The Department of Computer and Information Science
University of Genova, Via Dodecaneso 35 - 16146 Genova Italy
**DISI Technical report no. DISI-TR-02-03**

# Soft Transition from Probabilistic to Possibilistic Fuzzy Clustering

Francesco Masulli [1] Stefano Rovetta [2]

April 19, 2002

[1]Department of Computer Science, University of Pisa, Italy, and INFM, the National Institute for the Physics of Matter. E-mail: masulli@di.unipi.it

[2]Department of Computer and Information Sciences, University of Genoa, Italy, and INFM, the National Institute for the Physics of Matter. E-mail: rovetta@disi.unige.it

**Abstract**

We discuss the *graded possibilistic model*. We review some clustering algorithms derived from the basic *c*-Means and introduce a formalism to provide an alternative, unified perspective on these clustering algorithms, focused on the memberships rather than on the cost function. An interesting case is the concept of graded possibility. Its formulation includes as the two extreme cases the "probabilistic" assumption and the "possibilistic" assumption. A possible formulation can be stated as an interval equality constraint enforcing both the normality condition and the required graded possibilistic condition. We outline a basic example of graded possibilistic clustering algorithm.

The experimental demonstrations presented aim at highlighting the different properties attainable through appropriate implementation of a suitable graded possibilistic model.

**Keywords:** Possibilistic clustering, Clustering methods, Possibility theory, Fuzzy clustering, Fuzzy statistics and data analysis.
**PACS codes** 07.05.Mh 07.05.Kf
**MSC codes** 68T05 62H30

# 1 Introduction

The clustering problem is usually stated as the task of partitioning a set of data vectors or patterns $X = \{x_k\}$, $k \in \{1,\ldots,n\}$, $x_k \in \mathbb{R}^n$ by attributing each data point $x_k$ to a subset $\omega_j \subset X$, $j \in \{1,\ldots,c\}$, defined by its *centroid* $y_j \in \mathbb{R}^n$. This attribution is made based on a given distance $d(\cdot,\cdot)$.

The most widely used clustering method is probably the *Fuzzy c-Means/Fuzzy ISODATA* [1][2][3] (FCM) algorithm, which is a "fuzzy relative" to the simple $c$-Means technique [4]. FCM defines the $\omega_j$ as fuzzy partitions of the data set $X$. Variations over this basic scheme try to overcome some of its well-known limitations. The *Deterministic Annealing* (or *Maximum Entropy*) approach [5][6] does not minimize a simple cost term, but a compound cost function which is the sum of a distortion term $\hat{E}$ and an entropic term $-H$ (see the next section for the mathematical definitions). The optimization is done by fixing a constant value for one of the two terms and minimizing the other; then this step is iterated for decreasing values of the constant, until a global optimum is reached. This alleviates the false minima problem of standard $c$-Means and (to a lesser extent) of FCM.

In decision-making and classification applications, algorithms should feature several desirable properties in addition to the basic discrimination or decision function. For instance, it is normally required that in certain configurations a decision is not made (*pattern rejection*). This situation typically occurs in the presence of outliers. This problem is very well-known and well studied (see for instance [7][8][9]), and is tackled in a convenient way within the framework of soft-computing, fuzzy, and neural approaches [10][11][12].

However, the clustering problem as stated above implies that the outlier rejection property cannot be achieved. This is because the membership values are constrained to sum to 1. By giving up the requirement for strict partitioning, and by resorting to a "mode seeking" algorithm, Krishnapuram and Keller proposed the so-called *possibilistic approach* [13][14], where this constraint is relaxed essentially to

$$u_{jk} \in [0,1] \quad \forall k, \forall j \tag{1}$$

With this model outlier rejection can be achieved, but at the expense of a clear cluster attribution and other computational drawbacks. The same issue of analysing the membership interactions on a local basis, as opposed to the global effects induced by the probabilistic model, is considered in [15].

Finally, we should mention that there are many other popular variations involving modified distance criteria to improve robustness or to account for different cluster geometries (such as the Gustafson-Kessel method [16] or Bezdek's Fuzzy $c$-Varieties algorithm [17]), but the issues they address are not directly related to the present study.

In the remainder of this paper, we discuss the *graded possibilistic model*, which introduces notable flexibility in the clustering process, while at the same time allowing for some behaviors (such as outlier rejection) not attainable with standard approaches.

# 2 Some popular clustering algorithms: a unified view

## 2.1 The $c$-Means family

We will now review some clustering algorithms derived from the basic $c$-Means: ("hard") $c$-Means (HCM) [4], entropy-constrained fuzzy clustering by Deterministic Annealing

(DA) [5], Possibilistic $c$-Means with an entropic cost term (PCM-II) [14], Fuzzy $c$-Means (FCM) [2]. All of these techniques are based on minimizing the following cost function:

$$\hat{E} = \sum_{j=1}^{c} \sum_{k=1}^{n} u_{jk} d_{jk}. \tag{2}$$

(this includes also FCM, although in the usual formulation this is not evident; see the Appendix). We will refer collectively to these algorithms as the $c$-Means (CM) family.

Here $u_{jk} \in U$ is the degree of membership of pattern $x_k$ to cluster $\omega_j$ and $Y = \{y_1, \ldots, y_c\}$. $\hat{E}$ can be termed approximation error in data analysis problems, distortion or quantization error in signal processing contexts, energy in physical analogies, risk in decision-theoretic and statistical learning frameworks.

