Soft Clustering: Why and How-To

Stefano Rovetta[1] and Francesco Masulli[1,2](✉)

[1] DIBRIS, University of Genova, Via Dodecaneso 35, 16146 Genoa, Italy
{stefano.rovetta,francesco.masulli}@unige.it
[2] Sbarro Institute for Cancer Research and Molecular Medicine, Temple University,
Philadelphia, PA, USA

Abstract. Despite the huge success of machine learning methods in the last decade, a crucial issue is to control the support of the data used in inference, so that data that are too far from the training set are given low confidence by default. The most important class that features this ability is that of prototype-based methods which are based on clustering or vector quantization as a representation learning model. This paper surveys a family of popular soft clustering methods, framing them in a unified formalism. It also discusses the peculiarities of each of them. A large fraction of the paper is devoted to clarifying the role of model parameters and to providing some guidelines on how to set up these parameters.

Keywords: Fuzzy clustering · Possibilistic clustering · Graded possibilistic clustering · Model parameters

1 Introduction

Despite the huge success of machine learning methods in the last decade, several issues remain unsolved. Machine learning usually focuses on black-box models which suffer from lack of explainability and, dually, from difficulty in using prior knowledge. In the specific case of deep learning [15] an additional issue is that theories of generalisation are apparently not applicable. In fact, adversarial machine learning techniques [16] seem to prove that generalisation ability is actually low in deep neural networks, and that bad quality outputs can easily be produced with high confidence. This is a very serious issue when machine learning is used in life-critical contexts like autonomous vehicle guidance or condition-based monitoring in predictive maintenance of sensitive plants.

In view of these problems, it is imperative to control the support of the data used in inference, so that data that are too far from the training set are given low confidence by default. The most important class that features this ability is that of prototype-based methods which are based on clustering or vector quantization as a representation learning model.

In the literature, these methods have been used extensively [2,23,26,27] although they may appear to be less popular than other approaches (in particular deep learning and support vector machines). As noted, clustering, and specifically soft clustering, is a key component.

R. Fullér et al. (Eds.): WILF 2018, LNAI 11291, pp. 67–82, 2019.
https://doi.org/10.1007/978-3-030-12544-8_6

In this perspective, this paper surveys a family of popular soft clustering methods, framing them in a unified formalism. It also discusses the peculiarities of each of them. A large fraction of the paper is devoted to clarifying the role of model parameters and to providing some guidelines on how to set up these parameters.

2 Soft Clustering

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

Some methods also employ a relational approach by measuring observation-observation closeness [9,13]; these are not considered here, but we cite them for completeness.

The following definitions deal with real-valued quantities and crisp sets, and therefore the symbols $\in$ and $\cup$ have the usual crisp-set-theoretic meaning:

Definition 1 (Fuzzy and possibilistic partitions [4]**).** *Given a set* $X = \{x_1, x_2, \ldots, x_n\}$ *of data items, a set* $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$*, and a membership function* $u(x, \omega)$*,* $x \in X$*, with* $0 \leq u(x, \omega) \leq 1 \ \forall x \in X$*,* $\forall \omega \in \Omega$*, the pair* (Ω, u) *is:*

- *A* ***possibilistic partition*** *if*

$$u(x, \omega) \in \mathbb{R} \quad \forall x, \forall \omega \qquad \text{and} \qquad 0 < \sum_{i=1}^{c} u(x, \omega_i) < c \quad \forall x \tag{1}$$

- *A* ***fuzzy partition*** *if it is a possibilistic partition with*

$$\sum_{i=1}^{c} u(x, \omega_i) = 1 \quad \forall x \tag{2}$$

- *A* ***crisp partition*** *if it is a fuzzy partition with*

$$\max_i u(x, \omega_i) = 1 \quad \forall x. \tag{3}$$

□

In the case of central clustering, partitions are represented by centroids.

Definition 2 (Central clustering). *A* ***central clustering*** *is a (crisp, fuzzy, possibilistic) partition of a metric data space* Ξ *whose membership functions are monotonically dependent on the similarity of objects to a set of centroids* $\{y_1, \ldots, y_c\} \subset \Xi$*.* □

Some methods not dealt with in this work, for instance those based on medoids or landmarks, require $\{y_1, \ldots, y_m\} \subset X$.

Central clustering is especially interesting as a concept representation tool because it can be learned from a training set X and applied to the whole data space Ξ. Many other approaches to clustering do not possess this *out-of-sample extension* property and can therefore only be used to partition the given data set.

The most widely used fuzzy clustering method is probably the *Fuzzy c-Means/Fuzzy ISODATA* [6,12,29] (FCM) algorithm, which is a "fuzzy relative" to the simple c-Means technique [5]. FCM defines the ω_j as fuzzy partitions of the data set X.

