Objective Functions for Fuzzy Clustering

Christian Borgelt

Abstract Fuzzy clustering comprises a family of prototype-based clustering methods that can be formulated as the problem of minimizing an objective function. These methods can be seen as "fuzzifications" of, for example, the classical c-means algorithm, which strives to minimize the sum of the (squared) distances between the data points and the cluster centers to which they are assigned. However, it is well known that in order to "fuzzify" such a crisp clustering approach, it is not enough to merely allow values from the unit interval for the variables encoding the assignments of the data points to the clusters (that is, for the elements of the partition matrix): the minimum is still obtained for a crisp data point assignment. As a consequence, additional means have to be employed in the objective function in order to obtain actual degrees of membership. This paper surveys the most common fuzzification means and examines and compares their properties.

1 Introduction

The general objective of *clustering* or *cluster analysis* [14, 23, 26, 20] is to group given objects in such a way that objects from the same cluster are as similar as possible, while objects from different clusters are as dissimilar as possible. In order to formalize the notion of similarity, so that it becomes mathematically treatable, it is usually expressed as a *distance measure* between points (or vectors) representing the objects in a metric space, usually $\mathbb{R}^m$. Two objects are then seen as the more similar, the smaller the distance between the data points that represent them.

A common approach to describe the clusters is to use *prototypes* that capture the location and possibly also the shape and size of the clusters in the data space. With such an approach the general objective of clustering can be reformulated as the task

Christian Borgelt
European Centre for Soft Computing, Edificio de Investigación, c/ Gonzalo Gutiérrez Quirós s/n, 33600 Mieres, Asturias, Spain, e-mail: christian@borgelt.net

to find a set of cluster prototypes together with an assignment of the data points to them, so that the data points are as close as possible to their assigned prototypes. By formalizing this approach, and using for the prototypes only points in the data space that represent the *cluster centers*, one obtains immediately the objective function of classical c-means clustering [1, 19, 32]: simply sum the (squared) distances of the data points to the center of the cluster to which they are assigned. The c-means algorithm then strives to minimize this objective function.

Unfortunately, c-means clustering always partitions the data, that is, each data point is assigned to one cluster and one cluster only. This is often inappropriate, as it can lead to somewhat arbitrary cluster boundaries and certainly does not treat points properly that lie between two (or more) clusters without belonging to any of them unambiguously. Solutions to this problem consist in either using a probabilistic approach, like applying the expectation maximization (EM) algorithm to a mixture of Gaussians (see, for example, [11, 15, 6]), or to employ one of the different "fuzzifications" of the classical crisp scheme (see, for instance, [37, 13, 2, 4, 20, 7]).

In this paper I focus on the latter approach, that is, on how the objective function of classical c-means clustering can be modified in order to obtain graded cluster memberships. I survey different methods that have been suggested in the literature and examine and compare their properties. The remainder of this paper is organized as follows: Section 2 introduces the presuppositions made and the notation used in this paper. Section 3 briefly reviews the formal basis of the classical c-means algorithm. The following two sections discuss the main classes of "fuzzification" approaches: Section 4 explores membership transformation and Section 5 examines membership regularization as tools to obtain graded memberships from a modified objective function. Finally, Section 6 draws conclusions from the discussion.

2 Presuppositions and Notation

We are given a data set $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ with n data points, each of which is an m-dimensional real-valued vector, that is, $\forall j; 1 \leq j \leq n : \mathbf{x}_j = (x_{j1}, \dots, x_{jm}) \in \mathbb{R}^m$. These data points are to be grouped into c clusters, each of which is described by a prototype $\mathbf{c}_i$, $i = 1, \dots, c$. The set of all prototypes is denoted by $\mathbf{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_c\}$. I confine myself here to cluster prototypes that consist merely of a cluster center, that is, $\forall i; 1 \leq i \leq c : \mathbf{c}_i = (c_{i1}, \dots, c_{im}) \in \mathbb{R}^m$. The assignment of the data points to the cluster centers is encoded as a $c \times n$ matrix $\mathbf{U} = (u_{ij})_{1 \leq i \leq c; 1 \leq j \leq n}$, which is often called the *partition matrix*. In the crisp case, a matrix element $u_{ij} \in \{0, 1\}$ states whether data point $\mathbf{x}_j$ belongs to cluster $\mathbf{c}_i$ or not. In the fuzzy case, $u_{ij} \in [0, 1]$ states the degree to which $\mathbf{x}_j$ belongs to $\mathbf{c}_i$ (degree of membership).

In this paper I also confine myself to the (squared) Euclidean distance as the measure for the distance between a data point $\mathbf{x}_j$ and a cluster center $\mathbf{c}_i$, that is,

$$d_{ij}^2 \;=\; d^2(\mathbf{c}_i, \mathbf{x}_j) \;=\; (\mathbf{x}_j - \mathbf{c}_i)^\top (\mathbf{x}_j - \mathbf{c}_i) \;=\; \sum_{k=1}^{m} (x_{jk} - c_{ik})^2.$$

A common alternative is the (squared) Mahalanobis distance with a cluster specific covariance matrix Σ_i [18, 17], that is, $d_{ij}^2 = (\mathbf{x}_j - \mathbf{c}_i)^\top \Sigma_i^{-1} (\mathbf{x}_j - \mathbf{c}_i)$. However, this choice adds at least a shape parameter and in some approaches also a size parameter to the cluster prototypes (see, for example, [4, 20, 7]). Nevertheless, extending the approaches to this distance measure is usually fairly straightforward. An extension to the L_1-distance [24], that is, to $d_{ij} = \sum_{k=1}^{m} |x_{jk} - c_{ik}|$, or to other Minkowski metrics is less simple to achieve, but certainly beyond the scope of this paper.