Miyamoto and Mukaidono [18] show that these algorithms are obtained by adding to the basic cost $\hat{E}$ in (2) either regularization terms or the maximum-entropy term

$$-H = \sum_{j=1}^{c} \sum_{k=1}^{n} u_{jk} \log u_{jk} \tag{3}$$

which represents the (negative) entropy of the clustering defined by $Y, U$. We propose an alternative perspective to interpret these techniques and to unify them with the possibilistic approach.

In clustering problems the focus is commonly placed on the analysis of data and clusters themselves, rather than on minimization of a global error criterion. We are often more interested in characterizing (hopefully significant) groups of data than in representing the details of the data with a faithful approximation. As an example, *model-based* clustering approaches focus on cluster modeling rather than performance optimization, and the cluster identification technique called *Alternating Cluster Estimation* [19] does not even assume the existence of a cost function. The key observation is that clustering is typically an exloratory phase of data analysis, and more accurate statistical testing should be placed in a subsequent phase.

Therefore we will introduce a formalism to provide an alternative, unified perspective on these clustering algorithms, focused on the memberships $u_{jk}$ rather than on the cost function. We will show that, apart from the possible addition of an entropic term, these algorithms are characterized by specific *feasible regions* for the membership values.

## 2.2 A unifying formalism

A CM clustering problem is defined by fixing the pair $\{J, \psi\}$, where:

- $J$ is the cost function
- $\psi$ is the constraint on the set of cluster memberships, such that

$$\psi(u_{1k}, \ldots, u_{ck}) = 0 \quad \forall k \in \{1, n\}$$

All the CM algorithms considered define either:

$$J = \hat{E} \tag{4}$$

or:

$$J = \hat{E} - H. \tag{5}$$

Table 1: The CM family of clustering algorithms

| | $J$ | $\psi$ | $v_{jk}$ | $Z_k$ | *Notes* |
|---|---|---|---|---|---|
| **DA** | $\hat{E} - H$ | $\sum_{j=1}^{c} u_{jk} - 1$ | $e^{-d_{jk}/\beta}$ | $\sum_{j=1}^{c} v_{jk}$ | $\beta \in \mathbb{R}$, $\beta > 0$ is the inverse temperature parameter to be increased during the "annealing" process. |
| **PCM-II** | $\hat{E} - H$ | $0$ | $e^{-d_{jk}/\beta_j}$ | $1$ | $\beta_j \in \mathbb{R}$, $\beta_j > 0$ are cluster width parameters to be selected a priori before optimization. |
| **FCM** | $\hat{E}$ | $\sum_{j=1}^{c} u_{jk}^{1/m} - 1$ | $1/d_{jk}$ | $\left(\sum_{j=1}^{c} v_{jk}^{1/(m-1)}\right)^{m-1}$ | $m \in \mathbb{R}$, $m > 1$ is the fuzzification parameter. |
| **HCM** | $\hat{E}$ | $\sum_{j=1}^{c} u_{jk} - 1$ | *See note* | *See note* | $v_{jk}$ and $Z_k$ can be written as for FCM, but their values have to be computed in the limit for $m \to 1$. |

Moreover, all the CM algorithms considered require that $u_{jk} \in [0,1]\ \forall j \in \{1,c\}$, $\forall k \in \{1,n\}$ (normality condition).

Let $v_{jk}$ be the solution of a CM problem with constraint $\psi$ removed (formally this can be implemented with $\psi \equiv 0$). We propose to call $v_{jk}$ the *free membership* of pattern $x_k$ in cluster $\omega_j$.

As a consequence of these definitions, for all the CM algorithms considered the cluster centroids $Y$ are computed as:

$$y_j = \frac{\sum_{k=1}^{n} u_{jk} x_k}{\sum_{k=1}^{n} u_{jk}} \tag{6}$$

which characterizes the *c*-Means principle and therefore the CM family. The memberships are computed as:

$$u_{jk} = \frac{v_{jk}}{Z_k}, \tag{7}$$

where $Z_k$ is the (generalized) partition function.

## 2.3 Review of the CM family

With the above set of definitions, the CM algorithms of interest are compactly described as in Table 1.

All algorithms are fuzzy techniques, since they adopt the concept of "partial membership" in a set. HCM itself can be cast without imposing the constraint of binary memberships. The relationships among these algorithms are clear from the table.

A method to allow for non-extreme solutions is the maximum entropy criterion, which is implemented in the DA and PCM-II algorithms. They are related by the use

of the entropic term $-H$, implying a parameter $\beta_j$. This parameter is different for each cluster and fixed in PCM-II, while it is constant for all clusters and varying with the algorithm progress in DA. However, these differences are not of a fundamental nature.

In the optimization perspective, the parameters $\beta_j$ arise from the Lagrange multiplier related to the entropic term in the Lagrangian. From the standpoint of the cluster model, they are related to cluster width. In PCM-II this term is even more crucial, since membership values are not constrained ($\psi \equiv 0$). Membership values are thus allowed to be simultaneously all zero and a means of biasing the solution toward nontrivial values is necessary.

The entropic term in the cost gives rise to free memberships having the functional form

$$v_{jk} = e^{-d_{jk}/\beta_j}, \tag{8}$$

which characterizes both DA and PCM-II.

An alternative way to obtain non-extreme solutions is having a nonlinear, convex Lagrangian by introducing nonlinear constraints. The memberships of the standard FCM formulation are equivalent to our $u_{jk}^{1/m}$, rather than $u_{jk}$. Apart from this constant transformation, our alternative formulation is equivalent and shows that the FCM problem optimizes the same cost function as HCM, but its feasible region is nonlinear ($\psi$ is nonlinear). This allows non-extreme solutions by acting on the membership model.