Well-known limitations of FCM include the need for fixing a fuzziness parameter in addition to the number of centroids, dependency on the initialisation, convergence to possibly bad-quality local solutions, the consequent need for many restarts, and a membership function profile that may not discriminate sharply enough between close and far points.

Variations over this basic scheme try to overcome some of these limitations. All of the following methods have membership functions that involve exponentials rather than powers of distance, which are sharper (for a discussion about this point see for instance [19]).

The *Maximum Entropy* (ME) approach, usually but not necessarily associated to the *Deterministic Annealing* optimisation procedure [24,25], does not minimize a simple cost term, but a compound cost function which is the sum of a distortion term $\hat{E}$ and an entropic term $-H$ (see the next section for the mathematical definitions). The optimization is done by fixing a constant value for one of the two terms and minimizing the other; then this step is iterated for decreasing values of the constant, until a global optimum is reached. This alleviates the false minima problem of standard c-Means and (to a lesser extent) of FCM.

In decision-making and classification applications, algorithms should feature several desirable properties in addition to the basic discrimination or decision function. For instance, it is usually required that in certain configurations a decision is not made (*pattern rejection*). This situation typically occurs in the presence of outliers. This problem is very well-known and well studied (see for instance [7,8,11]), and is tackled in a convenient way within the framework of soft-computing, fuzzy, and neural approaches [10,17,23].

However, the clustering problem as stated above implies that the outlier rejection property cannot be achieved. This is because the membership values are constrained to sum to 1. By giving up the requirement for strict partitioning, and by resorting to a "mode seeking" algorithm, Krishnapuram and Keller proposed the so-called *possibilistic approach* [18,19], where this constraint is relaxed essentially to

$$u_{jl} \in [0,1] \quad \forall l, \forall j \tag{4}$$

With this model outlier rejection can be achieved, but at the expense of a clear cluster attribution and other computational drawbacks. The same issue of analysing the membership interactions on a local basis, as opposed to the global effects induced by the probabilistic model, is considered in [14].

An additional clustering model that can be thought of as a generalization of all those outlined above can be devised starting from the following observations.

Crisp partitions constrain membership in a very strong way: For a given object, memberships to all clusters must be zero except one. Fuzzy partitions relax this constraint in the sense that all membership can be non-zero, provided that their sum is still one. This means that membership to one cluster directly affects the membership to all other clusters. Finally, possibilistic partitions don't impose any constraint on memberships.

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. To model this idea, we could require the membership to one cluster to have an influence on the other memberships, but not so strong as to determine it directly.

This brings us to the concept of *graded possibility.* An example of such concept 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 close to 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.

These ideas form the rationale of the Graded possibilistic c Means clustering methods, described in the following.

3 Some Popular Clustering Algorithms: A Unified View

3.1 The c-Means Family

We will now review some clustering algorithms derived from the basic c-Means: ("hard" or "crisp") c-Means (HCM) [5], Minimum-Entropy fuzzy clustering by Deterministic Annealing (ME) [24], Possibilistic c-Means with an entropic cost term (PCM-II) [19], Fuzzy c-Means (FCM) [12], Graded Possibilistic c-Means (GPCM) [21]. All of these techniques are based on minimizing the following cost function:

$$\hat{E} = \sum_{j=1}^{c} \sum_{l=1}^{n} u_{jl} d_{jl}. \tag{5}$$

(this includes also FCM, although in the usual formulation this is not evident; see [22]). We will refer collectively to these algorithms as the c-Means (CM) family.

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

Miyamoto and Mukaidono [22] show that these algorithms are obtained by adding to the basic cost $\hat{E}$ in (5) either regularization terms or the maximum-entropy term

$$-H = \sum_{j=1}^{c} \sum_{l=1}^{n} u_{jl} \log u_{jl} \tag{6}$$

which represents the (negative) entropy of the clustering defined by Y, U.

Figure 1 shows how the effect of fuzziness parameters on the objective function corresponds to regularization.

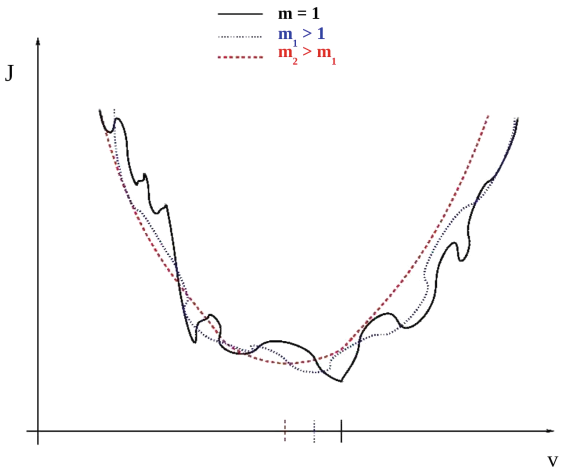


Fig. 1. Regularizing effect of fuzziness parameters on the objective function.