3 Classical *c*-Means Clustering

As already stated, classical c-means clustering strives to find, for a given data set $\mathbf{X}$, a set $\mathbf{C}$ of cluster centers and a partition matrix $\mathbf{U}$, such that the objective function

$$J(\mathbf{X}, \mathbf{C}, \mathbf{U}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} d_{ij}^2$$

is minimized under the constraints $\forall i; 1 \leq i \leq c : \forall j; 1 \leq j \leq n : u_{ij} \in \{0, 1\}$ and $\forall j; 1 \leq j \leq n : \sum_{i=1}^{c} u_{ij} = 1$. These constraints ensure that each data point is assigned to one cluster and to one cluster only (crisp partition of the data set).

Since the minimum cannot be found directly using analytical means, an *alternating optimization* scheme is employed. At the beginning the cluster centers are initialized randomly, for example, by selecting c data points arbitrarily or by sampling c points from some distribution on the data space. Then the two steps of *partition matrix update* (data point assignment) and *cluster center update* are iterated until convergence, that is, until the cluster centers do not change anymore.

In the partition matrix update each data point $\mathbf{x}_j$ is assigned to the cluster $\mathbf{c}_i$, the center of which is closest to it, that is, the partition matrix is updated according to

$$u_{ij} = \begin{cases} 1, \text{ if } i = \operatorname{argmin}_{i=1}^{c} d_{ij}^2, \\ 0, \text{ otherwise.} \end{cases}$$

In the cluster center update each cluster center is recomputed as the mean of the data points that were assigned to it (hence the name c-means clustering), that is,

$$\mathbf{c}_i = \frac{\sum_{j=1}^{n} u_{ij} \mathbf{x}_j^2}{\sum_{j=1}^{n} u_{ij}}.$$

This update process is guaranteed to converge and usually does so after fairly few steps. However, it is fairly sensitive to the initial conditions (i.e. the initial cluster centers), due to which it can yield undesired results, which are caused by local minima of the objective function. In order to handle this drawback, it is usually recommended to execute the clustering algorithm multiple times and take the best result, that is, the result that yields the smallest value of the objective function.

In order to obtain *degrees of membership*, it may seem, at first sight, to be sufficient to simply extend the allowed range of values of the u_{ij} from the set $\{0,1\}$ to the real interval $[0,1]$, but to make no changes to the objective function itself. However, this is not the case: the optimum of the objective function is obtained for a crisp assignment, regardless of whether we enforce a crisp assignment or not.

This can easily be demonstrated as follows: let $k_j = \operatorname{argmin}_{i=1}^{c} d_{ij}^2$, that is, let k_j be the index of the cluster center closest to the data point $\mathbf{x}_j$. Then it is

$$\begin{aligned} J(\mathbf{X},\mathbf{C},\mathbf{U}) \;&=\; \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\, d_{ij}^2 \;\;\geq\;\; \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\, d_{k_j j}^2 \;=\; \sum_{j=1}^{n} d_{k_j j}^2 \underbrace{\sum_{i=1}^{c} u_{ij}}_{=1 \text{ (due to the constraints)}} \\ &=\; \sum_{j=1}^{n} \left(1 \cdot d_{k_j j}^2 + \sum_{\substack{i=1\\ i\neq k_j}}^{c} 0 \cdot d_{ij}^2 \right). \end{aligned}$$

Therefore it is best to set $\forall j; 1 \leq j \leq n : u_{k_j j} = 1$ and $u_{ij} = 0$ for $1 \leq i \leq c$, $i \neq k_j$. In other words: the objective function is minimized by assigning each data point crisply to the closest cluster, even though we allowed for degrees of membership.

4 Fuzzification by Membership Transformation

Since we cannot obtain degrees of membership by merely expanding the range of values of the u_{ij}, we have to modify the objective function if we desire graded assignments. The most common approach is to apply a transformation to the membership degrees, that is, to use an objective function of the form

$$J(\mathbf{X},\mathbf{C},\mathbf{U}) = \sum_{i=1}^{c}\sum_{j=1}^{n} h(u_{ij})\, d_{ij}^2,$$

where h is a convex function on the real interval $[0,1]$. This general form was first studied in [27], where the convexity of h was derived as follows: for simplicity, we confine ourselves to two clusters $\mathbf{c}_1$ and $\mathbf{c}_2$ and consider the terms of the objective function that refer to a single data point $\mathbf{x}_j$. That is, we consider $J(\mathbf{x}_j,\mathbf{c}_1,\mathbf{c}_2,u_{1j},u_{2j}) = h(u_{1j})\, d_{1j}^2 + h(u_{2j})\, d_{2j}^2$ and study how it behaves for different values u_{1j} and u_{2j}. Note that a crisp assignment should not be ruled out categorically, namely if the distances d_{1j} and d_{2j} differ significantly. Hence we assume that d_{1j} and d_{2j} differ only slightly, so that a graded assignment is actually desired.