HCM can be cast as a linear programming problem, or a special case of the FCM problem, therefore its solutions are found on the border of the feasibility region defined by $\psi$. This means that the resulting memberships have extreme values within $[0,1]$: either 0 or 1. This behavior, whereby the "hard" nature of this algorithm is inherent and not a result of additional constraints, is easy to understand in the optimization perspective, and is due to the absence of nonlinear components in the Lagrangian.

# 3 The graded possibilistic model

## 3.1 The concept of graded possibility

The classic membership model (either hard or fuzzy) implements the concept of partitioning a set into disjoint subsets. This is done through the so-called "probabilistic constraint" by setting $\psi(u_{1k},\ldots,u_{ck}) = \sum_{j=1}^{c} u_{jk} - 1$. Each membership is therefore formally equivalent to the probability that an experimental outcome coincides with one of $c$ mutually exclusive events.

The possibilistic approach implies instead that each membership is formally equivalent to the probability that an experimental outcome coincides with one of $c$ mutually *independent* events. This is due to the complete absence of a constraint on the set of membership values ($\psi \equiv 0$).

However, it is possible (and in practice it is frequent) that pairs of events are not mutually independent, but are not completely mutually exclusive either. Instead, events can provide *partial information* about other events. Of course, this is a problem-dependent situation and accounting for it may or may not be appropriate.

An interesting case of partial information, in the context of the present research, is the concept of *graded possibility*. The standard possibilistic approach to clustering implies that all membership values are independent. In contrast, the graded possibilistic model assumes that, when one of the $c$ membership values is fixed, the other $c-1$ values are constrained into a given interval contained in $[0,1]$.

Clearly, this situation includes the possibilistic model, and also encompasses the standard ("probabilistic") approach.

An example of such graded possibility is given by a glass and by the fuzzy concepts of "full" and "empty". If the glass is full or almost full, its membership to the concept "empty" should clearly be around zero, and similarly for the empty or almost empty case. However, if the glass is half filled, it is much more difficult to assess the membership in the concept "empty" with similar confidence. The profile of the membership functions in this case should be decided according to further considerations.

In summary, in these intermediate cases the membership function should not be constrained by the cost function, but should be arbitrary to a certain degree.

## 3.2 Modeling graded possibility

A class of constraints $\psi$, which includes the probabilistic and the possibilistic cases, can be expressed by the following unified formulation:

$$\psi = \sum_{j=1}^{c} u_{jk}^{[\xi]} - 1, \tag{9}$$

where $[\xi]$ is an interval variable representing an arbitrary real number included in the range $[\underline{\xi}, \overline{\xi}]$. This interval equality should be interpreted as follows: there must exist a scalar exponent $\xi^* \in [\underline{\xi}, \overline{\xi}]$ such that the equality $\psi = 0$ holds.

This constraint enforces both the normality condition and the required probabilistic or possibilistic constraints; in addition, for nontrivial finite intervals $[\xi]$, it implements the required graded possibilistic condition.

The constraint presented above can be implemented in various ways. A particular implementation is as follows: the extrema of the interval are written as a function of a running parameter $\alpha$, where

$$\underline{\xi} = \alpha \qquad \overline{\xi} = \frac{1}{\alpha} \tag{10}$$

and

$$\alpha \in [0,1] \tag{11}$$

This formulation includes as the two extreme cases:

- The "probabilistic" assumption:

$$\alpha = 1$$

$$[\xi] = [1,1] = 1$$

$$\sum_{j=1}^{c} u_{jk} = 1$$

- The "possibilistic" assumption:

$$\alpha = 0$$

$$[\xi] = [0,\infty]$$

$$\sum_{j=1}^{c} u_{jk}^{0} \geq 1 \qquad \sum_{j=1}^{c} u_{jk}^{\infty} \leq 1$$

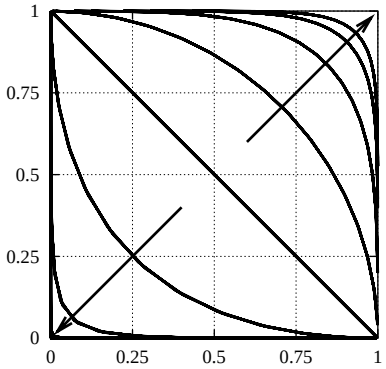


Figure 1: Bounds of the feasible region for $u_{jk}$ for different values of $\alpha$ (decreasing from 1 to 0 along the direction of the arrows)

The latter case can be better understood as the limit of the process of bringing $\alpha \to 0$. The interval exponent $[\xi]$ expands, so that the actual value can be any arbitrary number between $\alpha$ and $1/\alpha$. Therefore, each equation containing an interval is equivalent to a set of two inequalities:

$$\sum_{j=1}^{c} u_{jk}^{\alpha} \geq 1 \qquad \sum_{j=1}^{c} u_{jk}^{1/\alpha} \leq 1.$$

This is graphically depicted in Figure 1, where the bounds of the feasible regions are plotted, for $c = 2$, for values of $\alpha$ which decrease in the direction of the arrows.

In the first limit case, the feasible values for $u_{jk}$ must lie on a one-dimensional set (a line segment). In the second limit case, the feasible values for $u_{jk}$ are in the unity square, a two-dimensional set. In intermediate cases, the feasible values are on two-dimensional sets which however do not fill the whole square, but are limited to an eye-shaped area around the line segment.

Another implementation of the interval constraint is used in the outlier rejection application as explained in Subsection 5.4. In this case the upper extremum of $[\xi]$ is fixed to 1 and the lower extremum is $\alpha$.

