Soft Transition from Probabilistic to Possibilistic Fuzzy Clustering

Francesco Masulli and Stefano Rovetta

Abstract

In the fuzzy clustering literature, two main types of membership are usually considered: a relative type, termed *probabilistic*, and an absolute or *possibilistic* type, indicating the strength of the attribution to any cluster independent from the rest. There are works addressing the unification of the two schemes. Here we focus on providing a model for the transition from one schema to the other, to exploit the dual information given by the two schemes, and to add flexibility for the interpretation of results. We apply an uncertainty model based on interval values to memberships in the clustering framework, obtaining a framework that we term *graded possibility*. We outline a basic example of graded possibilistic clustering algorithm and add some practical remarks about its implementation. The experimental demonstrations presented highlight the different properties attainable through appropriate implementation of a suitable graded possibilistic model. An interesting application is found in automated segmentation of diagnostic medical images, where the model provides an interactive visualization tool for this task.

Index Terms

Possibilistic clustering, Clustering methods, Fuzzy clustering, Fuzzy statistics and data analysis.

I. I

The problem of clustering [1], [2] is usually stated in the context of exploratory data analysis, where it is used as an aid to understand the structure of a given phenomenon represented by a collection of experimental data. In this framework, it is important that the clustering results are given in a form that helps understanding the properties of the phenomenon.

F. Masulli and S. Rovetta are with the Department of Computer and Information Sciences, University of Genova, Italy, and CNISM.

Corresponding author: Stefano Rovetta, DISI, University of Genova, Via Dodecaneso 35 – 16146 Genova, Italy. Fax: +39-10-353 6699. E-mail: rovetta@disi.unige.it

One of the most prominent clustering paradigms (as witnessed by the number of existing variations) is the principle adopted in the *c*-Means [3], [4] algorithm, also termed "central clustering". This paradigm implies partitioning a set of data vectors or patterns $X = \{\mathbf{x}_k\}$, $k \in \{1, \ldots, n\}$, $\mathbf{x}_k \in \mathbb{R}^n$ by attributing each data point $\mathbf{x}_k$ to a subset (cluster) $\omega_j \subset X$, $j \in \{1, \ldots, c\}$, defined by its *centroid* $\mathbf{y}_j \in \mathbb{R}^n$. This attribution is made on the basis of a given distance $d(\cdot, \cdot)$. The partition thus obtained is then generalizable to the whole data space by means of an implicit Voronoi tessellation $\Omega = \{\omega_1, \ldots, \omega_c\}$ with the centroids as its Voronoi sites.

Membership u_{jk} of a data point $\mathbf{x}_k$ to a given cluster ω_j is often modeled as a fuzzy value [5], [6], [7], with crisp clustering becoming a marginal special case [8] which can always be obtained by a winner-take-all operation. It is interesting to note that, although crisp or "hard" *c*-Means is very popular because of its speed, the fuzzy approach is both less prone to local minima and more informative as a technique for data analysis. The set of all centroids $\mathbf{y}_j$ and the set of all memberships to clusters for any given point constitute the description of the data under the structure suggested by our clustering algorithm.

In the literature two main types of membership are usually considered [9]: a relative type of membership, sometimes termed *probabilistic*, which for a given point indicates to what proportion it should be attributed to each cluster; and an absolute or *possibilistic* type, indicating the strength of the attribution to any cluster independent from the rest. There are works addressing the unification of the two schemes or exploitation of both schemes in a single method [10], [11], [12], [13], [14], especially in a robust clustering framework. Here instead we focus on providing a model for the *transition* from one schema to the other. Our main aim is to exploit the dual information given by the two schemes, and to add a degree of flexibility for the interpretation of results in an exploratory data analysis task. However, the proposed method features also some interesting properties from the robustness standpoint.

This paper is organized as follows: first we briefly review the "Maximum Entropy" and "Possibilistic" approaches to clustering (Section II), then we introduce the concept of graded possibility and its application to clustering (Section III). A possible implementation is discussed in Section IV, along with suggested variations. Section V proposes some applications in standard clustering problems, whereas Section VI introduces an application to interactive post-processing in diagnostic image segmentation. Conclusions are drawn in Section VII.

II. T "M E " "P "

The central clustering paradigm is implemented in several, diverse algorithms. Often these algorithms are based on objective function minimization [15], [14]. This helps in understanding the behavior and

features of each technique. However, there are also many examples of algorithms for which no objective function is defined (e.g. Alternating Cluster Estimation [16] or the Self Organizing Map [17]). In this section we will discuss two interesting algorithms, the Maximum Entropy clustering method by Rose et al. [18], [19], and the Possibilistic approach by Krishnapuram and Keller [20], [21], [22].

Both methods are based on the following definition of cluster centroids Y:

$$\mathbf{y}_j = \frac{\sum_{k=1}^{n} u_{jk}\mathbf{x}_k}{\sum_{k=1}^{n} u_{jk}}, \tag{1}$$

which is the same as in the c-Means methods, and is obtained independently on the definition of distance. However, since the memberships u_{jk} are computed in different ways, the resulting centroids are not the same.

In the Maximum Entropy approach, points and centroids are regarded as a formal physical system, and memberships of points are interpreted as probabilities of belonging to a given cluster. The centroids are found by fixing the energy of the system, and then minimizing the negative entropy of the partition $\Omega = \{\omega_1, \ldots, \omega_c\}$, constrained by the energy, so that it represents the most general (less imbalanced) partition.

Using Lagrange multipliers, the objective function is therefore expressed as $J'_{\mathrm{ME}} = \sum_{j=1}^{c}\sum_{k=1}^{n} u_{jk}\log u_{jk} + \beta\sum_{j=1}^{c}\sum_{k=1}^{n} u_{jk}d_{jk}$, where we use the shorthand notation $d_{jk} = d(\mathbf{y}_j, \mathbf{x}_k)$. We can divide by β and add the normality constraint $\sum_{j=1}^{c} u_{jk} = 1\ \forall k$, which is needed by the interpretation of u_{jk} as probabilities (see also [5]). The objective function can thus be rewritten as

$$J_{\mathrm{ME}} = \sum_{j=1}^{c}\sum_{k=1}^{n} u_{jk}d_{jk} + \frac{1}{\beta}\sum_{j=1}^{c}\sum_{k=1}^{n} u_{jk}\log u_{jk} + \sum_{k=1}^{n}\lambda_k\left(1 - \sum_{j=1}^{c} u_{jk}\right) \tag{2}$$

and therefore interpreted as “minimize the energy for a fixed entropy level”, as proposed in [23]. The terms containing coefficients β and λ_k are constraints and the coefficients themselves are Lagrange multipliers.

Energy is simply the distortion term found in many central clustering algorithms, so this formulation is closer to the familiar clustering formulation in which distortion is minimized under additional constraints.

Memberships are computed as:

$$u_{jk} = \frac{e^{u_{jk}/\beta}}{Z_k}, \tag{3}$$

where Z_k is called the partition function and is computed as

$$Z_k = \sum_{l=1}^{c} e^{u_{lk}/\beta}, \tag{4}$$

as a result of having enforced the normality constraint. It is worth stressing that the ME approach is related to the Fuzzy c-Means algorithm by a simple change of variables [15].

