The Graded Possibilistic Clustering Model

Francesco Masulli[1,2] and Stefano Rovetta[1,3]

(1) INFM, the National Institute for the Physics of Matter, Italy
(2) Department of Computer Science, University of Pisa, Via F. Buonarroti 2 - 56127 Pisa Italy
(3) Department of Computer and Information Sciences, University of Genoa, Via Dodecaneso 35 - 16146 Genova Italy

E-mail: masulli@di.unipi.it - rovetta@disi.unige.it

Abstract— **This paper presents the *graded possibilistic model*. After reviewing some clustering algorithms derived from *c*-Means, we provide a unified perspective on these clustering algorithms, focused on the memberships rather than on the cost function. Then the concept of graded possibility is introduced. This is a *partially* possibilistc version of the fuzzy clustering model, as compared to Krishnapuram and Keller's possibilistic clustering. We outline a basic graded possibilistic clustering algorithm and highlight the different properties attainable by means of experimental demonstrations.**

I. Introduction

Clustering algorithms underlie a whole family of neural network techniques, including for instance kohonen's Self Organizing Maps and Learning Vector Quantization. In this framework, clustering problems are 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 $d(\cdot,\cdot)$.

A very widely used clustering method is the *Fuzzy c-Means* [1] (FCM) algorithm, a "fuzzy relative" to the simple *c*-Means technique [2]. FCM defines the ω_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 [3] 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 $-H$. 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. With this technique it is possible to alleviate the false minima problem.

In decision-making and classification applications, algorithms should feature several desirable properties in addition to the basic decision function. For instance, it is often required that in certain configurations a decision is not made (*pattern rejection*), typically in the presence of outliers. This problem is very well-known and well studied (e.g. see [4][5][6]), and is tackled in a convenient way within the framework of soft-computing, fuzzy, and neural approaches [7][8][9].

However, the clustering problem as stated above implies that the outlier rejection property cannot be achieved. This is because the membership values u_{jk} are constrained to sum to 1 (the *probabilistic* model). 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* [10][11], 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 analyzing the membership interactions on a local basis, as opposed to the global effects induced by the probabilistic model, is considered in [12].

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.

II. Unified view of some clustering algorithms

A. The c-Means family

We will now review some clustering algorithms derived from the basic *c*-Means: ("hard") *c*-Means (HCM) [2], entropy-constrained fuzzy clustering by Deterministic Annealing (DA) [3], Possibilistic *c*-Means with an entropic cost term (PCM-II) [11], Fuzzy *c*-Means (FCM) [13]. 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 ref. [14]). 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 ω_j and $Y = \{y_1,\ldots,y_c\}$. $\hat{E}$ can be termed approximation error, distortion or quantization error, energy, or risk, depending on the application and the nature of the problem.

Miyamoto and Mukaidono [15] 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 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.

0-7803-7898-9/03/$17.00 ©2003 IEEE

B. A unifying formalism

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

- J is the cost function
- ψ is the constraint on the set of cluster memberships, such that

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

All the CM algorithms considered define either $J=\hat{E}$ or $J=\hat{E}-H$. 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 without constraint ψ (formally this can be implemented with $\psi\equiv 0$). We propose to call v_{jk} the *free membership* of pattern x_k in cluster ω_j.

Therefore for all the CM algorithms considered the cluster centroids Y are computed as:

$$y_j=\frac{\sum_{k=1}^{n}u_{jk}x_k}{\sum_{k=1}^{n}u_{jk}} \tag{4}$$

characterizing the c-Means principle and therefore the CM family. Memberships are computed as:

$$u_{jk}=\frac{v_{jk}}{Z_k}, \tag{5}$$

where Z_k is the (generalized) partition function.

These CM algorithms are summarized in Table I.

C. Review of the CM family

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 β_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.

In the optimization perspective, the parameters β_j arise from the Lagrange multiplier related to the entropic term. They are related to cluster width. In PCM-II their role is crucial, since membership values are not constrained ($\psi\equiv 0$) and are thus allowed to be simultaneously all zero; 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

$$v_{jk}=e^{-d_{jk}/\beta_j}, \tag{6}$$

which characterizes both DA and PCM-II.

An alternative way to obtain non-extreme solutions is introducing nonlinear constraints. The memberships of FCM 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 (ψ is nonlinear). This allows non-extreme solutions by acting on the membership model.