# 4 Sample algorithm

In this section we outline a basic example of graded possibilistic clustering algorithm. This is an application of the ideas in the previous section. However, it is possible to apply many variations to this algorithm, so that appropriate properties can be obtained. Some of these variations will be presented and demonstrated in the experimental section.

For the proposed algorithm implementations, the free membership function has been selected as in the DA and PCM-II algorithms:

$$v_{jk} = e^{-d_{jk}/\beta_j}. \tag{12}$$

The generalized partition function can be defined as follows:

$$Z_k = \sum_{j=1}^{c} v_{jk}^{\zeta} \tag{13}$$

where:

$$\begin{aligned} \zeta = 1/\alpha \qquad & \text{if} \qquad \textstyle\sum_{j=1}^{c} v_{jk}^{1/\alpha} > 1 \\ \zeta = \alpha \qquad & \text{if} \qquad \textstyle\sum_{j=1}^{c} v_{jk}^{\alpha} < 1 \\ \zeta = 1 \qquad & \text{else.} \end{aligned}$$

These definitions ensure that, for $\alpha = 1$, the algorithm reduces to standard DA, whereas in the limit case for $\alpha = 0$, the algorithm is equivalent to PCM-II.

In both cases, the required value for the $\beta_j$ can be assessed from previous experiments, possibly in an independent way for each cluster (as done in PCM), or gradually lowered in an iterated application of the algorithm (as done in DA).

---

**Algorithm: Graded possibilistic clustering**
select c
select alphastep $\in \mathbb{R}$
randomly initialize $y_j$
for $\alpha = 1$ downto 0 by alphastep do
**begin**
    compute $v_{jk}$ using (12)
    compute $Z_k$ using (13)
    compute $u_{jk} = v_{jk}/Z_k$
    **if** stopping criterion satisfied **then stop**
    **else** compute $y_j$ using (6)

**end**

---

# 5 Demonstrations and applications

## 5.1 Experimental demonstrations

The experimental demonstrations presented here aim at highlighting the different properties attainable through appropriate implementation of a suitable graded possibilistic model. These include: demonstration of the concept of graded possibility, use of a-priori knowledge, outlier rejection.

## 5.2 Demonstration of the Graded Possibilistic approach

To show the properties of graded possibilistic clustering we use the toy training set shown in Figure 2. It is a simple, two-dimensional data set composed of 2 Gaussian-distributed clusters (50 points each), with centers indicated by the larger, black squares. Centers are located at (.7,.7) for cluster 1 and (.3,.3) for cluster 2. All data lie in the unit square.

We run the graded possibilistic clustering algorithm in 10 steps, with $\overline{\xi} = 1/\alpha$ and $\underline{\xi} = \alpha$ as in the sample algorithm of Section 4, and $\alpha$ decreasing from 1 to 0. We analyze the resulting memberships for different settings of the constraints.

We focus on memberships of three representative points. Point #9 in the data set is located at (.3,.3), i.e., it coincides with one cluster center. Point #10 is at (.53,.51), half-way from each center. Point #67 is at (.84,.34), quite far from both centers.

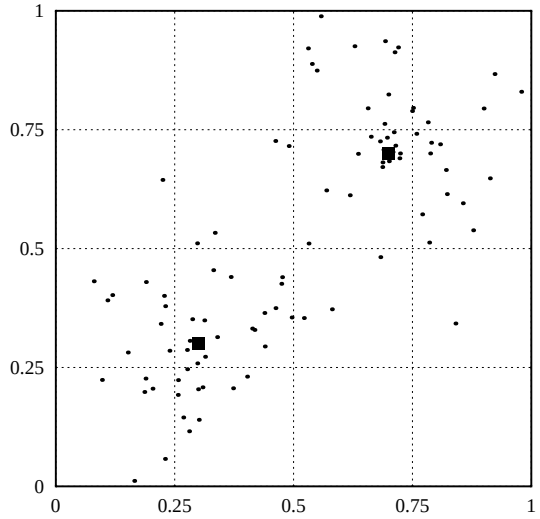


Figure 2: Toy problem to evaluate the behaviour of the algorithm.

Figure 3 shows the membership of each of these three data points in cluster 1 (solid line) and in cluster 2 (dashed line) for various steps of the clustering algorithms, corresponding to decreasing values of $\alpha$ from 1 to 0.

Point #9 is clearly attributed to cluster 2. Its distance is so small that its membership are "stuck" at 0 (for cluster 1) and 1 (for cluster 2), repectively.

Point #10 should be attributed to both clusters with approximately the same membership value. However, since it is on the separating boundary, it is far from any cluster, so that, when $\alpha$ decreases and the model becomes more possibilistic, the memberships also decrease from .6 and .4 to .15 and .25 (respectively for clusters 1 and 2).

Point #67 is clearly an outlier. However, in the first step of the algorithm, it is classified as belonging in cluster 1 with high degree (almost 1). In the further steps, with the transition to the possibilistic model, the values are reduced to about 0 and .07, respectively.

Figure 4 shows a 3-d plot of membership values to each cluster for every point in the plane, for the two extreme values $\alpha = 1$ and $\alpha = 0$. The different shape of the memberships, especially for points far from the separating line, is apparent. However Figure 5 clarifies further this result, by detailing the profile (intersection of the membership functions with the diagonal plane $x_1 = x_2$) for three values of $\alpha$, two extreme and one intermediate (1, .5, and 0).

A similar analysis is presented in Figure 6. This experiment is performed on the usual Iris dataset [20] obtained from the UCI Machine Learning Repository [21]. (The Iris problem is a 4-dimensional, 150-pattern data set with 3 classes represented by 50 patterns each.)