The Possibilistic approach is based on removing any equality constraint on the sum of memberships. This is replaced by a set of loose requirements, expressed for instance as follows:

$$u_{jk} \in [0,1] \quad \forall j \; \forall k \tag{5}$$

$$0 < \sum_{k=1}^{n} u_{jk} < n \qquad \forall j \tag{6}$$

$$\forall k \; \exists j : u_{jk} > 0 \tag{7}$$

Roughly speaking, these requirements simply imply that no cluster be empty and each pattern be assigned to at least one cluster. This turns a standard fuzzy clustering procedure into a mode seeking algorithm [20]. If the interpretation of memberships in the probabilistic model is that of probability of one of c mutually exclusive events, then in the possibilistic approach memberships can be viewed as the probability of any of c independent events (or "typicality"). For an introduction, important references, and another interpretation of the Possibilistic Approach (in the framework of robust clustering), we address the reader to the on-line tutorial indicated in Reference [24].

The first formulation of the Possibilistic Clustering algorithm [20] includes a penalty term to avoid trivial solutions. This term simply discourages low membership values, introducing a bias to 1 for all memberships. In the second formulation [22], the penalty term considers the entropy of clusters as well as their overall membership values: $\sum_{j=1}^{c} \frac{1}{\beta_j} \sum_{k=1}^{n} \left(u_{jk} \log u_{jk} - u_{jk} \right)$, so that the objective function has the following form:

$$J_{\mathrm{PC}} = \sum_{j=1}^{c} \sum_{k=1}^{n} u_{jk} d_{jk} + \sum_{j=1}^{c} \frac{1}{\beta_j} \sum_{k=1}^{n} \left(u_{jk} \log u_{jk} - u_{jk} \right) \tag{8}$$

We will refer to this version of the algorithm, which we will indicate as PCM-II. Note that this approach is equivalent to a set of c independent estimation problems [8]. The parameters β_j, "temperature" in the entropic analogy, here can be termed "scale" to emphasize the fact that each cluster is an independent M-estimator, and can be individually set for each cluster, rather than being a global property of the whole system.

Under this model, the memberships are computed as:

$$u_{jk} = e^{-u_{jk}/\beta_j}. \tag{9}$$

If we keep the entropic analogy of the ME approach, and consider the scale parameter as a global property, we can set a single value β for all the β_j. In this case the only difference with Equation (3) is that $Z_k = 1$.

This establishes a connection with the ME approach which was pointed out in [15] and also considered in [14]. In the case of equal β_j's, the partition function is generalizable to a normalization term which takes on different values according to the constraints which are placed on the sum of memberships.

By comparison of the two above expressions (3) and (9), we can generalize the membership function as follows:

$$u_{jk} = \frac{v_{jk}}{Z_k}, \tag{10}$$

where $v_{jk} = e^{-u_{jk}/\beta_j}$ for both algorithms discussed. This is a term that is used for computing the membership function *prior* to applying the partition function. Since the partition function is used to enforce constraints, we refer to the term v_{jk} as *free membership*.

III. T G P

A. The concept of graded possibility

As we already noted, the classic membership model (either hard or fuzzy) implements the concept of partitioning a set into disjoint subsets through the probabilistic constraint $\sum_{j=1}^{c} u_{jk} = 1$. Each membership is therefore formally equivalent to the probability that an experimental outcome coincides with one of c mutually exclusive events. In the possibilistic approach instead each membership is formally equivalent to the probability that an experimental outcome coincides with one of c mutually *independent* events.

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. For instance, there may be a statistical correlation between two events, rather than a causal relation; or the probability of an event can be used to bound, but not to exactly compute, the probability of other events [25]. Of course, this is a problem-dependent situation and accounting for it may or may not be appropriate. Interval and ellipsoidal models of uncertainty represent an implementation of this principle [26].

These parametric uncertainty models are common in control and optimization, where they are used for bounding the region of possible values for the variables of interest. We apply the same technique to membership values in the clustering framework. What we obtain is a model that we term *graded possibility*, and since we model uncertainty of memberships, the technique is related to Type-2 fuzzy sets [27][28]. In particular, Interval Type-2 fuzzy sets [29] are based on the same type of modeling, whereby the representation of uncertainty is simplified through the use of intervals. Note that it is also possible to define centroids as uncertain quantities, as done for instance in Interval Vector Quantization [30]. This is an entirely different model, related to Type-1 fuzzy sets.

We will base our explanations on the following geometrical concept. The ordered set of memberships to each of the clusters $\{u_1, \ldots, u_c\}$ spans a (c-dimensional) space. Within this space sets of specific membership values are represented as points, and we will term *feasible region* the locus of all points that satisfy the constraints we impose. We remark that in general the feasible region will be inside the unit (hyper)cube $[0,1]^c$. We will draw diagrams illustrating feasible regions for $c = 2$.

The standard possibilistic approach to clustering implies that all membership values are independent: the feasible region for memberships is the whole unit cube. In probabilistic approaches, there is no uncertainty: the feasible region is the plane $\sum_{j=1}^{c} u_{jk} = 1$ (more precisely, its segment bounded by the unit cube). 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 region which is smaller than the unit cube, but larger than the above plane segment.

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

B. Modeling graded possibility

We propose to use the following uncertainty model for memberships. Let $[\xi]$ be an interval variable with upper and lower values indicated by $[\underline{\xi}, \overline{\xi}]$. The memberships should be subject to the following constraint:

$$\sum_{j=1}^{c} u_{jk}^{[\xi]} = 1, \tag{11}$$

This interval equality should be interpreted as follows: the equality is satisfied for any point (set of membership values) such that there exist a scalar number $\xi^* \in [\underline{\xi}, \overline{\xi}]$ for which the equality

$$\sum_{j=1}^{c} u_{jk}^{\xi^*} = 1, \tag{12}$$

is satisfied.

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, as we will show in the following.

Equation 11, containing an interval, is equivalent to a set of two inequalities:

$$\sum_{j=1}^{c} u_{jk}^{\underline{\xi}} \geq 1 \qquad \sum_{j=1}^{c} u_{jk}^{\overline{\xi}} \leq 1.$$

This formulation includes as the two extreme cases:

- The “probabilistic” assumption:

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

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

- The “possibilistic” assumption:

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

$$\sum_{j=1}^{c} u_{jk}^{\underline{\xi} \to 0} > 1 \qquad \sum_{j=1}^{c} u_{jk}^{\overline{\xi} \to \infty} < 1$$

In the latter case considerations are done in the limit rather than in a punctual fashion. This ensures that the membership values $u_{jk} = 0 \ \forall j \ \forall k$ are avoided, as required also in the original possibilistic formulation [20]. In the limit, the two inequalities are satisfied for any choice of $u_{jk} \in [0, 1]$, since in this case $u_{jk}^{\underline{\xi}} \to 1$ and $u_{jk}^{\overline{\xi}} \to 0$.

The constraint presented above can be implemented in various ways, according to the values we choose to adopt for $\underline{\xi}$ and $\overline{\xi}$. A particular implementation, with the minimum possible number of parameters, is as follows:

$$\underline{\xi} = \alpha \qquad \overline{\xi} = \frac{1}{\alpha} \tag{13}$$

where α is a single parameter such that $\alpha \in (0, 1]$, controlling the amount of uncertainty or “possibility level”.

The remarks above, with this specific implementation, are illustrated by means of Figure 1. This type of diagram is drawn assuming $c = 2$, and shows the boundaries of the feasible region for memberships. The axes represent membership values to clusters for any possible point **x**.

In this case, we plot a family of feasible regions for the graded possibilistic case. The regions are therefore parameterized by α, which decreases from 1 to 0 in the direction of the arrows. The feasible regions are those inside the eye-shaped contours. The eye opens as α shrinks, until it fills the whole square when α vanishes.

[Figure 1 about here.]