$J(\mathbf{x}_j,\mathbf{c}_1,\mathbf{c}_2,u_{1j},u_{2j})$ is minimized by choosing u_{1j} and u_{2j} appropriately. Exploiting $\sum_{i=1}^{c} u_{ij} = 1$ yields $J(\mathbf{x}_j,\mathbf{c}_1,\mathbf{c}_2,u_{1j}) = h(u_{1j})\, d_{1j}^2 + h(1-u_{1j})\, d_{2j}^2$. A necessary condition for a minimum is $\frac{\partial}{\partial u_{1j}} J(\mathbf{x}_j,\mathbf{c}_1,\mathbf{c}_2,u_{1j}) = h'(u_{1j})\, d_{1j}^2 - h'(1-u_{1j})\, d_{2j}^2 = 0$, where $'$ denotes taking the derivative w.r.t. the argument of the function. This leads to $h'(u_{1j})\, d_{1j}^2 = h'(1-u_{1j})\, d_{2j}^2$, which yields another argument that a graded assignment cannot be optimal without any function h: if h is the identity, we have $h'(u_{1j}) = h'(1-u_{1j}) = 1$ and thus the equation cannot hold if the distances differ.

For the further analysis let us assume, without loss of generality, that $d_{1j} < d_{2j}$, which implies $h'(u_{1j}) > h'(1-u_{1j})$. In addition, we know that $u_{1j} > u_{2j} = 1-u_{1j}$, because the degree of membership should be higher for the cluster that is closer. In other words, the function h must be the steeper, the greater its argument. Therefore it must be a convex function on the unit interval [27].

Since we confine ourselves to the Euclidean distance (see Section 2), we can already derive the update rule for the cluster centers, namely by exploiting that a necessary condition for a minimum of the objective function J is that the partial derivatives w.r.t. the cluster centers vanish. Therefore we have $\forall k; 1 \leq k \leq c:$

$$\nabla_{\mathbf{c}_k} J(\mathbf{X},\mathbf{C},\mathbf{U}) = \nabla_{\mathbf{c}_k} \sum_{i=1}^{c}\sum_{j=1}^{n} h(u_{ij})\,(\mathbf{x}_j - \mathbf{c}_i)^\top(\mathbf{x}_j - \mathbf{c}_i) = -2\sum_{j=1}^{n} h(u_{ij})(\mathbf{x}_j - \mathbf{c}_i) \overset{!}{=} 0.$$

Independent of the function h, it follows immediately

$$\mathbf{c}_i = \frac{\sum_{j=1}^{n} h(u_{ij})\mathbf{x}_j}{\sum_{j=1}^{n} h(u_{ij})}.$$

This update rule already shows one of the core drawbacks of a fuzzification by membership transformation, namely that the transformation function enters the update of the cluster centers. It would be more intuitive to use the membership degrees directly as the weights for the mean computation, which would also ensure that all data points enter with the same total unit weight (since $\sum_{i=1}^{c} u_{ij} = 1$ by definition). However, the weights are rather the transformed membership degrees $h(u_{ij})$, which gives unequal weight to the data points as they need not sum to 1.

It may be argued, though, that this effect can actually be desirable: due to the convexity of the function h the total weight $\sum_{i=1}^{c} h(u_{ij})$ of data points $\mathbf{x}_j$ with a less ambiguous assignment is higher than that of more ambiguously assigned data points. Hence in this scheme the locations of the cluster centers depend more strongly on the data points that are "typical" for the clusters. Such an effect is very much in the spirit of, for instance, robust regression techniques, in which data points receive a lower weight if they do not fit well to the regression function. This connection to robust statistical methods was explored in more detail, for example, in [10].

In order to derive the update rule for the partition matrix (and thus for the membership degrees u_{ij}) we need to know the exact form of the function h. The most common choice is $h(u_{ij}) = u_{ij}^2$, which leads to the standard objective function of fuzzy clustering [13]. The more general form $h(u_{ij}) = u_{ij}^w$ was introduced in [2]. The exponent w, $w > 1$, is called the *fuzzifier*, since it controls the "fuzziness" of the data point assignments: the higher w, the softer the boundaries between the clusters. This leads to the commonly used objective function [2, 4, 20, 7]

$$J(\mathbf{X},\mathbf{U},\mathbf{C}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^w d_{ij}^2.$$

The update rule for the membership degrees is now derived by incorporating the constraints $\forall j; 1 \leq j \leq n : \sum_{i=1}^{c} u_{ij} = 1$ with Lagrange multipliers into the objective function. This yields the Lagrange function

$$L(\mathbf{X},\mathbf{U},\mathbf{C},\Lambda) = \underbrace{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{w} d_{ij}^{2}}_{=J(\mathbf{X},\mathbf{U},\mathbf{C})} + \sum_{j=1}^{n} \lambda_j \left(1 - \sum_{i=1}^{c} u_{ij}\right),$$

where $\Lambda = (\lambda_1, \ldots, \lambda_n)$ are the Lagrange multipliers, one per constraint.

Since a necessary condition for a minimum of the Lagrange function is that the partial derivatives w.r.t. the membership degrees vanish, we obtain

$$\frac{\partial}{\partial u_{kl}} L(\mathbf{X},\mathbf{U},\mathbf{C},\Lambda) = w\, u_{kl}^{w-1}\, d_{kl}^{2} - \lambda_l \overset{!}{=} 0 \qquad \text{and thus} \qquad u_{kl} = \left(\frac{\lambda_l}{w\, d_{kl}^{2}}\right)^{\frac{1}{w-1}}.$$

Summing these equations over the clusters (in order to be able to exploit the corresponding constraints on the membership degrees, which are recovered from the fact that it is a necessary condition for a minimum that the partial derivatives of the Lagrange function w.r.t. the Lagrange multipliers vanish), we get

$$1 = \sum_{i=1}^{c} u_{ij} = \sum_{i=1}^{c} \left(\frac{\lambda_j}{w\, d_{ij}^{2}}\right)^{\frac{1}{w-1}} \qquad \text{and thus} \qquad \lambda_j = \left(\sum_{i=1}^{c} \left(w\, d_{ij}^{2}\right)^{\frac{1}{1-w}}\right)^{1-w}.$$

Therefore we finally have for the membership degrees $\forall i; 1 \leq i \leq c$: $\forall j; 1 \leq j \leq n$:

$$u_{ij} = \frac{d_{ij}^{\frac{2}{1-w}}}{\sum_{k=1}^{c} d_{kj}^{\frac{2}{1-w}}} \qquad \text{and thus for } w = 2\text{:} \qquad u_{ij} = \frac{d_{ij}^{-2}}{\sum_{k=1}^{c} d_{kj}^{-2}}.$$

This rule is fairly intuitive, as it updates the membership degrees according to the relative inverse squared distances of the data points to the cluster centers.