Here the profiles of memberships are plotted for 2 of the 3 clusters and for 2 of the 4 input dimensions, so that two-dimensional analysis is again possible. The figure shows membership profiles for $\alpha = 1.0$, $\alpha = 0.5$, and $\alpha = 0.0$. It is possible to tune the desired trade-off between the possibilistic clustering and the partitioning behavior, by deciding to what extent the algorithm should be forced to make a decision on data points on the decision border or on the exterior part of the data distribution.

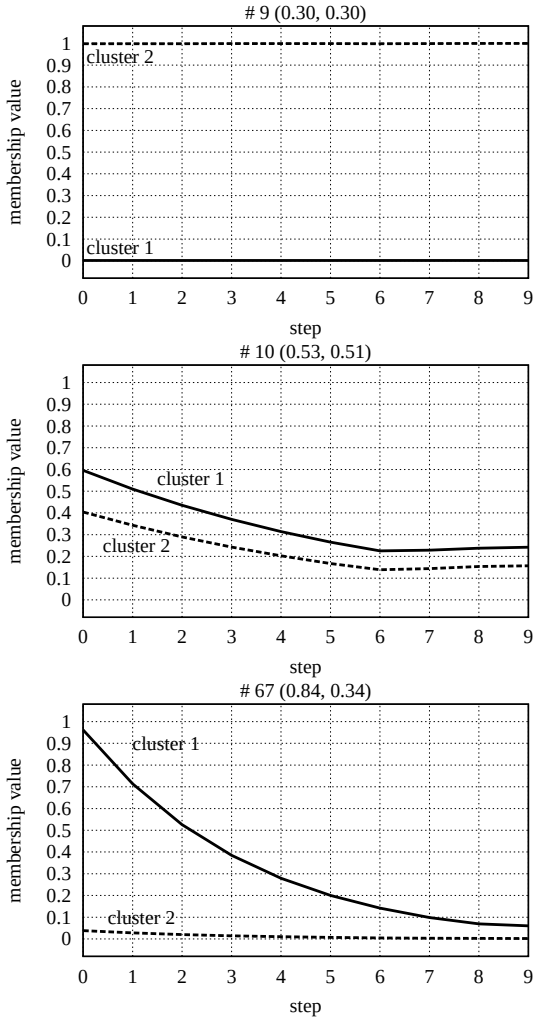


Figure 3: Memberships of points #9, #10, and #67 in each cluster.

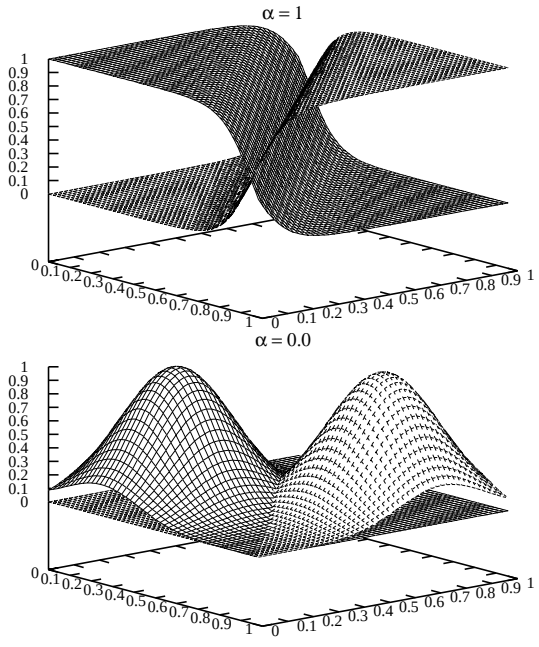


Figure 4: Plot of memberships for $\alpha = 1$ (above) and for $\alpha = 0$ (below).

## 5.3 Using a-priori knowledge

This experimental demonstration illustrates the use of a suitable value for $\alpha$ to improve the results with respect to the extreme cases (probabilistic and pure possibilistic). In this case the optimum value is inferred from the results but not used (for lack of a test set); in real applications it can be estimated on the training set prior to use on new data.

We show sample results from the following unsupervised classification experiment. First, the graded possibilistic clustering procedure was applied to the Iris data set. Only one cluster center per class was used ($c = 3$). Then the cluster memberships were "defuzzified" by setting the maximum to 1 and the other two to 0. Subsequently, the hard memberships were used to associate class labels to each cluster (by majority). Finally, the classification error was evaluated.

The classification error percentages as a function of $\alpha$ are shown in Figure 7. Although these are only a sample of the results, which may have been different in other runs, the profile of the graph was qualitatively almost constant in all trials. The best classification performance with $c = 3$ was 7.3% error, which means 11 mistaken points.

In all experiments this value was obtained for *intermediate* values of $\alpha$, between 0.3 and 0.7. In other words, the graded possibilistic model was able to catch the true distributions of data better than either the probabilistic or the possibilistic approaches. The pure possibilistic case gave rise (as in the results presented in the figure) to a percentage of cases with overlapping cluster centers, in accordance with previous experimental observations [14].