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

$$[\xi] = [\alpha, 1]\,. \tag{14}$$

C. A graded possibilistic approach to clustering

Unfortunately, solving the clustering problem with the objective function approach is hard under the graded possibilistic model. However it is possible to exploit the generalized membership function formulation (10) to find suitable sets of membership values complying with the graded possibilistic constraint (11) for any value of α. In this way, we obtain an iterative procedure whose updating step is not directly related to the optimization of a cost function. This approach was previously followed in many works, the two previously cited [16][17] being only an example.

We first select the free membership function as:

$$v_{jk} = e^{-d_{jk}/\beta_j}, \tag{15}$$

so that it is equivalent to the ME and PCM-II algorithms.

We observe then that, as we have already seen, in the possibilistic case, v_{ij} coincides with u_{ij} (point A in Figure 2), whereas in the probabilistic case it is necessary to project point A on the feasible region $\sum_{j=1}^{c} u_{jk} = 1$ along the straight line through the origin and A, obtaining point C. This is actually the effect of the normalization obtained by defining the generalized partition function in the standard way, as in (4).

In intermediate cases, we need to distinguish whether the point falls within the eye-shaped feasible region, or it falls outside. In the first case constraint (10) is already satisfied, while in the second case it is necessary again to project point A onto the border of the feasible region, obtaining point B.

To implement this variable behavior, an appropriate definition of the generalized partition function is required. Therefore we define Z_k as follows:

$$\begin{aligned} Z_k &= \left(\sum_{j=1}^{c} v_{jk}^{1/\alpha}\right)^{\alpha} \quad \text{if} \quad \sum_{j=1}^{c} v_{jk}^{1/\alpha} > 1 \\ Z_k &= \left(\sum_{j=1}^{c} v_{jk}^{\alpha}\right)^{1/\alpha} \quad \text{if} \quad \sum_{j=1}^{c} v_{jk}^{\alpha} < 1 \\ Z_k &= 1 \quad \text{else.} \end{aligned} \tag{16}$$

These definitions ensure that, for $\alpha = 1$, the representation properties of the method reduce to those of ME, whereas in the limit case for $\alpha = 0$, the representation properties are equivalent to those of PCM-II.

Note that the three conditions are mutually exclusive for $\alpha \in (0, 1)$. A sketch of proof is as follows: we want to prove that, given a set of c quantities u_j, $j \in \{1, \ldots, c\}$ such that $u_j \in (0, 1)\ \forall j$, the following implications:

$$\sum_{j=1}^{c} u_j^{\alpha} < 1 \Rightarrow \sum_{j=1}^{c} u_j^{1/\alpha} > 1, \tag{17}$$

$$\sum_{j=1}^{c} u_j^{1/\alpha} > 1 \Rightarrow \sum_{j=1}^{c} u_j^{\alpha} < 1, \tag{18}$$

and

$$\sum_{j=1}^{c} u_j^{1/\alpha} = 1 \Rightarrow \sum_{j=1}^{c} u_j^{\alpha} = 1, \tag{19}$$

are all false.

Suppose for the sake of simplicity that $\alpha \in (0, 1)$. This implies that $u_j^{\alpha} > u_j^{1/\alpha}\ \forall j$.

Now, suppose that $\sum_{j=1}^{c} u_j^{\alpha} < 1$. From the above relation it is also $\sum_{j=1}^{c} u_j^{\alpha} > \sum_{j=1}^{c} u_j^{1/\alpha}$, therefore $\sum_{j=1}^{c} a_j^{1/\alpha} > 1$ cannot be true, so (17) is proved false.

By a similar argument also (18) can be proved false, while (19) is not true in general, but only for $\alpha = 1/\alpha$, i.e., $\alpha = 1$.

D. A note on parameter selection

The proposed method depends on one parameter, α, which is to be selected according to some criterion. At this point we should note that this parameter allows the user to introduce his/her own bias into the representation of cluster, by permitting the "soft transition" between the probabilistic and possibilistic paradigms.

As far as the other parameters involved in the formulae, they are the same as in more standard techniques, so that usual center and radius estimation techniques apply. This is especially true for the number of clusters, for which many methods are available. Since it is not the aim of this study to focus on this point, in the experiments presented below we have either used a number of clusters known from previous work, or (in the last example, Section VI) adopted a naive reduction approach, by starting from an overspecified number of clusters, performing a number of runs, and reducing clusters until a reasonable stability was achieved (centroids were found approximately in the same positions for most runs).

The idea of starting from an overspecified number of clusters and then progressively reducing it has been thoroughly studied elsewhere. It can be incorporated in a clustering algorithm [31] and also combined with a preliminary "inflation" or growth phase [32]. The process is of course driven by appropriate criteria, which may be local to clusters or may apply to the global partition, and which characterize the individual techniques.

[Figure 2 about here.]

IV. B G P

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.

Although we are mainly interested in describing the knowledge representation properties of the presented model, we will also add some practical remarks about its implementation.

Algorithm BGPC: Basic Graded Possibilistic Clustering

select c

select niter $\in \mathbb{N}$

select $\alpha \in [0, 1]$

initialize y_j

for iter = 1 to niter **begin**

  β = β(iter)

  compute v_{jk} using (15)

  compute Z_k using (16)

  compute u_{jk} using (10)

  if stopping criterion satisfied **then stop**

  else compute y_j using (1)

end

This basic algorithm can be modified in a number of standard ways. For our experiments we usually apply a multi-trial version in which new initializations are performed `ntrials` times, a selection criterion is applied, and the result of the best trial is kept. The selection criterion can be designed according to the application; usually it is a cluster validity estimate [33], [34], [35], [36]. In classification tasks, where cluster centroids are labeled at the end of training, such labeling can be used to perform the evaluation on the training set or a test set, actually turning the BGPC algorithm into a supervised one.

The value for β can be assessed from previous experiments, possibly in an independent way for each cluster (as done in PCM-II – formulas must then use β_j instead of β), estimated from data with appropriate scale estimation techniques [37], or gradually lowered in an iterated application of the algorithm (as done in the deterministic annealing approach to ME). In our experiments we used an initial value for β, to

be decreased according to an exponential decay schedule [38] of type $\beta = \beta_i (\beta_f/\beta_i)^{\texttt{iter/niter}}$ (starts at $\beta = \beta_i$ and stops at $\beta = \beta_f$). In this formulation, the number of iterations also influences the actual "annealing" schedule, since variations of β from one iteration to the next are smaller if the number of allowed iterations (`niter`) is larger.

If α is fixed, convergence relies on the size of the variation of β. If β is changed abruptly from one iteration to the other, the error surface is modified and in the next iteration the error may rise. However, since each iteration is a batch update, random fluctuations in the empirical error landscape do not occur as in stochastic optimization algorithms [39], [32]. It is also possible to implement a full deterministic annealing procedure by substituting the single update step for each value of β with a sequence of updates until convergence as in the original ME formulation [18], [19], where deterministic annealing was used. However this may not be advisable when α is low, since there is reduced or no competition among prototypes, so that centroids which are coincident are not pulled apart in further iterations and "phase transitions" [19] can not occur. Stochastic ("on-line") optimization could be used for very large datasets, where a full scan of the training set at each iteration could be expensive.

An interesting variation to the simple scheme outlined in the BGPC algorithm is obtained by the modification of α across the iterations. If α starts at $\alpha_i = 1$ and decays to a lower value $\alpha_f \in [0, 1)$, the procedure is similar to initializing the algorithm with a few probabilistic iterations (with α close to 1). The best results have been experimentally obtained with a decay rate which is linear, or steepest at the last iterations (e.g. $(\alpha_i - \alpha_f)\left(1 + e^{\texttt{iter}-\texttt{niter}}\right) + \alpha_f$).