However, this rule also has the disadvantage that it necessarily yields a graded assignment. Regardless of how far a data point is from a cluster center, it will always receive a non-vanishing degree of membership to the corresponding cluster. The undesirable results that can be caused by this property in the presence of clusters with fairly uneven numbers of members have been demonstrated clearly in [27].

In addition, it was revealed in [27] that the reason lies essentially in the fact that $h'(u_{ij}) = \frac{\mathrm{d}}{\mathrm{d}u_{ij}} u_{ij}^{w} = w u_{ij}^{w-1}$ vanishes at $u_{ij} = 0$. This suggests the idea to use a transformation function that does not have this property and thus allows, at least for sufficiently large distance relationships, a crisp assignment of data points to cluster centers. In [27] the function $h(u_{ij}) = \alpha u_{ij}^{2} + (1-\alpha) u_{ij}$, $\alpha \in (0,1]$, or, with a more easily interpretable parametrization, $h(u_{ij}) = \frac{1-\beta}{1+\beta} u_{ij}^{2} + \frac{2\beta}{1+\beta} u_{ij}$, $\beta \in [0,1)$, was suggested as such a transformation. It relies on the standard function $h(u_{ij}) = u_{ij}^{2}$ and mixes it with the identity to avoid a vanishing derivative at zero. The parameter β is, for two clusters, the ratio of the smaller to the larger squared distance, at and below which we get a crisp assignment [27]. It therefore takes the place of the fuzzifier w: the smaller β, the softer the boundaries between the clusters.

The update rule for the membership degrees is derived in essentially the same way as for $h(u_{ij}) = u_{ij}^w$, although one has to pay attention to the fact that crisp assignments are now possible and thus some membership degrees may vanish. The detailed derivation, which I omit here, can be found in [27] or in [7]. It yields

$$u_{ij} = \frac{u'_{ij}}{\sum_{k=1}^{c} u'_{kj}} \qquad \text{with} \qquad u'_{ij} = \max\left\{0,\ d_{ij}^{-2} - \frac{\beta}{1+\beta(c_j-1)} \sum_{k=1}^{c_j} d_{\varsigma(k)j}^{-2}\right\},$$

where $\varsigma : \{1,\ldots,c\} \to \{1,\ldots c\}$ is a mapping function for the cluster indices such that $\forall i; 1 \le i < c : d_{\varsigma(i)j} \le d_{\varsigma(i+1)j}$ (that is, ς sorts the distances ascendingly) and

$$c_j = \max\left\{k \,\middle|\, d_{\varsigma(k)j}^{-2} > \frac{\beta}{1+\beta(k-1)} \sum_{i=1}^{k} d_{\varsigma(i)j}^{-2}\right\}$$

is the number of clusters to which the data point $\mathbf{x}_j$ has a non-vanishing membership. This update rule is fairly interpretable, as it still assigns membership degrees essentially according to the relative inverse squared distances to the clusters, but subtracts an offset from them, which makes crisp assignments possible.

5 Fuzzification by Membership Regularization

We have seen that transforming the membership degrees in the objective function has the disadvantage that the transformation function appears in the update rule for the cluster centers. In order to avoid this drawback, one may try to achieve a fuzzification by leaving the membership degrees in their weighting of the (squared) distances untouched. Graded memberships are rather achieved by adding a regularization term to the objective function, which pushes the minimum away from a crisp assignment. Most commonly, the objective function then takes the form

$$J(\mathbf{X},\mathbf{C},\mathbf{U}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij} d_{ij}^2 + \gamma \sum_{i=1}^{c}\sum_{j=1}^{n} f(u_{ij}),$$

where f is a convex function on the real interval $[0,1]$. The parameter γ takes the place of the fuzzifier w: the higher γ, the softer the boundaries between the clusters.

To analyze this objective function, we use the same basic means as in the preceding section: we confine ourselves to two clusters $\mathbf{c}_1$ and $\mathbf{c}_2$ and consider the terms of the objective function that refer to a single data point $\mathbf{x}_j$, that is, we consider $J(\mathbf{x}_j,\mathbf{c}_1,\mathbf{c}_2,u_{1j},u_{2j}) = u_{1j}d_{1j}^2 + u_{2j}d_{2j}^2 + \gamma f(u_{1j}) + \gamma f(u_{2j})$. Since $u_{2j} = 1 - u_{1j}$, it is $J(\mathbf{x}_j,\mathbf{c}_1,\mathbf{c}_2,u_{1j}) = u_{1j}d_{1j}^2 + (1-u_{1j})d_{2j}^2 + \gamma f(u_{1j}) + \gamma f(1-u_{1j})$. A necessary condition for a minimum is $\frac{\partial}{\partial u_{1j}} J(\mathbf{x}_j, \mathbf{c}_1, \mathbf{c}_2, u_{1j}) = d_{1j}^2 - d_{2j}^2 + \gamma f'(u_{1j}) - \gamma f'(1 - u_{1j}) = 0$, where $'$ denotes taking the derivative w.r.t. the argument of the function. This leads to the simple condition $d_{1j}^2 + \gamma f'(u_{1j}) = d_{2j}^2 + \gamma f'(1 - u_{1j})$.