The error levels can be categorized into three classes. The first is around the optimum (11 or 12 or occasionally 13 wrong classifications). The second, sometimes observed in the pure possibilistic case, is the case of overlapping clusters, with about

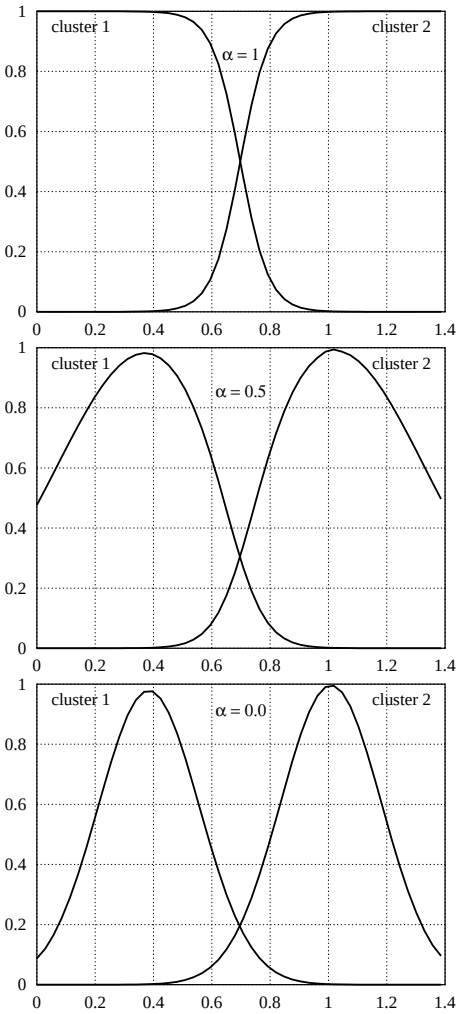


Figure 5: Projection of the same memberships as in Figure 4 on the diagonal of the square area. Above: $\alpha = 1$; middle: $\alpha = 0.5$; below: $\alpha = 0$.

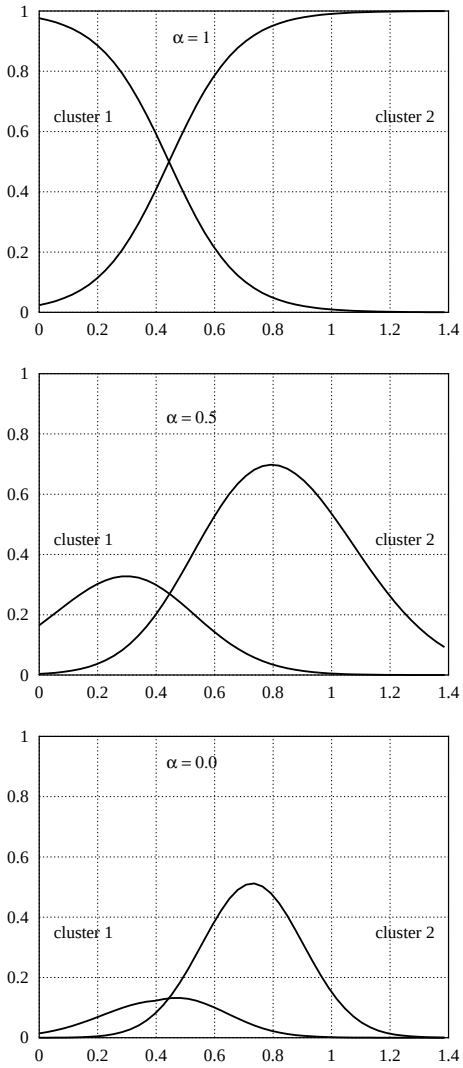


Figure 6: Two-dimensional plot of memberships for $\alpha = 1.0$ (above), for $\alpha = 0.5$ (middle), and for $\alpha = 0.0$ (below) for the Iris dataset (same analysis as in Figure 5).

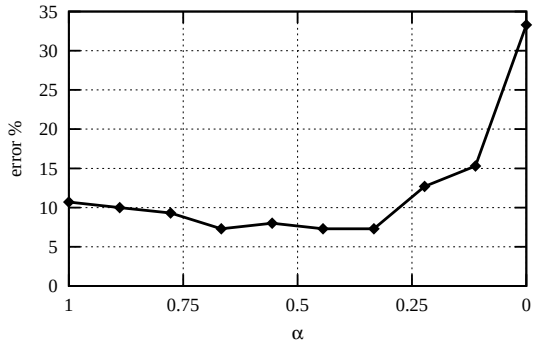


Figure 7: Error percentage plot for the unsupervised Iris classification.

33% error rate. The third, above 10%, is typical of the probabilistic case, where competition among clusters does not allow optimal placement of the cluster centers.

## 5.4 Outlier rejection

To implement the outlier rejection functionality, the feasible region should be made asymmetric:

$$\sum_{j=1}^{c} u_{jk} \leq 1 \qquad \text{and} \qquad \sum_{j=1}^{c} u_{jk}^{\alpha} \geq 1. \tag{14}$$

This ensures that the clustering model is as follows. When there is competition among the clusters, i.e., many memberships tend to be close to 1, the membership values are normalized to sum to 1 (first constraint). When memberships are all low, there is no clear attribution to any cluster, so they are free to take on low values (second constraint).

Rejection is then done by selecting a membership threshold, which could be different for each cluster and obtained for instance by statistical analysis of the individual cluster distribution. Patterns for which no membership in any cluster exceeds the appropriate threshold are rejected.

However, even without explicit outlier analysis, the algorithm becomes very robust with respect to the presence of outliers.

The experiments involve a set of three Gaussian clusters, plus a very wide background data distribution (see Figure 8). Data are in the unit square; there are 600 data points of which a given percentage is clustered in 3 Gaussian clusters (again centers are marked with black squares), and the others are spread in the background, with higher density in the proximity of the unit square corners and perimeter. The proportion of outliers to clustered points was varied from 10% to 90%.