III. The graded possibilistic model

A. 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 $\psi(u_{1k},\dots,u_{ck})=\sum_{j=1}^{c}u_{jk}-1$. 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 frequent that sets of events are neither mutually independent nor completely mutually exclusive either. Instead, events can be loosely related. Often this situation can be modeled by a statistical correlation.

Another interesting case of partial information 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.

B. Modeling graded possibility

A class of constraints ψ, including the probabilistic and the possibilistic cases, can be expressed as follows:

$$\psi=\sum_{j=1}^{c}u_{jk}^{[\xi]}-1, \tag{7}$$

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 exists 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. We suggest the following particular implementation which accounts for the probabilistic and possibilistic models as limit cases.

The extrema of the interval are written as a function of a running parameter α, where

$$\underline{\xi}=\alpha \qquad \overline{\xi}=\frac{1}{\alpha} \tag{8}$$

and

$$\alpha\in[0,1] \tag{9}$$

This formulation includes as the two extreme cases:

- The "probabilistic" assumption:

$$\alpha=1 \quad ; \quad [\xi]=[1,1]=1 \quad ; \quad \sum_{j=1}^{c}u_{jk}=1$$

TABLE I

THE CM FAMILY OF CLUSTERING ALGORITHMS

	J	ψ	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}/β_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 should be computed for $m \to 1$.

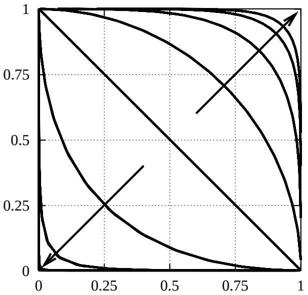


Fig. 1. Bounds of the feasible region for u_{jk} for different values of α (decreasing from 1 to 0 along the direction of the arrows)

- The "possibilistic" assumption:

$$\alpha = 0 \quad ; \quad [\xi] = [0, \infty] \quad ; \quad \sum_{j=1}^{c} u_{jk}^{0} \geq 1 \ , \ \sum_{j=1}^{c} u_{jk}^{\infty} \leq 1$$

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 depicted in Figure 1 ($c = 2$), where the bounds of the feasible regions are plotted for α decreasing in the direction of the arrows.

In the first limit case, the u_{jk} are constrained on a one-dimensional locus (a line segment). In the second limit case, the locus of the feasible values for u_{jk} is the unit square, which is two-dimensional. In intermediate cases, the loci of feasible values are two-dimensional, but they do not fill the whole square, being limited to eye-shaped areas (increasing with $\alpha \to 0$) around the line segment.

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

IV. 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. It is also possible to apply many variations to this algorithm to obtain specific properties.

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

$$v_{jk} = e^{-d_{jk}/\beta_j}. \tag{10}$$

The generalized partition function can be defined as follows:

$$Z_k = \sum_{j=1}^{c} v_{jk}^{\kappa} \tag{11}$$

where:

$$\begin{array}{lll} \kappa = 1/\alpha & \text{if} & \sum_{j=1}^{c} v_{jk}^{1/\alpha} > 1 \\ \kappa = \alpha & \text{if} & \sum_{j=1}^{c} v_{jk}^{\alpha} < 1 \\ \kappa = 1 & \text{else.} & \end{array}$$

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, β_j can be experimentally estimated (as in PCM) or iteratively "annealed" (as in DA).

```
Algorithm: Graded possibilistic clustering
select c $\in \mathbb{N}$, alphastep $\in \mathbb{R}$
randomly initialize $y_j$
for $\alpha = 1$ downto 0 by alphastep do
begin
  compute $v_{jk}$ using (10)
  compute $Z_k$ using (11)
  compute $u_{jk} = v_{jk}/Z_k$
  if stopping criterion satisfied then stop
  else compute $y_j$ using (4)
end
```

V. DEMONSTRATIONS AND APPLICATIONS

A. 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,

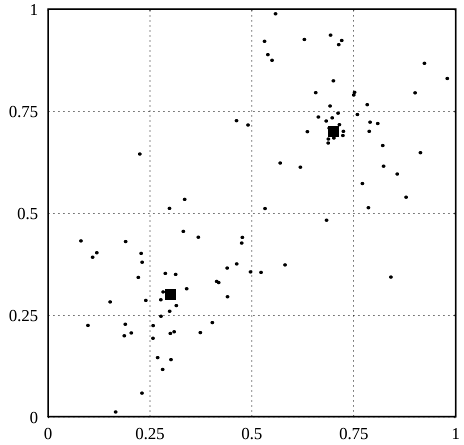


Fig. 2. Toy problem to evaluate the behaviour of the algorithm.