We now assume again, without loss of generality, that $d_{1j} < d_{2j}$, which implies $f'(u_{1j}) > f'(1 - u_{1j})$. In addition we know $u_{1j} > u_{2j} = 1 - u_{1j}$, because the degree of membership should be higher for the cluster that is closer. In other words, the function f must be the steeper, the greater its argument. Hence it must be a convex function on the unit interval in order to allow for graded memberships.

More concretely, we obtain $(d_{2j}^2 - d_{1j}^2)/\gamma = f'(u_{1j}) - f'(1 - u_{1j})$ as a condition for a minimum. Since f is a convex function on the unit interval, the maximum value of the right hand side is $f'(1) - f'(0)$. If $f'(1) - f'(0) < \infty$, we have the possibility of crisp assignments, because in this case there exist values for d_{1j}^2, d_{2j}^2 and γ such that the minimum of the function $J(\mathbf{x}_j, \mathbf{c}_1, \mathbf{c}_2, u_{1j})$ w.r.t. u_{ij} either does not exist or lies outside the unit interval. In such a situation the best choice is the crisp assignment $u_{1j} = 1$ and $u_{2j} = 0$ (still assuming that $d_{1j} < d_{2j}$).

To obtain the update rule for the cluster centers we can simply transfer the result from the preceding section, since the regularization term does not refer to the cluster centers. Therefore we have the simple rule (because here $h(u_{ij}) = u_{ij}$)

$$\mathbf{c}_i = \frac{\sum_{j=1}^n u_{ij} \mathbf{x}_j}{\sum_{j=1}^n u_{ij}}.$$

This demonstrates the advantage of a membership regularization approach, because the membership degrees are directly the weights with which the data points enter the mean computation that yields the new cluster center.

In order to derive the update rule for the membership degrees, we have to respect the constraints $\forall j; 1 \leq j \leq n : \sum_{i=1}^c u_{ij} = 1$. This is achieved in the usual way (cf. the preceding section) by incorporating them with Lagrange multipliers into the objective function. The resulting Lagrange function is

$$L(\mathbf{X}, \mathbf{U}, \mathbf{C}, \Lambda) = \underbrace{\sum_{i=1}^c \sum_{j=1}^n u_{ij} d_{ij}^2 + \gamma \sum_{i=1}^c \sum_{j=1}^n f(u_{ij})}_{= J(\mathbf{X}, \mathbf{C}, \mathbf{U})} + \sum_{j=1}^n \lambda_j \left(1 - \sum_{i=1}^c u_{ij}\right),$$

where $\Lambda = (\lambda_1, \ldots, \lambda_n)$ are the Lagrange multipliers, one per constraint.

Since a necessary condition for a minimum of the Lagrange function is that the partial derivatives w.r.t. the membership degrees vanish, we obtain

$$\frac{\partial}{\partial u_{kl}} L(\mathbf{X}, \mathbf{U}, \mathbf{C}) = d_{kl}^2 + \gamma f'(u_{kl}) - \lambda_l \overset{!}{=} 0 \qquad \text{and thus} \qquad u_{kl} = f'^{-1}\left(\frac{\lambda_l - d_{kl}^2}{\gamma}\right),$$

where $'$ denotes taking the derivative w.r.t. the argument of the function and f'^{-1} denotes the inverse of the derivative of the function f. In analogy to Section 4, the constraints on the membership degrees are now exploited to obtain $1 = \sum_{k=1}^c u_{kj} = \sum_{k=1}^c f'^{-1}((\lambda_j - d_{kj}^2)/\gamma)$. This equation has to be solved for λ_j and the result has to be used to substitute λ_l in the expression for the u_{kl} derived above. However, in order to do so, we need to know the exact form of the regularization function f.

The regularization functions f that have been suggested in the literature (concrete examples are studied below) can be seen as derived from a maximum entropy approach. That is, the term of the objective function that forces the u_{ij} to minimize the weighted sum of squared distances is complemented by a term that forces them to maximize the entropies of the distributions over the clusters, the u_{ij} describe for each data point. Thus the u_{ij} are pushed away from a crisp assignment, which has minimum entropy. Generally, such an approach starts from the objective function

$$J(\mathbf{X},\mathbf{C},\mathbf{U}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij} d_{ij}^2 - \gamma \sum_{j=1}^{n} H(\mathbf{u}_j),$$

where $\mathbf{u}_j = (u_{1j},\ldots,u_{cj})$ comprises the degrees of membership the data point $\mathbf{x}_j$ has to the different clusters. H computes their entropy, as $\mathbf{u}_j$ is, at least formally, a probability distribution, since it satisfies $\forall i; 1 \leq i \leq c : u_{ij} \in [0,1]$ and $\sum_{i=1}^{c} u_{ij} = 1$.