From the experimental results in Figure 9, obtained with an outlier-to-clustered ratio of 90%-10%, it is possible to compare the behavior of the graded possibilistic model with the behavior of standard "probabilistic" clustering. Centers found with the proposed model are clearly much closer to true cluster centers than those found with the "probabilistic" model (the residual error being due mostly to the random sampling, so that the barycenter of the data points in a given cluster does not coincide with the true cluster center). By inspection of the membership values, we have verified that this

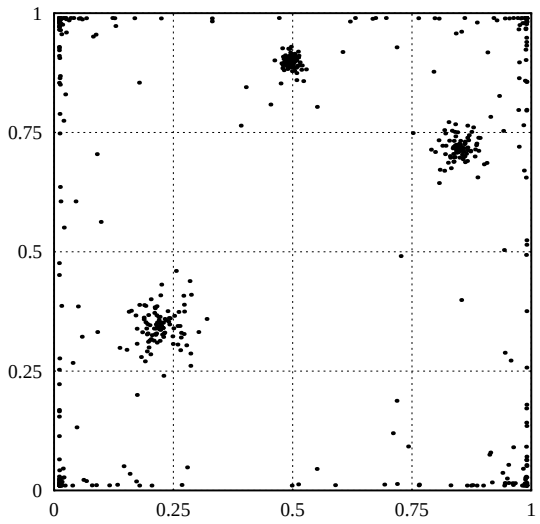


Figure 8: Dataset for the outlier rejection demonstration.

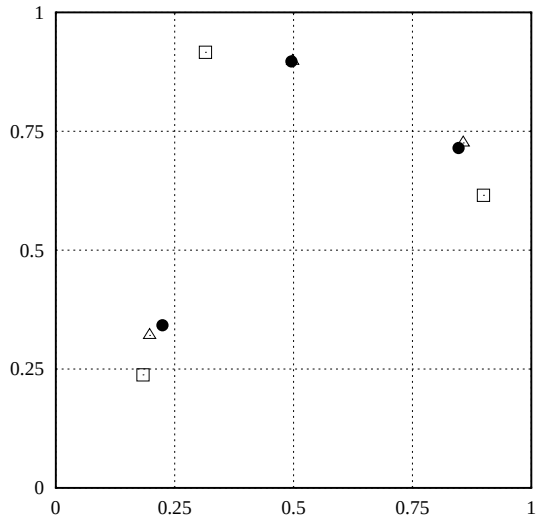


Figure 9: Results for the outlier rejection demonstration. Black circles: true cluster centers; triangles: centers found with $\alpha = 0$ (maximum rejection); squares: centers found with $\alpha = 1$ (no rejection).

is not a true possibilistic case: no two memberships ever approach 1 simultaneously. Therefore, either a pattern is rejected, or it is uniquely labeled.

# 6 Conclusion

The concept of graded possibility applied to clustering, which has been presented in this paper, allows the implementation of specific properties in the *c*-Means family of clustering techniques. With appropriate selection of some parameters, an entropy-constrained version of *c*-Means can implement partitioning, mode-seeking, constraining by prior knowledge, outlier rejection.

We expect that this flexible behavior will be proficently exploited in the currently active research areas of Web content analysis, document data mining, preprocessing of space mission data, DNA microarray data analysis, since they allow the clustering algorithm to be tailored on the specific demands of a given applicative problem.

# Appendix

We prove here that the proposed formulation of the FCM algorithm is equivalent to the standard formulation, in the sense that solutions of the necessary conditions for extrema of the Lagrangian for the two methods are coincident apart from a constant (fixed) transformation.

Let $\hat{\mathcal{L}}$ be the Lagrangian of the standard Fuzzy *c*-Means problem:

$$\hat{\mathcal{L}} = \sum_{k=1}^{n} \sum_{j=1}^{c} \hat{u}_{jk}^{m} d_{jk} + \sum_{k=1}^{n} \lambda_k \left( \sum_{j=1}^{c} \hat{u}_{jk} - 1 \right) \tag{15}$$

(where $\lambda_k$ is the Lagrange multiplier for the probabilistic constraint on the set of memberships of the *k*-th data point).

Let now $\mathcal{L}$ be the Lagrangian of the proposed Fuzzy *c*-Means problem formulation. In this case the error term of the Lagrangian is $\hat{E}$ as in HCM, and the feasible region for the memberships is modified. Therefore:

$$\mathcal{L} = \sum_{k=1}^{n} \sum_{j=1}^{c} u_{jk} d_{jk} + \sum_{k=1}^{n} \lambda_k \left( \sum_{j=1}^{c} u_{jk}^{1/m} - 1 \right) \tag{16}$$

The proof of the following proposition is straightforward, but for completeness we provide it with some algebraic details.

**Theorem 1** *Solving the FCM problem in the standard formulation is equivalent to solving the FCM problem in the proposed alternative formulation with the substitution* $\hat{u}_{jk}^{m} = u_{jk}$

*Proof:* It is well known that, in the standard problem setting for FCM, cluster centers and memberships are given respectively by

$$y_j = \sum_{k=1}^{n} \frac{u_{jk}^{m} x_k}{u_{jk}^{m}} \quad \text{and} \quad u_{jk} = \left[ \sum_{l=1}^{c} \left( \frac{d_{jk}}{d_{jl}} \right)^{1/(m-1)} \right]^{-1} \tag{17}$$

in the hypothesis of Euclidean distance measure (which we assume for simplicity of exposition, but in this context this does not affect the results in any significant way).