V. D

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

A. Demonstration of the Graded Possibilistic approach

To show the properties of graded possibilistic clustering we use the toy training set shown in Figure 3. 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.

[Figure 3 about here.]

We run the BGPC algorithm 10 times, with $\overline{\xi} = 1/\alpha$ and $\underline{\xi} = \alpha$ as in the sample algorithm of Section IV, and different values of $\alpha \in [0, 1]$. We analyze the resulting memberships for varying α.

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 4 shows the membership of each of these three data points in cluster 1 (solid line) and in cluster 2 (dashed line) for various values of α ranging from 1 to 0.

[Figure 4 about here.]

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, actually 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 .25 and .15 (respectively for clusters 1 and 2).

Point #67 is clearly an outlier. However, when α is close to 1, it is classified as belonging in cluster 1 with high degree (almost 1). When α is lowered, with the transition to the possibilistic model, the values are reduced to about 0 and .07, respectively.

Figure 5 shows membership values to each cluster for points along the diagonal of the data plane, for the values $\alpha = 1$, $\alpha = .5$, and $\alpha = 0$. The different shape of the memberships, especially for points far from the separating line, is apparent.

[Figure 5 about here.]

A similar analysis is presented in Figure 6. This experiment is performed on the usual Iris dataset [40] obtained from the UCI Machine Learning Repository [41]. (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 6 about here.]

B. Outlier rejection

Outlier management is very problem-dependent, as pointed out in [42], to the point that in some problems it could be more useful to reject good clusters and to retain only outliers [43]. Here however we assume that we are interested in finding Gaussian clusters in data, and all points outside these clusters are considered to be noise. This is a standard formulation of the problem [44], but it is not the only possible one [45].

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

$$\sum_{j=1}^{c} u_{jk} \leq 1 \qquad \text{and} \qquad \sum_{j=1}^{c} {u_{jk}}^{\alpha} \geq 1. \tag{20}$$

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).

This is an effective form of soft rejection similar to that of robust estimation methods based on weighting [37], [46], [12]. Hard rejection can be implemented by thresholding (reject on value of membership) or by counting (reject on number of points) [47], [48].

The experiments involve a set of three Gaussian clusters, plus a very wide background data distribution (see Figure 7). Data are in the unit square; there are 600 data points of which 10% are clustered in 3 Gaussian clusters, while the rest are spread in the background.

[Figure 7 about here.]

[Figure 8 about here.]

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. 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.

This result was obtained with β scheduled to decay linearly from 0.1 to 0.005 over 2000 iterations in 10 different trials. In the asymmetric-possibilistic case the result illustrated was obtained 10 times.

For the probabilistic case, with the parameters indicated, we obtained 10 times a trivial solution with coincident centroids, all located at the global average of data; the illustrated result is the best one and was possible only by lowering the initial β to 0.01, and was obtained 6 times out of 10 trials.

C. An application: characterization of leukemia

As a sample application to an interesting problem, we applied the BGPC algorithm to clustering of data from a set of DNA microarray experiments, as described in [49]; data are available at the following web address: `http://www-genome.wi.mit.edu/mpr/data_set_ALL_AML.html`.

Acute leukemia is a family of diseases including two different forms, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The DNA microarray technique allows a detailed profiling of the expression level of thousands of genes. In the work cited, Golub and collaborators analyzed a set of 7129 genes for correlation with each form of leukemia, and discovered that a set of 50 "best" genes (the number was arbitrarily chosen) can produce good classification results. They also found that a Self-Organizing Map can discover clusters which correspond well to the actual distinction between ALL and AML, and moreover it can find a further distinction between two types of the lymphoblastic form, T-lineage ALL and B-lineage ALL. This task was termed "class discovery" by the authors.

We tried to verify these results using the BGPC algorithm. First, we selected the 50 most informative genes pointed out by Golub *et al.* Then we performed clustering using 2 and 3 centroids on the training set, composed of 38 examples. However, we used the separate test set of 27 examples (as provided by the authors) to verify the results, and performed repeated training (50 times) to reduce the influence of initialization.

The results are reported in Table I. They were obtained with the asymmetric option described in V-B. As in the original work, after training we performed majority labeling of centroids (or "calibration"), and used them for classifying patterns. However the evaluation, as noted, was performed on the test set. The table contains the number c of centers used, the average error as computed on the test set, the frequency of results with 0 errors and the frequency of results with 1 error over the total 50 trials. These results show that 2 centers can approximate very well the basic distinction among ALL and AML, and that using 3 centers it is possible to reach errorless performance on the test set, albeit infrequently over the repeated trials.

[Table 1 about here.]

We do not draw conclusions in comparison with the technique used in the original work. They did

not obtain errorless clusters even on the training set; on the other hand, they used all genes, not only the most informative subset as we did. However our results can confirm the validity of their selection, since the performance with the 50 best genes is much better than the one reported with the whole set of genes.

VI. T G P -

[Figure 9 about here.]

An interesting application of the proposed model is found in automated segmentation of diagnostic images in medicine. This is an important application area for data clustering. Here we analyze slices from magnetic resonance imaging (MRI) volumes.

MRI is a widely used diagnostic imaging technique. Its success is mainly due to its low invasivity and its ability to detect contrast from different physical parameters, predominantly the spin-lattice relaxation time (T1), spin-spin relaxation time (T2), and proton spin density (PD) of the tissues being imaged. The output is usually obtained in the form of a three-dimensional, multi-modal image. We can analyze the volume or individual slices (two-dimensional images extracted from volumetric acquisition), as done in the present study. Figure 9 shows an example slice. Three-dimensional data sets are then obtained by recording the T1, T2, and PD values for each voxel or pixel.

Magnetic resonance imaging is very well suited for analyzing tissues within the brain and head, and is widely applied to detection of tumors or other pathologic tissues or trauma. Image segmentation in this context aims at separating homogeneous regions and subsequently at labeling them with tissue descriptions. Clustering is one of the methods used for the first task, whereas supervised methods (possibly including human inspection) are required for the second task. Diagnostic image segmentation with clustering has been the subject of previous work by one of the authors [50].

First, we perform clustering with varying α. The results for $\alpha = 0$, $\alpha = 0.7$, and $\alpha = 1$ are presented in figure 10. These are plots of data points in the three-dimensional space of features. Black squares represent centroids. The user can select the result that best fits the distribution of points. This is possible because three-dimensional problems can still be represented visually.

[Figure 10 about here.]

Then, once a suitable set of centroids has been selected, we apply the following interactive visualization method. The centroids obtained from training are retained. The membership values are instead recomputed according to values of α and β interactively set by the user. Images corresponding to the centroids are

therefore visualized as grey-level profiles, which can be adjusted for maximum clarity and interpreted by the operator.

Each centroid represents a homogeneous region (set of mutually close points in the space of MRI features), corresponding to a given "segment" or anatomically significant area. In the MRI slice, a segment need not appear as a concentrated area, but may have an arbitrary shape. One of the centroids will represent the region of interest for our analysis, the one which most plausibly characterizes the affected area. This area can be represented as a gray level image, as shown in Figure 11.

[Figure 11 about here.]

In the figure, several values of α (horizontally) and β (vertically) are shown. Depending on the α-β combination, the interesting area can be clearly outlined, or its background can be made more visible, if the user wants to see the area within its anatomical context. In this technique, background points are those points whose membership to the interesting cluster has an intermediate value. Their brightness is therefore most sensitive to variations in the parameters α and β.

The cluster centroids here play the role of templates to point out possible segments within the image; α and β can be varied by the user as a sort of "brightness/contrast" control acting on template similarity rather than intensity levels.