In order to develop the maximum entropy approach in more detail, we consider the generalized entropy proposed by Daróczy in [9]. Let $\mathbf{p} = (p_1,\ldots,p_r)$ be a probability distribution over r values. Then *Daróczy entropy* is defined as

$$H_\beta(\mathbf{p}) = \frac{2^{\beta-1}}{2^{\beta-1}-1}\sum_{i=1}^{r} p_i(1-p_i^{\beta-1}) = \frac{2^{\beta-1}}{2^{\beta-1}-1}\left(1-\sum_{i=1}^{r} p_i^\beta\right).$$

From this general formula the well-known *Shannon entropy* [38] can be derived as

$$H_1(\mathbf{p}) = \lim_{\beta\to 1} H_\beta(\mathbf{p}) = -\sum_{i=1}^{r} p_i \log_2 p_i.$$

Employing it in the entropy-regularized objective function leads to

$$J(\mathbf{X},\mathbf{C},\mathbf{U}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij} d_{ij}^2 + \gamma \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij} \ln u_{ij},$$

where the factor $1/\ln 2$ (which stems from the relation $\log_2 u_{ij} = \ln u_{ij}/\ln 2$) is incorporated into the factor γ, as the natural logarithm allows for easier mathematical treatment. That is, we have $f(u_{ij}) = u_{ij} \ln u_{ij}$ [25, 31, 33, 8] and therefore obtain $f'(u_{ij}) = 1 + \ln u_{ij}$ and $f'^{-1}(y) = e^{y-1}$. Using the latter in the formulas obtained above for deriving the update rule for the membership degrees yields

$$u_{ij} = \frac{e^{-d_{ij}^2/\gamma}}{\sum_{k=1}^{c} e^{-d_{kj}^2/\gamma}}.$$

As was pointed out in [35, 21], this update rule relates the approach very closely to the expectation maximization (EM) algorithm for Gaussian mixtures [11, 15, 6], since by setting $\gamma = 2\sigma^2$, we obtain exactly the formula for the expectation step. As a consequence, this update rule can be interpreted as computing the probability that a data point $\mathbf{x}_j$ was sampled from a Gaussian distribution centered at $\mathbf{c}_i$ and having the variance σ^2. In addition, since the update rule for the cluster centers coincides with the maximization step, this form of fuzzy clustering is actually indistinguishable from the expectation maximization algorithm for a mixture of Gaussians.

It should be noted that $f'(u_{ij}) = 1 + \ln u_{ij}$ implies $f'(1) - f'(0) = \infty$ and thus Shannon entropy regularization always yields graded assignments. However, this drawback is less harmful here, because $e^{-d_{ij}^2/\gamma}$ is much "steeper" than d_{ij}^{-2} and thus is less prone to produce undesired results (cf. also the discussion in [12]).

Another commonly used special case of Daróczy entropy is so-called *quadratic entropy*, which results if we set the parameter $\beta = 2$, that is,

$$H_2(\mathbf{p}) = 2\sum_{i=1}^{r} p_i(1-p_i) = 2 - 2\sum_{i=1}^{r} p_i^2.$$

Employing it in the entropy-regularized objective function leads to

$$J(\mathbf{X},\mathbf{C},\mathbf{U}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}d_{ij}^2 + \gamma\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2,$$

as the constant term 2 has no influence on the location of the minimum and thus can be discarded, and the factor 2 can be incorporated into the factor γ. That is, we have $f(u_{ij}) = u_{ij}^2$ [34] and therefore obtain $f'(u_{ij}) = 2u_{ij}$ and $f'^{-1}(y) = \frac{y}{2}$.

In order to derive the update rule for the memberships, one has to pay attention to the fact that $f'(1) - f'(0) = 2$. Therefore crisp assignments are possible and some membership degrees may vanish. However, the detailed derivation can easily be found by following, for example, the same lines as for the analogous approach in the preceding section, which also allowed for vanishing membership degrees.

The resulting membership degree update rule is $\forall i : 1 \le i \le c : \forall j : 1 \le j \le n :$

$$u_{ij} = \max\left\{0, \frac{1}{c_j}\left(1 + \sum_{k=1}^{c_j}\frac{d_{\varsigma(k)j}^2}{2\gamma}\right) - \frac{d_{ij}}{2\gamma}\right\},$$

where $\varsigma : \{1,\ldots,c\} \to \{1,\ldots c\}$ is a mapping function for the cluster indices such that $\forall i; 1 \le i < c : d_{\varsigma(i)j} \le d_{\varsigma(i+1)j}$ (that is, ς sorts the distances ascendingly) and

$$c_j = \max\left\{k \,\middle|\, \sum_{i=1}^{k} d_{\varsigma(i)j}^2 > k\,d_{kj} - 2\gamma\right\}$$

is the number of clusters to which the data point $\mathbf{x}_j$ has a non-vanishing membership. In this update rule 2γ can be interpreted as a reference distance relative to which all distances are judged. For two clusters, 2γ is the difference between the distances of a data point to the cluster centers, at and above which a crisp assignment is used. Clearly, this is equivalent to saying that the distances, if measured in 2γ units, must differ by less than 1 in order to obtain a graded assignment.

A disadvantage of this update rule is that it refers to the difference of the distances rather than their ratio, which seems more intuitive. As a consequence, a data point that has distance x to one cluster and distance y to the other is assigned in exactly the same way as a data point that has distance $x+z$ to the first cluster and distance $y+z$ to the second, regardless of the value of z (provided $z \geq -\min\{x,y\}$).

Alternatives to the discussed approaches modified the Shannon entropy term, using, for instance, $f(u_{ij}) = u_{ij} \ln u_{ij} + (1-u_{ij}) \ln(1-u_{ij})$ [42], or replaced it with Kullback-Leibler information divergence [30] to the (estimated) cluster probability distribution [22], that is, $f(u_{ij}) = u_{ij} \ln \frac{u_{ij}}{p_i}$ with $p_i = \frac{1}{n} \sum_{j=1}^{n} u_{ij}$.