In clustering problems the focus is commonly placed on the analysis of data and clusters themselves, rather than on minimization of a global error criterion. We are often more interested in characterizing (hopefully significant) groups of data than in representing the details of the data with a faithful approximation. As an example, *model-based* clustering approaches focus on cluster modeling rather than performance optimization, and the cluster identification technique called *Alternating Cluster Estimation* [28] does not even assume the existence of a cost function.

Therefore we will introduce a formalism to provide an alternative, unified perspective on these clustering algorithms, focused on the memberships u_{jl} rather than on the cost function.

3.2 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

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

All the CM algorithms considered here define either:

$$J = \hat{E} \tag{7}$$

or:

$$J = \hat{E} - H \tag{8}$$

where the cluster entropy acts as a regularizer.

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

Let v_{jl} be the solution of a CM problem with constraint ψ removed (formally this can be implemented with $\psi \equiv 0$). We call v_{jl} the *free membership* of pattern x_l in cluster ω_j.

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

$$y_j = \frac{\sum_{l=1}^{n} u_{jl} x_l}{\sum_{l=1}^{n} u_{jl}} \tag{9}$$

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

$$u_{jl} = \frac{v_{jl}}{Z_l}, \tag{10}$$

where Z_l is the (generalized) partition function, which is computed as a function of the conventional partition function $\zeta_l = \sum_{j=1}^{c} v_{lj}$:

$$Z_l = f(\zeta_l) \tag{11}$$

Since the specific form of $f()$ is given by the constraint ψ, a member of the CM family is equivalently defined by the pair (J, f) or (J, Z_l).

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

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

A method to allow for non-extreme solutions is the maximum entropy criterion, which is implemented in the ME 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 ME.

Table 1. The CM family of clustering algorithms

	J	v_{jl}	Z_l	*Notes*
ME	$\hat{E} - H$	$e^{-d_{jl}\beta}$	$\sum_{j=1}^{c} v_{jl}$	$\beta \in \mathbb{R}$, $\beta > 0$ is the inverse temperature parameter to be increased during the "annealing" process
PCM-II	$\hat{E} - H$	$e^{-d_{jl}\beta_j}$	1	$\beta_j \in \mathbb{R}$, $\beta_j > 0$ are cluster width parameters to be selected a priori before optimization or using heuristic criteria
FCM	$\hat{E}$	$1/d_{jl}$	$\left(\sum_{j=1}^{c} v_{jl}^{1/(m-1)}\right)^{m-1}$	$m \in \mathbb{R}$, $m > 1$ is the fuzzification parameter
HCM	$\hat{E}$	*See note*	*See note*	v_{jl} and Z_l can be written as for FCM, but their values have to be computed in the limit for $m \to 1$
GPCM	$\hat{E}$	$e^{-d_{jl}\beta_j}$	$\left(\sum_{j=1}^{c} v_{jl}\right)^{\alpha}$	$\beta_j \in \mathbb{R}$, $\beta_j > 0$ are cluster width parameters to be selected a priori before optimization or using heuristic criteria. $\alpha \in [0, 1]$ is the degree of probabilistic tendence

4 Membership Function Parametrization

All soft clustering methods require at least one model parameter, which in general terms decides the degree of fuzziness of the solution.

Since Miyamoto and Mukaidono [22] showed that the power membership function of FCM can be transformed into the exponential one of the other methods, the following discussion will only focus on the methods featuring the latter form, i.e., ME, PCM-II, GPCM.

4.1 Possible Parametrizations in the CM Family

The original formulation of free membership in ME features one global parameter β, interpreted as a global temperature, energy, disorder, or resolution.

$$v_{lj} = \exp\left(-\beta \|\boldsymbol{x}_l - \boldsymbol{y}_j\|^2\right) \tag{12}$$

The Deterministic Annealing optimization procedure fixes the temperature at each optimisation step, making it effectively a regularisation coefficient rather than a model parameter.

In contrast, PCM-II features one parameter β_j per centroid.

$$v_{lj} = \exp\left(-\beta_j \|\boldsymbol{x}_l - \boldsymbol{y}_j\|^2\right) \tag{13}$$

In this case the parametrization can be considered that of a system with non-constant energy, i.e., out of thermodynamic equilibrium.

It is also possible to write a free membership function with parameters that differ for each of the vector components of the centroid, although to the best of our knowledge no popular method from the literature features the anisotropic parametrizations described in the following.