This interactive visualization procedure seems very promising and features sufficient generality to be applied to other multimodal diagnostic imaging techniques as well.

VII. C

The present study demonstrates possible novel applications of the possibilistic clustering approach by introducing a method to obtain a soft transition from the possibilistic to the probabilistic models. This graded possibilistic approach is amenable to the same applications as the two previous models. In addition, the experiments suggest that there are cases in which the clustering performance can benefit of the added degree of flexibility. The discussion about the possible implementations also points out possible issues to be kept in mind while applying the technique.

However, the focus of the proposed method is on the representation capabilities. The experiment with the synthetic, bi-dimensional data set in Figure 3 is used to illustrate the behavior of the graded possibilistic membership as a function of the degree of possibility α and of the data position with respect to cluster centers. The problem of characterizing leukemia from DNA microarray data confirms the effectiveness of the proposed approach as a clustering method, by comparison with published results. Interactive post

processing in the segmentation of MRI images is another real-world application where the model is used to gain insight into the problem in a unique way. The adjustable representation capabilities can be exploited in this and other tasks (e.g., in outlier analysis) to obtain an additional perspective to cluster analysis. It is worth mentioning that other applications are currently being studied, such as the problem of variable selection in linear clustering [51].

A

The authors are partly supported by a grant by the Italian Ministry of Education, University and Research (PRIN 2004 project 2004062740), and by the Biopattern EU Network of Excellence.

R

[1] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, New Jersey, USA: Prentice Hall, 1988.

[2] A. Jain, M. Murty, and P. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

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

[4] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. L. Cam and J. Neyman, Eds., vol. I. University of California, January 1967, pp. 281–297.

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

[6] 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.

[7] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum, 1981.

[8] O. Nasraoui and R. Krishnapuram, "Crisp interpretations of fuzzy and possibilistic clustering algorithms," in *Proceedings of the 3rd European Congress on Intelligent Techniques and Soft Computing, Aachen, Germany*, vol. 3, August 1995, pp. 1312–1318.

[9] A. Baraldi and P. Blonda, "A survey of fuzzy clustering algorithms for pattern recognition. I," *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 29, pp. 778–785, 1999.

[10] H. Frigui and R. Krishnapuram, "A robust clustering algorithm based on m-estimator," in *Proceedings of the 1st International Conference on Neural, Parallel and Scientific Computations, Atlanta, USA*, vol. 1, May 1995, pp. 163–166.

[11] N. R. Pal, K. Pal, and J. C. Bezdek, "A mixed c-Means clustering model," in *FUZZIEEE97: Proceedings of the International Conference on Fuzzy Systems*. IEEE, Barcelona, Spain, 1997, pp. 11–21.

[12] R. N. Davé and R. Krishnapuram, "Robust clustering methods: a unified view," *IEEE Transactions on Fuzzy Systems*, vol. 5, no. 2, pp. 270–293, May 1997.

[13] H. Frigui and R. Krishnapuram, "A robust competitive clustering algorithm with applications in computer vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 450–465, May 1999.

[14] M. Ménard, V. Courboulay, and P.-A. Dardignac, "Possibilistic and probabilistic fuzzy clustering: unification within the framework of the non-extensive thermostatistics," *Pattern Recognition*, vol. 36, no. 6, pp. 1325–1342, June 2003.

[15] S. Miyamoto and M. 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] T. A. Runkler and J. 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.

[17] T. Kohonen, *Self-Organizing Maps*. Springer, 2001.

[18] K. Rose, E. Gurewitz, and G. Fox, "A deterministic annealing approach to clustering," *Pattern Recognition Letters*, vol. 11, pp. 589–594, 1990.

[19] ——, "Statistical mechanics and phase transitions in clustering," *Physical Review Letters*, vol. 65, pp. 945–948, 1990.

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

[21] M. Barni, V. Cappellini, and A. Mecocci, "Comments on 'A Possibilistic Approach to Clustering'," *IEEE Transactions on Fuzzy Systems*, vol. 4, no. 3, pp. 393–396, 1996.

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

[23] S. Miyamoto, "An overview and new methods in fuzzy clustering," in *IEEE Conference on Knowledge-Based Intelligent Electronic Systems*, vol. 1. Adelaide, Australia: IEEE, Piscataway NJ, USA, April 1998, pp. 33–40.

[24] O. Nasraoui, "A brief overview of robust clustering techniques." [Online]. Available: http://www.louisville.edu/˜o0nasr01/Websites/tutorials.html

[25] H. E. Kyburg, "Interval valued probabilities," electronic collection, at http://ippserv.ugent.be/, 1998, documentation of the Imprecise Probabilities Project.

[26] A. Bargiela, "Interval and ellipsoidal uncertainty models," in *Granular computing: an emerging paradigm*. Physica-Verlag GmbH, 2001, pp. 23–57.

[27] L. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning-1," *Information Sciences*, vol. 8, pp. 199–249, 1975.

[28] J. M. Mendel and R. I. John, "Type-2 fuzzy sets made simple," *IEEE Transactions on Fuzzy Systems*, vol. 10, no. 2, pp. 117–127, April 2002.

[29] D. Schwartz, "The case for an interval-based representation of linguistic truth," *Fuzzy Sets and Systems*, vol. 17, pp. 153–165, 1985.

[30] S. Ridella, S. Rovetta, and R. Zunino, "IAVQ–interval-arithmetic vector quantization for image compression," *IEEE Transactions on Circuits and Systems, Part II*, vol. 47, no. 12, pp. 1378–1390, December 2000.

[31] R. Krishnapuram and C.-P. Freg, "Fitting an unknown number of lines and planes to image data through compatible cluster merging." *Pattern Recognition*, vol. 25, no. 4, pp. 385–400, 1992.

[32] S. Ridella, S. Rovetta, and R. Zunino, "Plastic algorithm for adaptive vector quantization," *Neural Computing and Applications*, vol. 7, no. 1, pp. 37–51, 1998.

[33] R. Krishnapuram, H. Frigui, and O. Nasraoui, "Fuzzy and possibilistic shell clustering algorithms and their application to boundary detection and surface approximation – part ii," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 1, pp. 44–60, 1995.

[34] N. R. Pal and J. C. Bezdek, "On cluster validity for the Fuzzy c-Means model," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 3, pp. 370–379, 1995.

[35] ——, "Corrections to 'On cluster validity for the Fuzzy c-Means model'," *IEEE Transactions on Fuzzy Systems*, vol. 5, no. 1, pp. 152–153, 1997.

[36] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2/3, pp. 107–145, 2001.

[37] O. Nasraoui and R. Krishnapuram, "An improved Possibilistic *C* Means algorithm with finite rejection and robust scale estimation," in *Proceedings of the North American Fuzzy Information Processing Society Conference, Berkeley, USA*, 1996, pp. 395–399.

[38] H. Ritter, T. Martinetz, and K. Schulten, *Neuronale Netze*. München, Germany: Addison-Wesley, 1991.

[39] T. Martinetz, S. Berkovich, and K. Schulten, "'Neural gas' network for vector quantization and its application to time-series prediction," *IEEE Transactions on Neural Networks*, vol. 4, no. 4, pp. 558–569, 1993.

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

[41] C. Blake and C. Merz, "UCI repository of machine learning databases," 1998, URL: http://www.ics.uci.edu/~mlearn/MLRepository.html. [Online]. Available: http://www.ics.uci.edu/~mlearn/-MLRepository.html

[42] F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel, *Robust Statistics: the Approach Based on Influence Functions*. New York, USA: John Wiley & Sons, 1986.

[43] J. Han and M. Kamber, *Data Mining, Concepts and Techniques*. Morgan Kaufmann, San Francisco, 2001.