It has also been tried to use $f(u_{ij}) = u_{ij}^w$ [41, 36], but combined with $h(u_{ij}) = u_{ij}^w$ (to avoid technical complications), so that the objective function is effectively

$$J(\mathbf{X}, \mathbf{C}, \mathbf{U}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^w \, (d_{ij}^2 + \gamma).$$

Hence this is actually a hybrid approach that combines membership transformation and regularization. Another hybrid approach, proposed in [40], combines $h(u_{ij}) = u_{ij}^w$ and Shannon entropy regularization $f(u_{ij}) = u_{ij} \ln u_{ij}$. Finally, a generalized objective function was presented in [5] and analyzed in more detail in [43].

It should be noted, though, that the approach of [16], which is covered by the generalized objective function of [5] and based on

$$J(\mathbf{X}, \mathbf{C}, \mathbf{U}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^w d_{ij}^2 - \gamma \sum_{i=1}^{c} p_i^2 \qquad \text{with} \qquad p_i = \frac{1}{n} \sum_{j=1}^{n} u_{ij},$$

is *not* a membership regularization scheme, as it yields crisp assignments unless $w > 1$. In this approach the entropy term (which is added rather than subtracted) serves the purpose to choose the number of clusters automatically.

A closely related approach is *possibilistic clustering* [28, 29], which eliminates the constraints $\forall j; 1 \leq j \leq n : \sum_{i=1}^{c} u_{ij} = 1$ and is based on the objective function

$$J(\mathbf{X}, \mathbf{C}, \mathbf{U}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^w d_{ij}^2 + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{n} (1-u_{ij})^w.$$

Here the η_i are suitable positive numbers (one per cluster $\mathbf{c}_i$, $1 \leq i \leq c$) that determine the distance at which the membership degree of a point to a cluster is 0.5. They are usually initialized, based on the result of a preceding run of standard fuzzy clustering, as the average fuzzy intra-cluster distance $\eta_i = \sum_{j=1}^{n} u_{ij}^w d_{ij}^2 / \sum_{j=1}^{n} u_{ij}^w$ and may or may not be updated in each iteration [28].

Although this approach is useful in certain applications, it should be noted that the objective function of possibilistic clustering is truly optimized only if all clusters are identical [39], because the missing constraints decouple the clusters. Thus it actually *requires* that the optimization process gets stuck in a local optimum in order to yield useful results, which is a somewhat strange property.

6 Conclusions

Since classical *c*-means clustering does not yield graded data point assignments, even if one allows the membership variables to take values in the unit interval, the objective function has to be modified if graded assignments are desired. There are two fundamental approaches to this: transforming the membership degrees or adding a membership regularization term. In both cases variants can be derived that allow partially crisp assignments, that is, allow for vanishing membership degrees, as well as variants that enforce graded assignments regardless of the data. All of these variants have advantages and disadvantages: membership transformation suffers generally from the fact that the transformation function enters the cluster center update, but uses a fairly intuitive relative inverse squared distance scheme for the membership updates. Quadratic entropy regularization allows for vanishing membership degrees, but refers to distance differences rather than more intuitive distance ratios. Shannon entropy regularization leads to a procedure that is equivalent to the expectation maximization (EM) algorithm for a mixture of Gaussian and thus is not a specifically "fuzzy" approach anymore. However, judging from the discussion in [12] due to which the forced graded assignment is unproblematic, its practical advantages make it, in my personal opinion, the most recommendable approach.

References

1. Ball GH and Hall DJ (1967) A Clustering Technique for Summarizing Multivariate Data. *Behavioral Science* 12(2):153–155. J. Wiley & Sons, Chichester, United Kingdom
2. Bezdek JC (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, New York, NY, USA
3. Bezdek JC and Pal N (1992) *Fuzzy Models for Pattern Recognition.* IEEE Press, New York, NY, USA
4. Bezdek JC, Keller J, Krishnapuram R, and Pal N (1999) *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing.* Kluwer, Dordrecht, Netherlands
5. Bezdek JC and Hathaway RJ (2003) Visual Cluster Validity (VCV) Displays for Prototype Generator Clustering Methods. *Proc. 12th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE 2003, Saint Louis, MO)*, 2:875-880. IEEE Press, Piscataway, NJ, USA
6. Bilmes J (1997) A Gentle Tutorial on the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. *Tech. Report ICSI-TR-97-021.* University of Berkeley, CA, USA
7. Borgelt C (2005) *Prototype-based Classification and Clustering.* Habilitationsschrift, Otto-von-Guericke-University of Magdeburg, Germany
8. Boujemaa N (2000) Generalized Competitive Clustering for Image Segmentation. *Proc. 19th Int. Meeting North American Fuzzy Information Processing Society (NAFIPS 2000, Atlanta, GA)*, 133–137. IEEE Press, Piscataway, NJ, USA
9. Daróczy Z (1970). Generalized Information Functions. *Information and Control* 16(1):36–51. Academic Press, San Diego, CA, USA
10. Davé RN and Krishnapuram R (1997) Robust Clustering Methods: A Unified View. *IEEE Trans. on Fuzzy Systems 5 (1997)*, 270–293. IEEE Press, Piscataway, NJ, USA
11. Dempster AP, Laird N, and Rubin D (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society (Series B)* 39:1–38. Blackwell, Oxford, United Kingdom