Using one *vector* parameter per centroid, with one component β_{ji} per centroid j per component i of the space Ξ, we obtain the following free membership function:

$$v_{lj} = \exp\left(-\sum_{i=1}^{d}(x_{li} - y_{ji})^2 \beta_{ji}\right). \tag{14}$$

In this case, parameters β_{ji} form a $c \times d$ matrix

$$\boldsymbol{B} = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1d} \\ & \vdots & \\ & \beta_{ji} & \\ & \vdots & \\ \beta_{c1} & \cdots & \beta_{cd} \end{pmatrix} \tag{15}$$

and, indicating with $\boldsymbol{b}_j$ the j-th column of $\boldsymbol{B}$, the argument of the exponential can be written in vector-matrix notation:

$$v_{lj} = \exp\left(-(\boldsymbol{x}_l - \boldsymbol{y}_j)^T \operatorname{diag}(\boldsymbol{b}_j)(\boldsymbol{x}_l - \boldsymbol{y}_j)\right) \tag{16}$$

where $\operatorname{diag}(\boldsymbol{v})$ denotes the diagonal matrix that has vector $\boldsymbol{v}$ as its diagonal.

This case is equivalent to a non-equilibrium, anisotropic system with axis-parallel principal directions of anisotropy.

The most general parametrization is obtained when the principal directions of anisotropy are not necessarily the coordinate axes. In this case there is a matrix of coefficients for each centroid, not necessarily diagonal, using a generalised (Mahalanobis) distance [20]:

$$v_{lj} = \exp\left(-\sum_{i=1}^{d}\sum_{k=1}^{d}(x_{li} - y_{ji})(x_{lk} - y_{jk})\boldsymbol{B}_{jik}\right) \tag{17}$$

or

$$v_{lj} = \exp\left(-(\boldsymbol{x}_l - \boldsymbol{y}_j)^T \boldsymbol{B}_j (\boldsymbol{x}_l - \boldsymbol{y}_j)\right) \tag{18}$$

This case implies that the model parameters are contained in a rank-3 tensor of shape (c, d, d). For each j, the corresponding $d \times d$ slice $\boldsymbol{B}_j$ is analogous to an inverse covariance matrix as used in the multidimensional form of the Gaussian density function and consequently in the expression of the Mahalanobis distance.

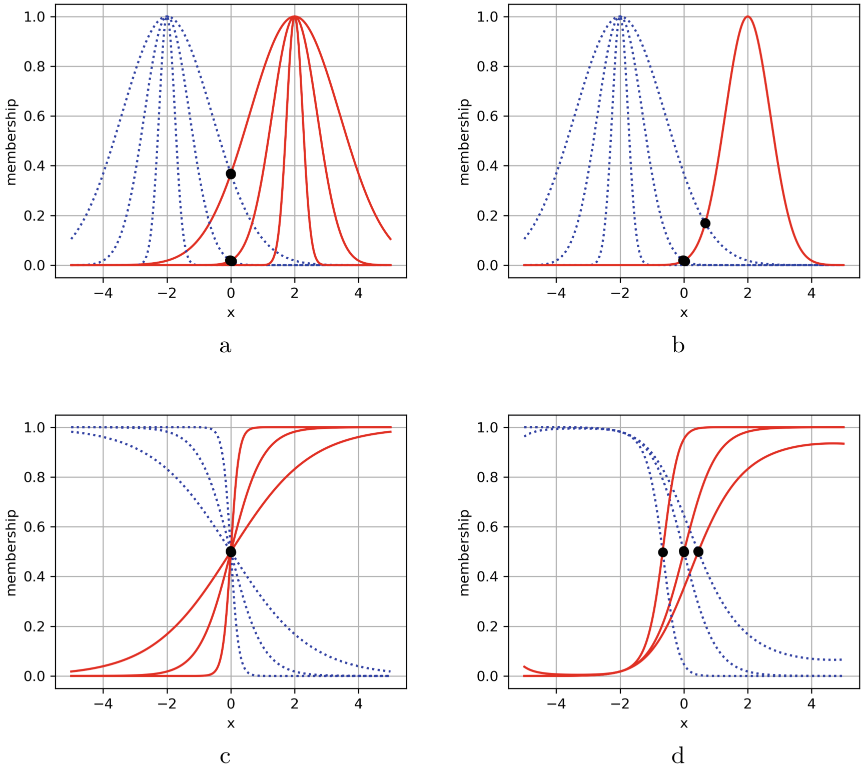
In addition to these model parameters, GPCM also has an additional parameter α that can be used to set the balance between a possibilistic and a probabilistic behaviour. In the first formulation [21], an interval-valued variable was used. In subsequent works, see for instance [3], a simpler formulation was adopted where $\alpha \in [0,1] \subset \mathbb{R}$.

