An Algorithm to Model Paradigm Shifting in Fuzzy Clustering

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

[1] INFM, Istituto Nazionale per la Fisica della Materia, 16146 Genova, Italy
[2] Dipartimento di Informatica, Università di Pisa, 56125 Pisa, Italy
masulli@di.unipi.it
[3] DISI, Università di Genova, 16146 Genova, Italy
rovetta@disi.unige.it

Abstract. The *graded possibilistic clustering paradigm* includes as the two extreme cases the "probabilistic" assumption and the "possibilistic" assumption adopted by many clustering algorithms. We propose an implementation of a graded possibilistic clustering algorithm based on an interval equality constraint enforcing both the normality condition and the required graded possibilistic condition. Experimental results highlight the different properties attainable through appropriate implementation of a suitable graded possibilistic model.

1 Introduction

Let $X = \{\mathbf{x}_k | \, k = 1, ..., n\}$ be the set of unlabeled samples; $Y = \{\mathbf{y}_j | j = 1, ..., c\}$ be the set of cluster centers (or prototypes); and $U = [u_{jk}]$ be the *fuzzy membership matrix*.

Many clustering approaches, such as C-Means (CM) [3], Fuzzy C-Means (PCM) [2], and Deterministic Annealing (DA) [9, 1], assume a *probabilistic constraint*, according to which the sum of the membership values of a point in all the clusters must be equal to one. 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.

In [5, 6], Krishnapuram and Keller showed the limits of the probabilistic approach to clustering and proposed a *possibilistic approach* to it. Their approach assumes the membership function of a point in a *fuzzy* set (or cluster) is absolute, i.e. it is an evaluation of a *degree of typicality* not depending on the membership values of the same point in other clusters.

Krishnapuram and Keller [5, 6] presented two version of a Possibilistic C-Means algorithm (PCM) that relax the probabilistic constraint, in order to allow a *possibilistic* interpretation of the membership function as a *degree of typicality*. In PCM, the elements of U fulfill the following conditions:

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

$$0 < \sum_{k=1}^{n} u_{jk} < n \;\; \forall \;\; j; \tag{2}$$

$$\bigvee_j \; u_{jk} > 0 \;\; \forall \;\; k. \tag{3}$$

Then, the possibilistic approach implies 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 ($\psi \equiv 0$).

Note that, due to lack of competitiveness among clusters, clustering algorithms based on the possibilistic approach, need of an initial distribution of prototypes in the feature space and the estimation some parameters, that can be obtained using a probabilistic clustering methods. E.g., in [5, 6], a Fuzzy C-Means initialization has been applied, while Masulli and Schenone [8] used a prototypes initialization based on the Capture Effect Neural Network(CENN) [4].

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 subset of the interval $[0, 1]$.

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

2 Modeling graded possibility

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

$$\psi = \sum_{j=1}^{c} u_{jk}^{[\xi]} - 1, \tag{4}$$

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 α, where

$$\underline{\xi} = \alpha \qquad \overline{\xi} = \frac{1}{\alpha} \tag{5}$$

and

$$\alpha \in [0, 1] \tag{6}$$

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:

$$\alpha = 0$$

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

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

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 α 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 α 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.

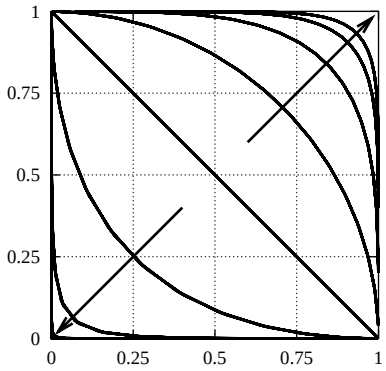

Fig. 1. Bounds of the feasible region for u_{jk} for different values of α (decreasing from 1 to 0 along the direction of the arrows)

3 The graded possibilistic clustering algorithm

In this section we outline a basic example of graded possibilistic clustering algorithm (Tab 1). 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:

$$v_{jk} = e^{-d_{jk}/\beta_j}. \tag{7}$$

The generalized partition function can be defined as follows:

$$Z_k = \sum_{j=1}^{c} v_{jk}^{\kappa} \tag{8}$$

where:

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

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. Note that in the implementation of the algorithm in Tab 1 the variation of α from 1 to 0 allow to obtain a probabilistic initialization of prototypes and a following refinement in a possibilistic sense.