The solution of $\nabla\mathcal{L} = 0$ is worked out as follows. From the condition $\frac{\partial\mathcal{L}}{\partial y_j} = 0$ we have:

$$\frac{\partial\mathcal{L}}{\partial y_j} = u_{jk}\frac{\partial d_{jk}}{\partial y_j} = 2u_{jk}(y_j - x_k) = 0 \tag{18}$$

so that

$$y_j = \frac{\sum_{k=1}^{n} u_{jk}y_j}{\sum_{k=1}^{n} u_{jk}}, \tag{19}$$

that is, the HCM centroid formula.

From the condition $\frac{\partial\mathcal{L}}{\partial u_{jk}} = 0$ we have:

$$\frac{\partial\mathcal{L}}{\partial u_{jk}} = d_{jk} + \lambda_k \frac{1}{m} u_{jk}^{\frac{1-m}{m}} \tag{20}$$

$$u_{jk} = \left(\frac{\lambda_k/m}{d_{jk}}\right)^{m/(m-1)} \tag{21}$$

From the condition $\frac{\partial\mathcal{L}}{\partial\lambda_k} = 0$ we have:

$$\sum_{j=1}^{c} u_{jk}^{1/m} = \sum_{j=1}^{c} \left(\frac{\lambda_k/m}{d_{jk}}\right)^{1/(m-1)} = 1, \tag{22}$$

therefore

$$\frac{\lambda_k}{m} = \left[\sum_{j=1}^{c} \left(\frac{1}{d_{jk}}\right)^{1/(m-1)}\right]^{1-m} \tag{23}$$

Substituting (23) into (21) yields

$$u_{jk} = \frac{(\lambda_k/m)^{m/(m-1)}}{d_{jk}^{m/(m-1)}} =$$

$$= \left[\sum_{l=1}^{c} \left(\frac{d_{lk}}{d_{jk}}\right)^{1/(m-1)}\right]^{-m} \tag{24}$$

The expressions in (19) and (24) differ from those in (17) only by a constant exponent $m$ in the memberships (that is, $\hat{u}_{jk}^{m} = u_{jk}$), which proves the assertion. ■

# References


[1] Enrique H. Ruspini, "A new approach to clustering", *Information and Control*, vol. 15, no. 1, pp. 22–32, 1969.

[2] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters", *Journal of Cybernetics*, vol. 3, pp. 32–57, 1974.

[3] James C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Plenum, New York, 1981.

[4] G.H. Ball and D.J. Hall, "ISODATA, an iterative method of multivariate analysis and pattern classification", *Behavioral Science*, vol. 12, pp. 153–155, 1967.

[5] Kenneth Rose, Eitan Gurewitz, and Geoffrey Fox, "A deterministic annealing approach to clustering", *Pattern Recognition Letters*, vol. 11, pp. 589–594, 1990.

[6] Kenneth Rose, Eitan Gurewitz, and Geoffrey Fox, "Statistical mechanics and phase transitions in clustering", *Physical Review Letters*, vol. 65, pp. 945–948, 1990.

[7] C.K. Chow, "An optimum character recognition system using decision function", *IRE Transactions on Electronic Computers*, vol. 6, pp. 247–254, 1957.

[8] C.K. Chow, "An optimum recognition error and reject tradeoff", *IEEE Transactions on Information Theory*, vol. 16, pp. 41–46, 1970.

[9] Richard O. Duda and Peter E. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York (USA), 1973.

[10] Hisao Ishibuchi and Manabu Nii, "Neural networks for soft decision making", *Fuzzy Sets and Systems*, vol. 115, no. 1, pp. 121–140, October 2000.

[11] Gian Paolo Drago and Sandro Ridella, "Possibility and necessity pattern classification using an interval arithmetic perceptron", *Neural Computing and Applications*, vol. 8, no. 1, pp. 40–52, 1999.

[12] Sandro Ridella, Stefano Rovetta, and Rodolfo Zunino, "K-winner machines for pattern classification", *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 371–385, March 2001.

[13] Raghu Krishnapuram and James M. Keller, "A possibilistic approach to clustering", *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, pp. 98–110, May 1993.

[14] Raghu Krishnapuram and James M. Keller, "The possibilistic *C*-Means algorithm: insights and recommendations", *IEEE Transactions on Fuzzy Systems*, vol. 4, no. 3, pp. 385–393, August 1996.

[15] Antonio Flores-Sintas, José M. Cadenas, and Fernando Martin, "Local geometrical properties application to fuzzy clustering", *Fuzzy Sets and Systems*, vol. 100, pp. 245–256, 1998.

[16] E. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a covariance matrix", in *Proc. IEEE Conf. Decision Contr.*, San Diego, USA, 1979, pp. 761–766.

[17] James C. Bezdek, "Detection and characterization of cluster substructure, II: fuzzy *c*-varieties and convex combinations thereof", *SIAM J. Appl. Math.*, vol. 40, no. 2, pp. 358–372, April 1981.

[18] Sadaki Miyamoto and Masao Mukaidono, "Fuzzy C-Means as a regularization and maximum entropy approach", in *Proceedings of the Seventh IFSA World Congress, Prague*, 1997, pp. 86–91.

[19] Thomas A. Runkler and James C. Bezdek, "Alternating cluster estimation: a new tool for clustering and function approximation", *IEEE Transactions on Fuzzy Systems*, vol. 7, no. 4, pp. 377–393, August 1999.

[20] Ronald A. Fisher, "The use of multiple measurements in taxonomic problems", *Annual Eugenics*, vol. 7, part II, pp. 179–188, 1936.

[21] C.L. Blake and C.J. Merz, "UCI repository of machine learning databases", 1998, URL: http://www.ics.uci.edu/∼mlearn/MLRepository.html.