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 IV, and α 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 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 α 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), respectively.

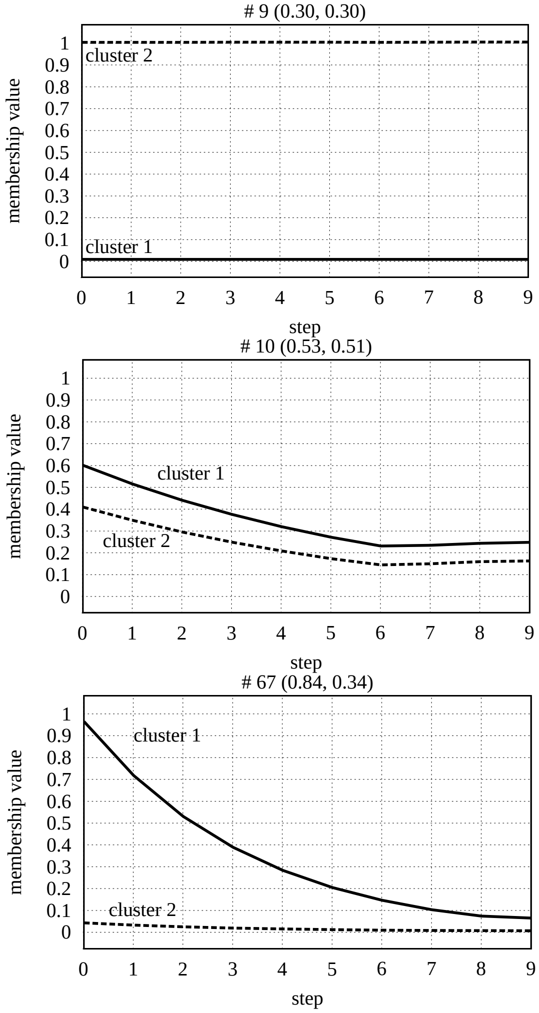
Point #10 should be attributed to both clusters with approximately the same membership value. However, since it is on the separating boundary, it is far from any cluster, so that, when α decreases and the model becomes more possibilistic, the memberships also decrease from .6 and .4 to .15 and .25 (respectively 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 the membership values along the line connecting the two cluster centers for three values of α, two extreme and one intermediate (1.0, 0.5, and 0.0).

A similar analysis is presented in Figure 5. However this experiment is performed on the usual Iris dataset obtained from the UCI Machine Learning Repository [16]. (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.


# 9 (0.30, 0.30)

# 10 (0.53, 0.51)

# 67 (0.84, 0.34)

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

B. Using a-priori knowledge

This experimental demonstration illustrates the use of a suitable value for α 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, in a semi-supervised setting.

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

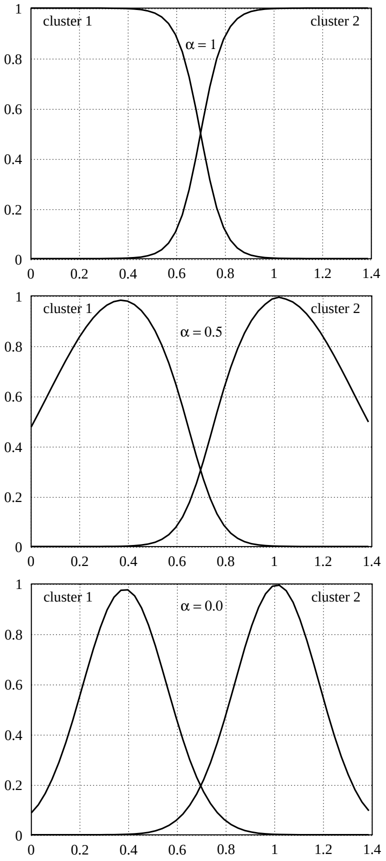


Fig. 4. Memberships along the line connecting the two cluster centers in the toy problem. Above: $\alpha = 1.0$; middle: $\alpha = 0.5$; below: $\alpha = 0.0$.

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 α are shown in Figure 6. 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 α, 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 [11]. We again note that in real applications α could be experimentally estimated on the training set.

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 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.