The required value for the β_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).

Table 1. Graded possibilistic clustering algorithm

```
select c
select alphastep ∈ ℝ
randomly initialize y_j
for α = 1 downto 0 by alphastep do
begin
    compute v_jk using (7)
    compute Z_k using (8)
    compute u_jk = v_jk/Z_k
    if stopping criterion satisfied  then  stop
    else compute the centroids y_j
end
```

4 Experimental analysis

In [7] we report some results aimed to highlighting the properties attainable through appropriate implementation of a suitable graded possibilistic model. The showed results demonstrated that:

1. the proposed implementation of the graded possibilistic model (Tab. 1) is able to correctly model the membership functions of data point without need of long experimental work, as necessary with the PCM, and
2. a very high outliers rejection is attainable, by setting the upper extremum of $[\xi]$ to 1 and the lower extremum to α.

In this section we illustrate a case of a-priori knowledge usage. We propose an experimental demonstration where we make 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.

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 α are shown in Figure 2.

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

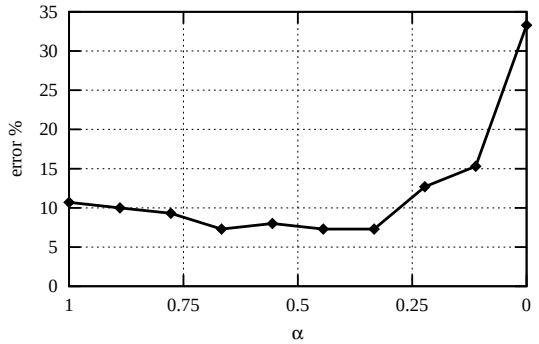

Fig. 2. Error percentage plot for the unsupervised Iris classification.

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 [6].

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.

5 Conclusions

The concept of graded possibility applied to clustering, which has been presented in this paper, is a flexible tool for knowledge representation. By tuning the level of possibility it is possible to represent overlapped clusters, as in standard possibilistic clustering, with the added capability to adapt the level of overlap to the problem at hand. This results in interesting rejection capabilities and in an adaptable trade-off between the mode-seeking and the partitioning behaviors of its two special cases – possibilistic and standard (probabilistic) fuzzy clustering.

Our current activities invvolve the application of this flexible behavior in the areas of Web content analysis, document data mining, DNA microarray data analysis. Deeper theoretical investigations are planned as well.

References

[1] G. Beni and X. Liu. A least biased fuzzy clustering method. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 16:954–960, 1994.

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

[3] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.

[4] F. Firenze and P. Morasso. The capture effect model: a new approach to self-organized clustering. In *Sixth International Conference. Neural Networks and their Industrial and Cognitive Applications. NEURO-NIMES 93 Conference Proceedings and Exhibition Catalog*, pages 65–54, Nimes, France, 1993.

[5] Raghu Krishnapuram and James M. Keller. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1(2):98–110, May 1993.

[6] Raghu Krishnapuram and James M. Keller. The possibilistic *C*-Means algorithm: insights and recommendations. *IEEE Transactions on Fuzzy Systems*, 4(3):385–393, August 1996.

[7] F. Masulli and S. Rovetta. Soft transition from probabilistic to possibilistic fuzzy clustering. Technical Report DISI-TR-02-03, Dipartimento di Informatica e Scienze dell'Informazione, Università di Genova (Italy), Italy, 2002. (*http:// www.disi.unige.it/ person/ MasulliF/ papers/ DISI-TR-02-03-masulli.pdf*).

[8] F. Masulli and A. Schenone. A fuzzy clustering based segmentation system as support to diagnosis in medical imaging. *Artificial Intelligence in Medicine*, 16:129–147, 1999.

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