4.2 Roles of Parameters

According to the original statistical mechanics analogy, the parameter β in EM can be interpreted as an inverse temperature. From the point of view of information representation, it plays the role of a degree of fuzziness: When β increases (i.e., temperature decreases), the memberships of data observations to clusters become crisper. Finally, from a geometrical interpretation, β is a global resolution parameter that defines the minimum distance between centroids to be considered as distinct; below this distance, centroids collapse into each other.

The limit cases are:

- for $\beta \to 0^+$, we have $u_{lj} = 1/c$ for all l, j, i.e., each instance is equally associated with each cluster;



a

b

c

d

Fig. 2. Effect of varying β or $\boldsymbol{b}$ on the membership of point x to cluster 1 (dotted blue) and 2 (solid red). (Color figure online)

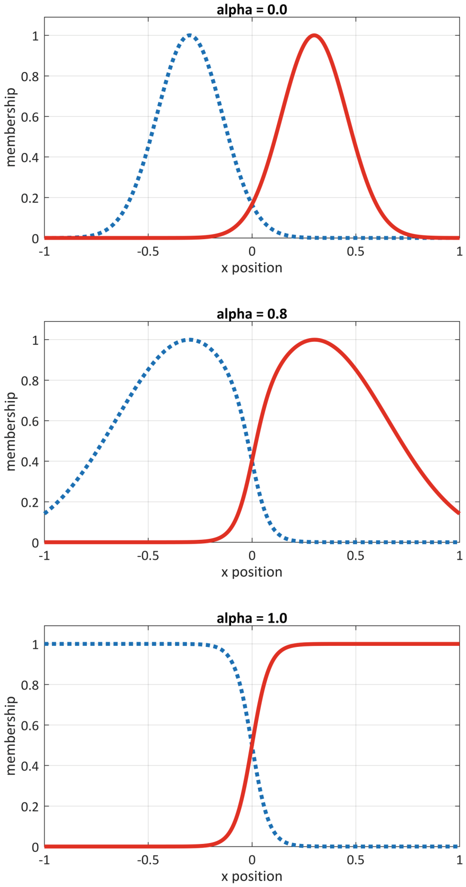


Fig. 3. Effect of varying α on the membership of point x to cluster 1 (dotted blue) and 2 (solid red). (Color figure online)

– for $\beta \to +\infty$, we have $u_{lj} = 1$ if $\boldsymbol{x}_l \in \omega_j$, and $u_{lk} = 0$ for all $k \neq j$, $k \in [1, c]$, i.e., each instance is associated with only one cluster (hard limit).

In the case of individual β_j per cluster, the size of clusters is affected individually. However, in all cases that are not purely possibilistic, the memberships influence each other via the partition function. This has an effect on the critical position for an observation, the point where its maximum membership switches from one centroid to another.

In Fig. 2 the effect of changing the temperature or resolution parameters is illustrated in a two-centroid case. Membership to the two centroids are plotted in different styles. The critical points are marked in black for each choice of parameter values. On the left (graphs a, c) a single global parameter β is used, assigning it three different values; on the right (graphs b, d) individual parameters for each centroid are used, resulting in a vector $\boldsymbol{b} = [\beta_1, \beta_2]$, and only β_1

is changed, again using three values. The top graphs (a, b) are the possibilistic cases; the bottom graphs (c, d) are the probabilistic ones. The effect of having different resolution factors for different centroids on the critical point is clearly visible in graphs b and d.

The global model parameter $\alpha \in [0, 1]$ sets the nature of the clustering model, with $\alpha = 0$ corresponding to a fully possibilistic model (pure mode-seeking), $\alpha = 1$ to a probabilistic model, and intermediate values corresponding to a partly possibilistic behaviour where the generalized partition function does not normalize the sum of memberships to a fixed value of 1 but to a value that depends on the values of all free memberships. An illustration of the effect of varying α in a 2-cluster problem is presented in Fig. 3.

4.3 Factorisation of Parameters

As already noted, the single parameter β of ME is used both as a model parameter, acting on the structure of the final clustering, and as an optimisation parameter, influencing the convergence of the optimisation itself.

It may be useful to express the two concepts in an uncoupled way to allow both actions simultaneously. To this end, we rewrite the most general parametrization (rank-3 tensor) as

$$\beta_{jik} = b\overline{\beta}_{jik} \tag{19}$$

where $\overline{\beta}_{jik}$ expresses the relative magnitude of parameters with respect to each other and b is a global scale factor that can be used for annealing. Disregarding a change of units, all choices for this decomposition are equivalent; we can fix the ideas by setting $\max\{\overline{\beta}_{jik}\} = 1$ which results in $\max\{\beta_{jik}\} = b$, i.e., the global scale parameter is the magnitude of the largest β_{jik}.

In the following we discuss some possible criteria to estimate the model parameters just discussed.

5 Setting the Model Parameters

With respect to the optimization, model parameters can be set beforehand, at each iteration, or at the end. While setting the parameters before the beginning only works in the presence of a good initialisation, the criteria here presented can easily be applied during the iterations or after their end.

