “Robust Regression Using Iteratively Reweighted Least-Squares”, Commun. Statist.-Theor. Meth., vol. A6, no. 9, pp. 813-827, 1977.
4. P. J. Huber, *Robust Statistics*, John Wiley & Sons, New York, 1981.
5. R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering", *IEEE Transactions on Fuzzy Systems*, Vol. 1, No. 2, May 1993, pp. 98-110.

6. R. Krishnapuram and J. M. Keller, “Fuzzy and Possibilistic Clustering Methods for Computer Vision,” in *Neural and Fuzzy Systems*, S. Mitra, M. Gupta, and W. Kraske (Ed.), SPIE Institute Series, Vol. IS 12, 1994, pp.133-159.
7. K. Rose, E. Gurewitz and G. Fox, “A deterministic annealing approach to clustering”, *Pattern Recognition Letters*; Vol. 11, No. 9, Sep 1990, pp. 589-594.