[44] R. N. Davé, "Characterization and detection of noise in clustering," *Pattern Recognition Letters*, vol. 12, no. 11, pp. 657–664, November 1991.

[45] F. Angiulli and C. Pizzuti, "An extension to possibilistic fuzzy cluster analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 203–215, February 2005.

[46] O. Nasraoui and R. Krishnapuram, "A robust estimator based on density and scale optimizations and its application to clustering," in *FUZZIEEE96: Proceedings of the International Conference on Fuzzy Systems*. IEEE, New Orleans, USA, 1996, pp. 1031–1035.

[47] E. Knorr and R. Ng, "Algorithms for mining distance-based outliers in large datasets," *VLDB Journal*, vol. 8, no. 2/3, pp. 237–253, 2000.

[48] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proceedings of ACM International Conference on Management of Data (SIGMOD00)*, 2000, pp. 427–438.

[49] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, October 1999.

[50] F. Masulli and A. Schenone, "A fuzzy clustering based segmentation system as support to diagnosis in medical imaging," *Artificial Intelligence in Medicine*, vol. 16, no. 2, pp. 129–147, 1999.

[51] K. Honda, H. Ichihashi, F. Masulli, and S. Rovetta, "Graded possibilistic approach to variable selection in linear fuzzy clustering," in *Proceedings of the 15th IEEE International Conference on Fuzzy Systems, Reno, USA*, June 2005, pp. 985–990.

L F

1 Bounds of the feasible region for memberships in the case $c = 2$, for a generic point $\mathbf{x}$, with α decreasing from 1 to 0 along the direction of the arrows. 21
2 How to compute the value of memberships for a point $\mathbf{x}$ ($c = 2$). Point A: unconstrained membership (equivalent to PCM-II). Point B: membership constrained on the graded possibilistic feasible region boundary. Point C: membership constrained on the standard feasible region (equivalent to ME). 22
3 Toy problem to evaluate the behavior of the algorithm. 23
4 Memberships of points #9, #10, and #67 in each cluster. 24
5 Plot of memberships along the diagonal of the square area. Above: $\alpha = 1$; middle: $\alpha = 0.5$; below: $\alpha = 0$. 25
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). 26
7 Dataset for the outlier rejection demonstration. 27
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). (Note that some triangles are hidden by true cluster centers since they are almost coincident.) 28
9 A MRI slice. From left to right: T1-weighted, T2-weighted, proton density channels. The nature of tissues can be inferred by their combined response to the three stimulations (channels), so that by comparing the three images different anatomical entities can be outlined. 29
10 Training results for the dataset in Figure 9 for $\alpha = 1$ (above), $\alpha = 0.7$ (middle), $\alpha = 0$ (below). 30
11 Using the proposed model as an interactive visualization tool for image segmentation. . . 31

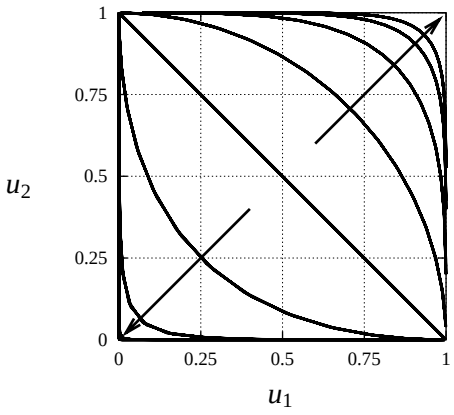

Fig. 1. Bounds of the feasible region for memberships in the case $c = 2$, for a generic point $\mathbf{x}$, with α decreasing from 1 to 0 along the direction of the arrows.

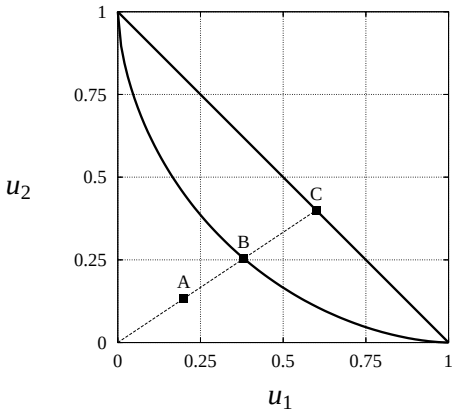

Fig. 2. How to compute the value of memberships for a point **x** ($c = 2$). Point A: unconstrained membership (equivalent to PCM-II). Point B: membership constrained on the graded possibilistic feasible region boundary. Point C: membership constrained on the standard feasible region (equivalent to ME).

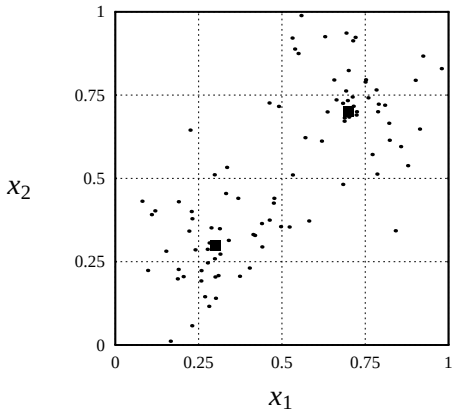

Fig. 3. Toy problem to evaluate the behavior of the algorithm.

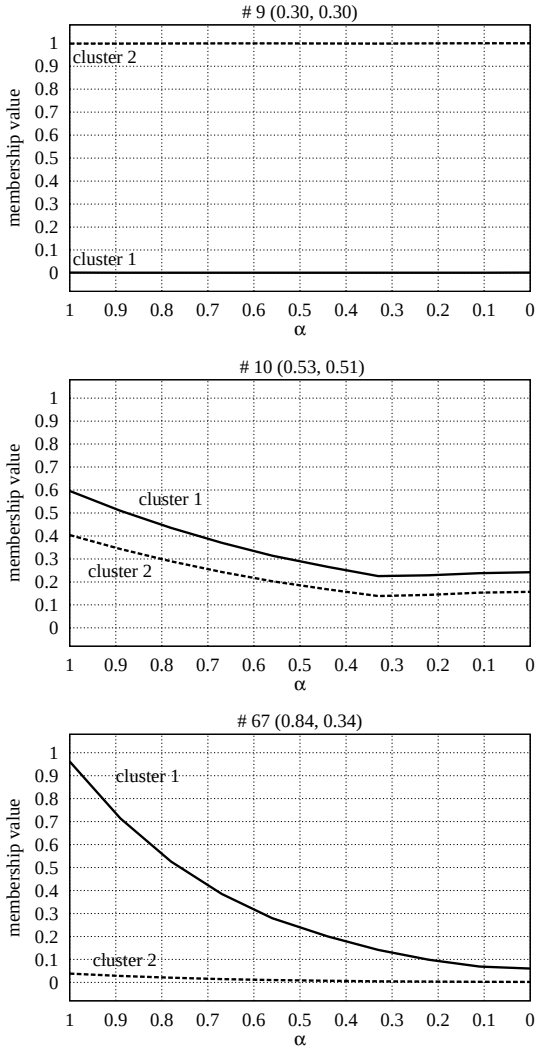

Fig. 4. Memberships of points #9, #10, and #67 in each cluster.