By necessity, all criteria ultimately depend on some user-selected parameters. The focus of the methods that are discussed in this section is to reduce the number of these parameters to a minimum and to provide an intuitive interpretation to make it possible for the user to assign meaningful values to these residual degrees of freedom.

In the following we only cover the case of vector scale parameter, $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_c]$. The scalar case is similar but obviously simpler, and the matrix and tensor cases are not as common.

5.1 Setting the Resolution Parameters Using Free Memberships $\boldsymbol{v}$

Criteria for setting $\boldsymbol{\beta}$ can be obtained by analysing inter-centroid distance and imposing a bias toward fuzzy solutions, similarly to what was done in the possibilistic approach in [18,19]. The first proposed method uses free memberships. For each centroid $\boldsymbol{y}_j$ we measure the free membership to its cluster ω_j of all other centroids:

$$v(\boldsymbol{y}_h, \boldsymbol{y}_j) = \exp\left(-\|\boldsymbol{y}_h - \boldsymbol{y}_j\|^2 \beta_j\right) \tag{20}$$

Note that this measure is taken using $\boldsymbol{y}_j$ as a reference and is asymmetric, i.e., $v(\boldsymbol{y}_h, \boldsymbol{y}_j) \neq v(\boldsymbol{y}_j, \boldsymbol{y}_h)$.

We define the minimal-overlap condition by setting a threshold $t \in (0, 1)$. Membership of centroid h to centroid j should not be larger than this threshold. Enforcing this for the nearest centroid guarantees that this is true also for all other centroids. To guarantee absolutely no overlap, the value should be $t = 1/2$. Other values can be used if some overlap is acceptable ($t > 1/2$) or if narrower boundaries are desired ($t < 1/2$).

The criterion is therefore:

$$\begin{aligned} & \max_{h \neq j} \; v(\boldsymbol{y}_h, \boldsymbol{y}_j) \leq t \\ \Rightarrow \quad & \max_{h \neq j} \; \exp\left(-\|\boldsymbol{y}_h - \boldsymbol{y}_j\|^2 \beta_j\right) \leq t \\ \Rightarrow \quad & \min_{h \neq j} \; \|\boldsymbol{y}_h - \boldsymbol{y}_j\|^2 \beta_j \geq -\ln t \end{aligned} \tag{21}$$

Let $h^* = \arg\min_{h \neq j} \; \|\boldsymbol{y}_h - \boldsymbol{y}_j\|$. Note that being the nearest neighbour is not a symmetric relation, so in general β_j and β_{h^*} will be different.

The above inequality yields the final criterion:

$$\Rightarrow \quad \beta_j = -\frac{\ln t}{\|\boldsymbol{y}_{h^*} - \boldsymbol{y}_j\|^2} \tag{22}$$

where the numerator can be used as a global degree of freedom, for instance for regularisation or annealing during the optimization (see Subsect. 4.3).

5.2 Setting the Resolution Parameters Using Memberships $\boldsymbol{u}$

In this case the function to be used is the fuzzy probabilistic one:

$$u(\boldsymbol{y}_k, \boldsymbol{y}_j) = \frac{\exp\left(-\|\boldsymbol{y}_h - \boldsymbol{y}_j\|^2 \beta_j\right)}{\sum_{k=1, k \neq j}^{c} \exp\left(-\|\boldsymbol{y}_k - \boldsymbol{y}_j\|^2 \beta_j\right)} \tag{23}$$

In this case the minimal-overlap condition:

$$\max_{h \neq j} \; u(\boldsymbol{y}_h, \boldsymbol{y}_j) \leq t \tag{24}$$

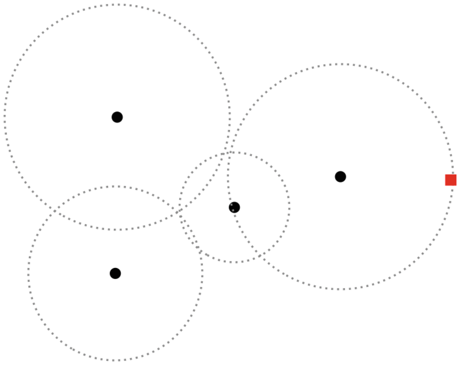


Fig. 4. Setting the value of parameter α by assigning a desired outlier membership. (Color figure online)

is much less simple to solve for β_j. However the value of the partition function (the denominator) can be estimated by using a very rough approximation. We fix an integer number c_{NN} between 1 and c. Among the centroids, we decide to take into account the nearest c_{NN}. The value of $v(\boldsymbol{y}_h, \boldsymbol{y}_j)$ for the neighbours is approximated as:

- For the c_{NN} nearest neighbours, $v(\boldsymbol{y}_h, \boldsymbol{y}_j) \approx 1$
- For the remaining $1 - c_{\mathrm{NN}}$ (farthest) neighbours, $v(\boldsymbol{y}_h, \boldsymbol{y}_j) \approx 0$

So we can estimate $\sum_{k=1}^{c} v(\boldsymbol{y}_k, \boldsymbol{y}_j)$ to be approximately equal to the number c_{NN} of neighbours sufficiently close to j. The criterion thus obtained is:

$$\beta_j = -\frac{\ln\,(c_{\mathrm{NN}} t)}{\|\boldsymbol{y}_{h^*} - \boldsymbol{y}_j\|^2} \tag{25}$$

where the numerator, a positive real number, can again be used as a global degree of freedom.

5.3 Setting the Possibility Degree α with an Outlier Rejection Criterion

In contrast to the resolution parameters, it is difficult to visualize the effect of α on cluster shape in geometric terms. This is a global parameter that influences the global configuration of clusters and interacts with the other model parameters.

A guideline for the selection of α is to set it in relation to the desired degree of outlier rejection. An outlier is an observation that has low membership to all clusters. We remark that outlier rejection is a crucial property to avoid meaningless generalisation due to extrapolation. However, complete outlier insensitivity makes the clustering model miss potentially meaningful observations. So our goal here is to set a desired worst-case membership u^* sufficiently small so as to clearly indicate outliers, but still sufficiently large to allow some effect of outliers in the centroid equations.

Supposing that the resolution parameters have been fixed, it is possible to calculate v_j for a point that lies on the border of the support of clusters. In Fig. 4 dotted circles are loci of constant free membership v, meaning that all points falling on dotted lines have the same free membership to the cluster to which the circle is referred. We want to assign the final membership u of the outlier (red square) a given value $u^* \leq v_j$ by setting the value of α.

Under the simplifying hypothesis that $v_h = 0 \ \forall h \neq j$, so that $Z = \sum_{h=1}^{c} (v_h)^\alpha = v_j^\alpha$:

$$\begin{aligned} & \frac{v_j}{v_j^\alpha} = u^* \\ \Rightarrow \quad & v_j^{1-\alpha} = u^* \\ \Rightarrow \quad & \alpha = 1 - \log u^* / \log v_j \end{aligned} \tag{26}$$

5.4 Setting the Possibility Degree α as an Independent Parameter

The value of α can also be assigned independently as a degree of freedom for regularisation or annealing. However, since it acts as an exponent, the effect of changes is much stronger when close to 1 than close to 0. Experimentally, it can be observed that the values between 0.9 and 1 are the most interesting, with values below 0.75 establishing an essentially pure possibilistic behaviour.

It is therefore advisable to set the value of α by means of an auxiliary variable that is related to it logarithmically. A suggested technique is to set $a \in [0, 1]$ so that

$$\alpha = (\log_2(a + 1))^{0.2} \tag{27}$$

where the exponent 0.2 is chosen such that, for $a = 0.5$, $\alpha \approx 0.9$. In this way the interesting range $(0.9, 1.0)$ is mapped onto half the range of variation of the control variable a. See Fig. 5 for a graph illustrating this effect.

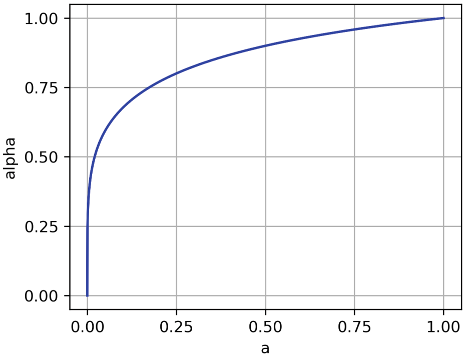


Fig. 5. Setting the value of parameter α via an auxiliary variable a.

6 Conclusion

In this paper we have reviewed a family of central soft clustering methods. Their relevance as feature learning methods for subsequent recognition, approximation, and forecasting tasks has been mentioned.

A key issue of these variations over HCM is the larger number of model parameters. Therefore, several criteria for setting these parameters have been discussed.

Current work on this topic involves the on-line adaptation of model parameters to non-stationary stream learning [1,3].