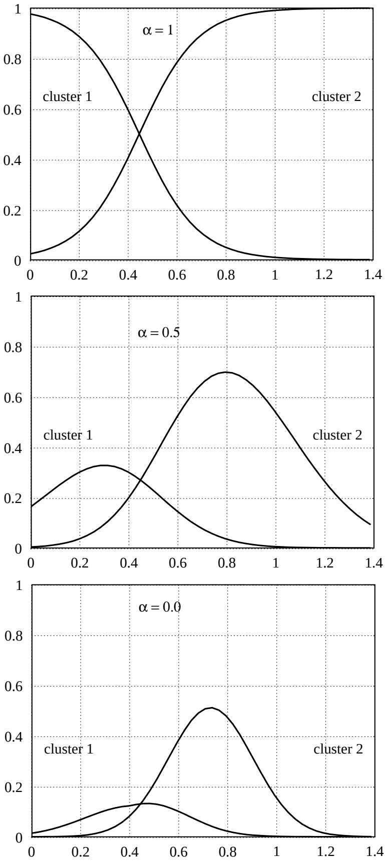


Fig. 5. Two-dimensional plot of memberships for $\alpha = 1.0$ (above), for $\alpha = 0.5$ (middle), and for $\alpha = 0.0$ (below) for the Iris dataset (same analysis as in Figure 4).

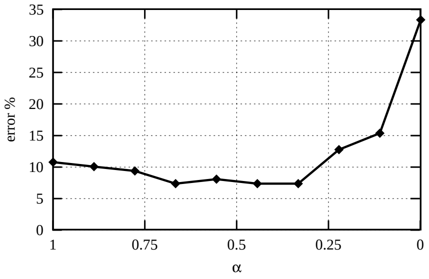


Fig. 6. Error percentage plot for the unsupervised Iris classification.

C. Outlier rejection

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

$$\sum_{j=1}^{c} u_{jk} \le 1 \qquad \text{and} \qquad \sum_{j=1}^{c} u_{jk}^{\alpha} \ge 1. \tag{12}$$

When there is competition among the clusters (many memberships approach 1) the membership values are normalized to sum to 1 by the 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 done by selecting a membership threshold, possibly different for each cluster. Patterns for which no membership in any cluster

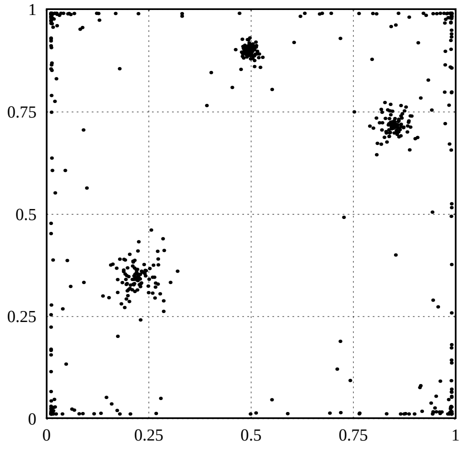


Fig. 7. Dataset for the outlier rejection demonstration.

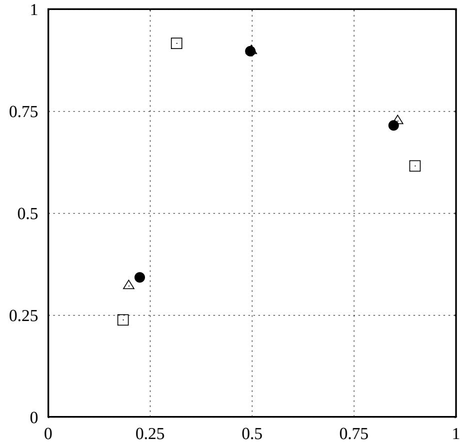


Fig. 8. Results for the outlier rejection demonstration. Black circles: true cluster centers; triangles: centers found with $\alpha = 0$ (maximum rejection); squares: centers found with $\alpha = 1$ (no rejection).

exceeds the appropriate threshold are rejected.

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 7). There are 600 data points, in the unit square, 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 8, 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 is not a true possibilistic case: no two memberships ever approach 1 simultaneously. Therefore, either a pattern is rejected, or it is uniquely labeled.

VI. 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. This flexible behavior can be exploited in several currently active research areas, often featuring a clustering step as an essential component.

References

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

[2] 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.

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

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

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

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Francesco Masulli and Stefano Rovetta, “Soft transition from probabilistic to possibilistic fuzzy clustering”, Tech. Rep. DISI-TR-03-02, Department of Computer and Information Sciences, University of Genoa, Italy, April 2002, URL: http://www.disi.unige.it/person/RovettaS/research/techrep/DISI-TR-02-03.ps.gz.

[15] 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.

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