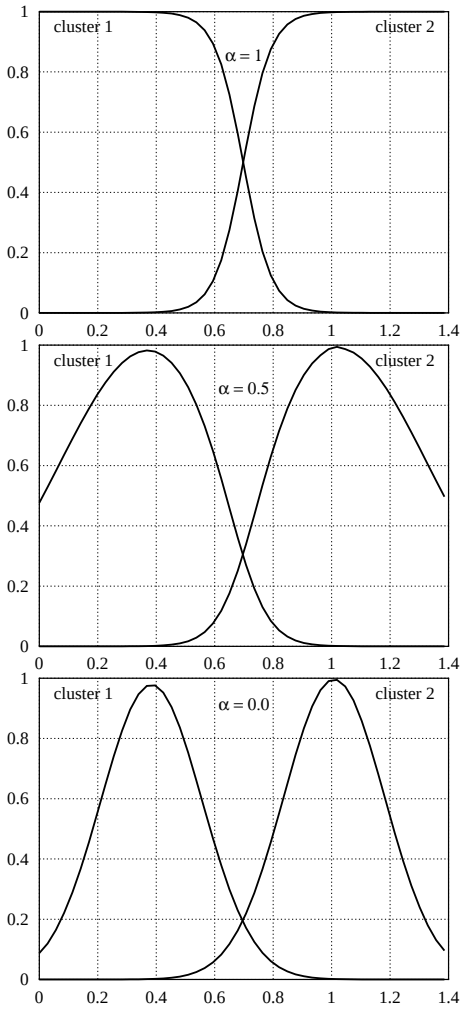

Fig. 5. Plot of memberships along the diagonal of the square area. Above: $\alpha = 1$; middle: $\alpha = 0.5$; below: $\alpha = 0$.

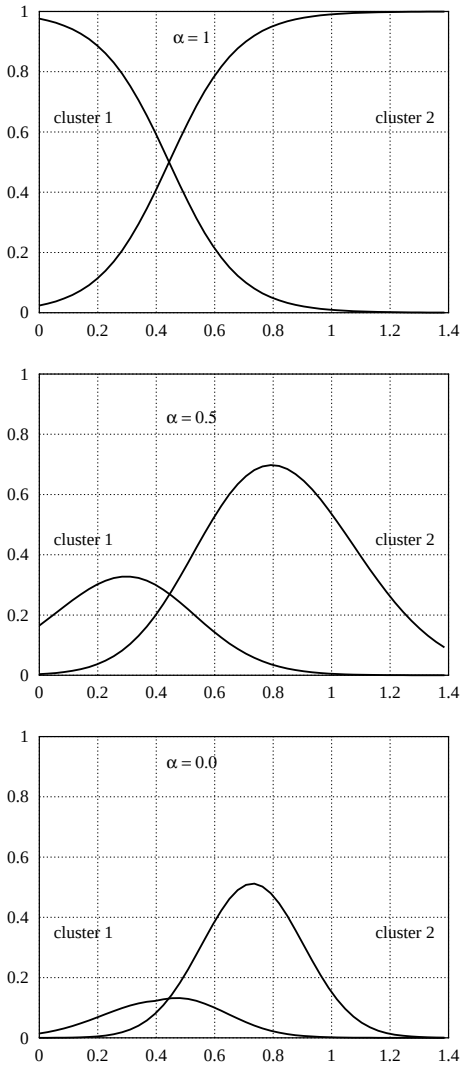

Fig. 6. Two-dimensional plot of memberships for $\alpha = 1.0$ (above), for $\alpha = 0.5$ (middle), and for $\alpha = 0.0$ (below) for the Iris dataset (same analysis as in Figure 5).

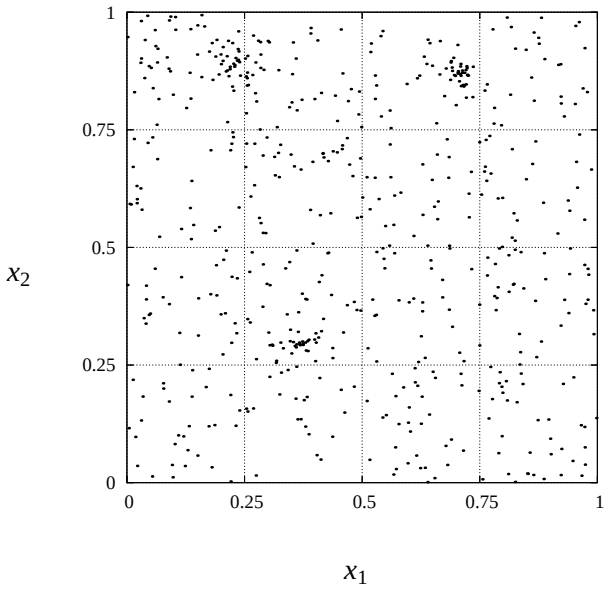

Fig. 7. Dataset for the outlier rejection demonstration.

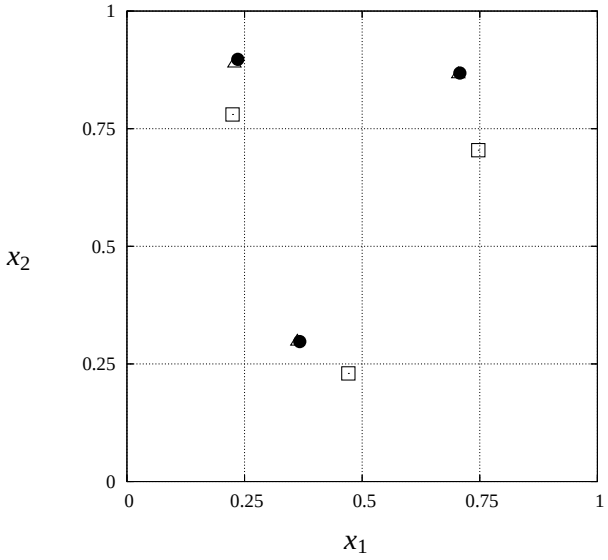

Fig. 8. Results for the outlier rejection demonstration. Black circles: true cluster centers; triangles: centers found with $\alpha = 0$ (maximum rejection); squares: centers found with $\alpha = 1$ (no rejection). (Note that some triangles are hidden by true cluster centers since they are almost coincident.)

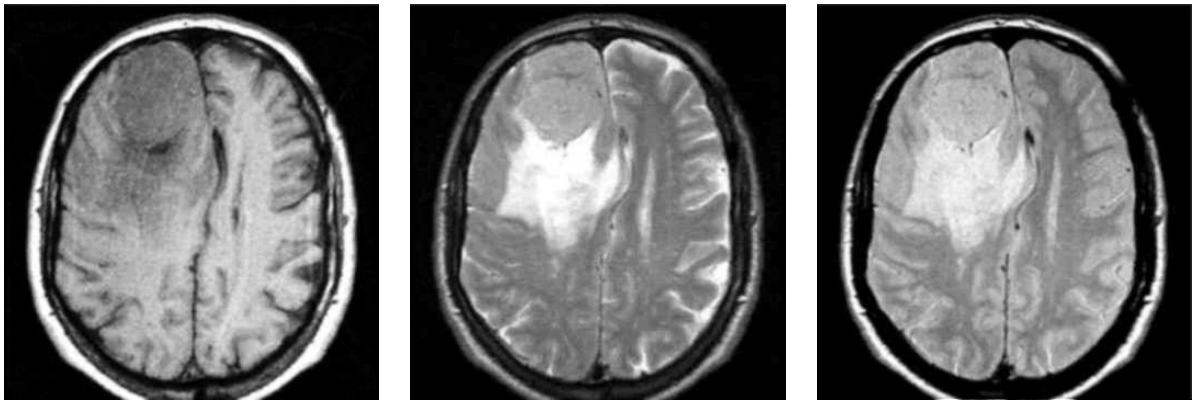

Fig. 9. A MRI slice. From left to right: T1-weighted, T2-weighted, proton density channels. The nature of tissues can be inferred by their combined response to the three stimulations (channels), so that by comparing the three images different anatomical entities can be outlined.