References

1. Abdullatif, A., Masulli, F., Rovetta, S., Cabri, A.: Graded possibilistic clustering of non-stationary data streams. In: Petrosino, A., Loia, V., Pedrycz, W. (eds.) WILF 2016. LNCS (LNAI), vol. 10147, pp. 139–150. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52962-2_12
2. Abdullatif, A., Rovetta, S., Masulli, F.: Layered ensemble model for short-term traffic flow forecasting with outlier detection. In: 2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI), pp. 1–6, September 2016. https://doi.org/10.1109/RTSI.2016.7740573
3. Abdullatif, A., Masulli, F., Rovetta, S., Cabri, A.: A fuzzy clustering approach to non-stationary data streams learning. In: Lintas, A., Rovetta, S., Verschure, P.F., Villa, A.E. (eds.) ICANN 2017, Part II, pp. 768–769. Springer, Cham (2017)
4. Anderson, D.T., Bezdek, J.C., Popescu, M., Keller, J.M.: Comparing fuzzy, probabilistic, and possibilistic partitions. IEEE Trans. Fuzzy Syst. **18**(5), 906–918 (2010). https://doi.org/10.1109/TFUZZ.2010.2052258
5. Ball, G., Hall, D.: ISODATA, an iterative method of multivariate analysis and pattern classification. Behav. Sci. **12**, 153–155 (1967)
6. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell (1981)
7. Chow, C.K.: An optimum character recognition system using decision function. IRE Trans. Electron. Comput. **6**, 247–254 (1957)
8. Chow, C.: An optimum recognition error and reject tradeoff. IEEE Trans. Inf. Theory **16**, 41–46 (1970)
9. Corsini, P., Lazzerini, B., Marcelloni, F.: A new fuzzy relational clustering algorithm based on the fuzzy C-means algorithm. Soft Comput. **9**(6), 439–447 (2005). https://doi.org/10.1007/s00500-004-0359-6
10. Drago, G.P., Ridella, S.: Possibility and necessity pattern classification using an interval arithmetic perceptron. Neural Comput. Appl. **8**(1), 40–52 (1999)
11. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley, New York (1973)
12. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybern. **3**, 32–57 (1974)
13. Filippone, M., Camastra, F., Masulli, F., Rovetta, S.: A survey of kernel and spectral methods for clustering. Pattern Recogn. **40**(1), 176–190 (2008)
14. Flores-Sintas, A., Cadenas, J.M., Martin, F.: Local geometrical properties application to fuzzy clustering. Fuzzy Sets Syst. **100**, 245–256 (1998)

15. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
16. Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I., Tygar, J.D.: Adversarial machine learning. In: Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, AISec 2011, pp. 43–58. ACM, New York (2011). https://doi.org/10.1145/2046684.2046692
17. Ishibuchi, H., Nii, M.: Neural networks for soft decision making. Fuzzy Sets Syst. **115**(1), 121–140 (2000)
18. Krishnapuram, R., Keller, J.M.: A possibilistic approach to clustering. IEEE Trans. Fuzzy Syst. **1**(2), 98–110 (1993)
19. Krishnapuram, R., Keller, J.M.: The possibilistic C-means algorithm: insights and recommendations. IEEE Trans. Fuzzy Syst. **4**(3), 385–393 (1996)
20. Mahalanobis, C.P., et al.: On the generalised distance in statistics. Proc. National Inst. Sci. India **2**(1), 49–55 (1936)
21. Masulli, F., Rovetta, S.: Soft transition from probabilistic to possibilistic fuzzy clustering. IEEE Trans. Fuzzy Syst. **14**(4), 516–527 (2006). https://doi.org/10.1109/TFUZZ.2006.876740
22. Miyamoto, S., Mukaidono, M.: Fuzzy C-means as a regularization and maximum entropy approach. In: Proceedings of the Seventh IFSA World Congress, Prague, pp. 86–91 (1997)
23. Ridella, S., Rovetta, S., Zunino, R.: K-winner machines for pattern classification. IEEE Trans. Neural Netw. **12**(2), 371–385 (2001)
24. Rose, K., Gurewitz, E., Fox, G.: A deterministic annealing approach to clustering. Pattern Recogn. Lett. **11**, 589–594 (1990)
25. Rose, K., Gurewitz, E., Fox, G.: Statistical mechanics and phase transitions in clustering. Phys. Rev. Lett. **65**, 945–948 (1990)
26. Rovetta, S., Masulli, F.: Online spectral clustering and the neural mechanisms of concept formation. In: Bassis, S., Esposito, A., Morabito, F.C. (eds.) Advances in Neural Networks: Computational and Theoretical Issues. SIST, vol. 37, pp. 61–72. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18164-6_7
27. Rovetta, S., Masulli, F., Cabri, A.: Measuring clustering model complexity. In: Lintas, A., Rovetta, S., Verschure, P.F.M.J., Villa, A.E.P. (eds.) ICANN 2017. LNCS, vol. 10614, pp. 434–441. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68612-7_49
28. Runkler, T.A., Bezdek, J.C.: Alternating cluster estimation: a new tool for clustering and function approximation. IEEE Trans. Fuzzy Syst. **7**(4), 377–393 (1999)
29. Ruspini, E.H.: A new approach to clustering. Inf. Control **15**(1), 22–32 (1969)