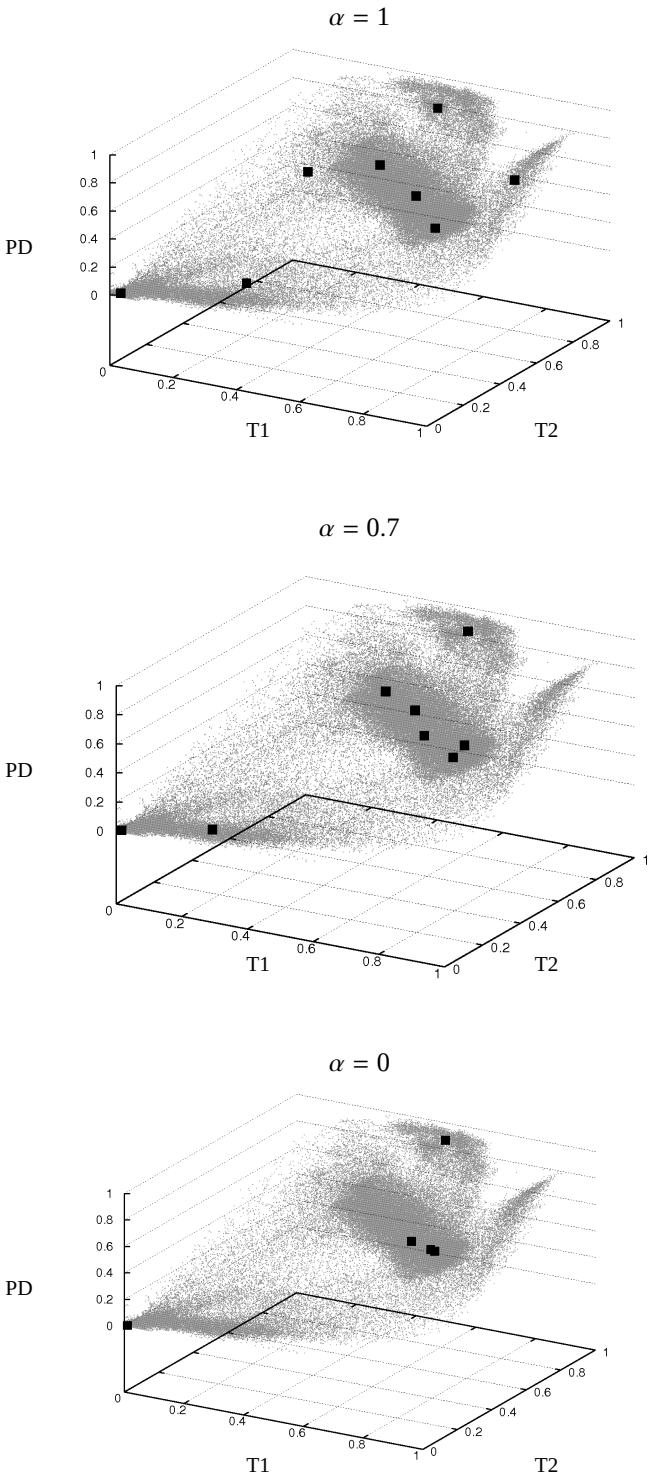

Fig. 10. Training results for the dataset in Figure 9 for $\alpha = 1$ (above), $\alpha = 0.7$ (middle), $\alpha = 0$ (below).

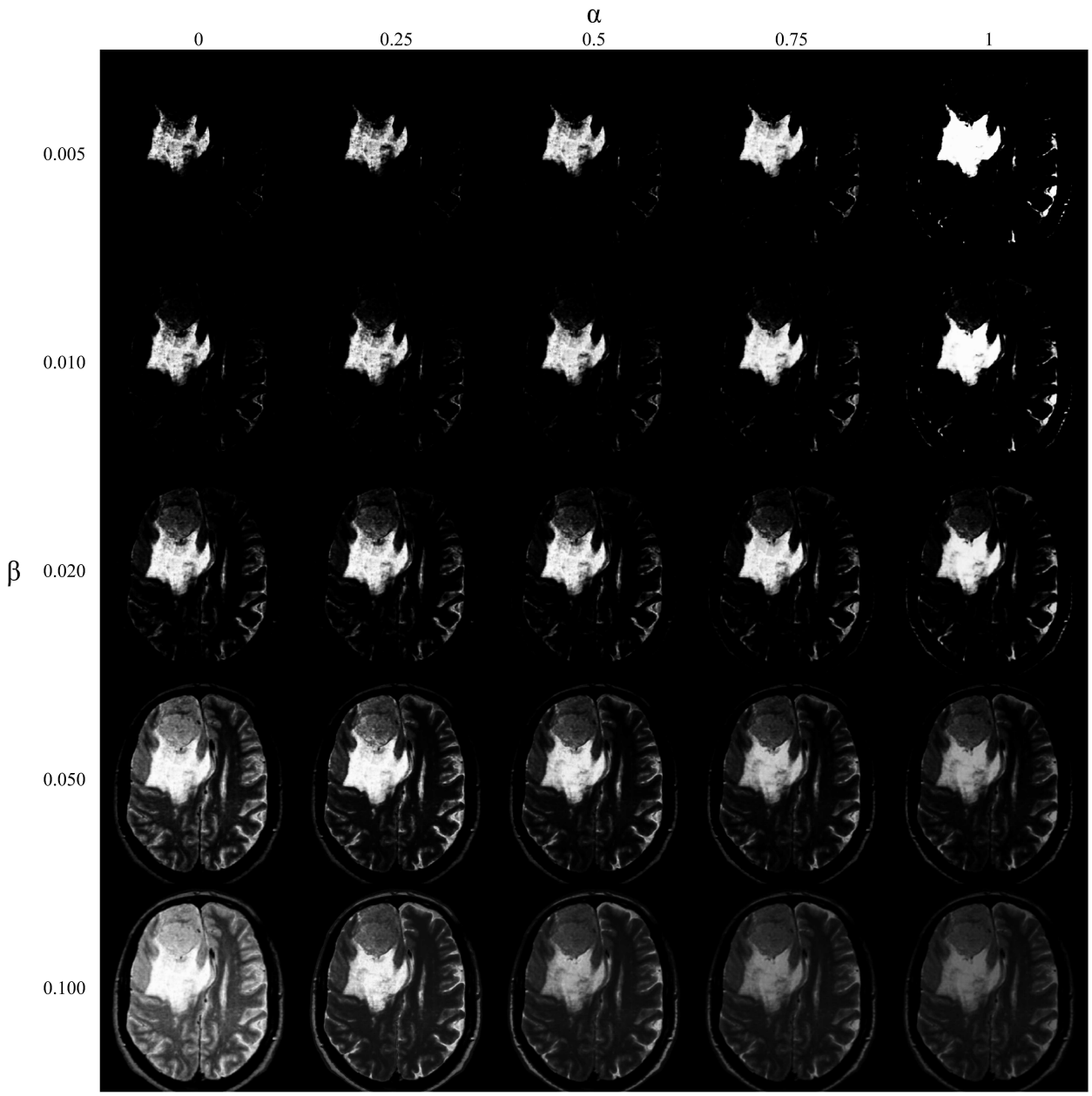

Fig. 11. Using the proposed model as an interactive visualization tool for image segmentation.

L T

I Results on the characterization of leukemia 33

TABLE I

R

c	**Avg. error**	**Freq. 0 errors**	**Freq. 1 error**
2	3.70%	0/50	50/50
3	3.48%	3/50	47/50