12. Döring C, Borgelt C, and Kruse R (2005) Effects of Irrelevant Attributes in Fuzzy Clustering. *Proc. 14th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE'05, Reno, NV)*, 862–866. IEEE Press, Piscataway, NJ, USA
13. Dunn JC (1973) A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics* 3(3):32–57. American Society for Cybernetics, Washington, DC, USA. Reprinted in [3], 82–101
14. Everitt BS (1981) *Cluster Analysis*. Heinemann, London, United Kingdom
15. Everitt BS and Hand DJ (1981) *Finite Mixture Distributions*. Chapman & Hall, London, United Kingdom
16. Frigui H and Krishnapuram R (1997) Clustering by Competitive Agglomeration. *Pattern Recognition* 30(7):1109–1119. Pergamon Press, Oxford, United Kingdom
17. Gath I and Geva AB (1989) Unsupervised Optimal Fuzzy Clustering. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)* 11:773–781. IEEE Press, Piscataway, NJ, USA. Reprinted in [3], 211–218
18. Gustafson EE and Kessel WC (1979) Fuzzy Clustering with a Fuzzy Covariance Matrix. *Proc. of the IEEE Conf. on Decision and Control (CDC 1979, San Diego, CA)*, 761–766. IEEE Press, Piscataway, NJ, USA. Reprinted in [3], 117–122
19. Hartigan JA and Wong MA (1979) A k-Means Clustering Algorithm. *Applied Statistics* 28:100–108. Blackwell, Oxford, United Kingdom
20. Höppner F, Klawonn F, Kruse R, and Runkler T (1999) *Fuzzy Cluster Analysis*. J. Wiley & Sons, Chichester, United Kingdom
21. Honda K and Ichihashi H (2005) Regularized Linear Fuzzy Clustering and Probabilistic PCA Mixture Models. *IEEE Trans. Fuzzy Systems* 13(4):508–516. IEEE Press, Piscataway, NJ, USA
22. Ichihashi H, Miyagishi K, and Honda K (2001) Fuzzy c-Means Clustering with Regularization by K-L Information. *Proc. 10th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE 2001, Melbourne, Australia)*, 924–927. IEEE Press, Piscataway, NJ, USA
23. Jain AK and Dubes RC (1988) *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ, USA
24. Jajuga K (2003) L_1-norm Based Fuzzy Clustering. *Fuzzy Sets and Systems* 39(1):43–50. Elsevier, Amsterdam, Netherlands
25. Karayiannis NB (1994) MECA: Maximum Entropy Clustering Algorithm. *Proc. 3rd IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE 1994, Orlando, FL)*, I:630–635. IEEE Press, Piscataway, NJ, USA
26. Kaufman L and Rousseeuw P (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. J. Wiley & Sons, New York, NY, USA
27. Klawonn F and Höppner F (2003) What is Fuzzy about Fuzzy Clustering? Understanding and Improving the Concept of the Fuzzifier. *Proc. 5th Int. Symposium on Intelligent Data Analysis (IDA 2003, Berlin, Germany)*, 254–264. Springer-Verlag, Berlin, Germany
28. Krishnapuram R and Keller JM (1993) A Possibilistic Approach to Clustering. *IEEE Trans. on Fuzzy Systems* 1(2):98–110. IEEE Press, Piscataway, NJ, USA
29. Krishnapuram R and Keller JM (1996) The Possibilistic c-Means Algorithm: Insights and Recommendations. *IEEE Trans. on Fuzzy Systems* 4(3):385–393. IEEE Press, Piscataway, NJ, USA
30. Kullback S and Leibler RA (1951) On Information and Sufficiency. *Annals of Mathematical Statistics* 22:79–86. Institute of Mathematical Statistics, Hayward, CA, USA
31. Li RP and Mukaidono M (1995) A Maximum Entropy Approach to Fuzzy Clustering. *Proc. 4th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE 1994, Yokohama, Japan)*, 2227–2232. IEEE Press, Piscataway, NJ, USA
32. Lloyd S (1982) Least Squares Quantization in PCM. *IEEE Trans. Information Theory* 28:129–137. IEEE Press, Piscataway, NJ, USA
33. Miyamoto S and Mukaidono M (1997) Fuzzy c-Means as a Regularization and Maximum Entropy Approach. *Proc. 7th Int. Fuzzy Systems Association World Congress (IFSA'97, Prague, Czech Republic)*, II:86–92

34. Miyamoto S and Umayahara K (1998) Fuzzy Clustering by Quadratic Regularization. *Proc. IEEE Int. Conf. on Fuzzy Systems/IEEE World Congress on Computational Intelligence (WCCI 1998, Anchorage, AK)*, 2:1394–1399. IEEE Press, Piscataway, NJ, USA
35. Mori Y, Honda K, Kanda A, and Ichihashi H (2003) A Unified View of Probabilistic PCA and Regularized Linear Fuzzy Clustering. *Proc. Int. Joint Conf. on Neural Networks (IJCNN 2003, Portland, OR)* I:541–546. IEEE Press, Piscataway, NJ, USA
36. Özdemir D and Akarun L (2002) A Fuzzy Algorithm for Color Quantization of Images. *Pattern Recognition* 35:1785–1791. Pergamon Press, Oxford, United Kingdom
37. Ruspini EH (1969) A New Approach to Clustering. *Information and Control* 15(1):22–32. Academic Press, San Diego, CA, USA. Reprinted in [3], 63–70
38. Shannon CE (1948) The Mathematical Theory of Communication. *The Bell System Technical Journal* 27:379–423. Bell Laboratories, Murray Hill, NJ, USA
39. Timm H, Borgelt C, Döring C, and Kruse R (2004) An Extension to Possibilistic Fuzzy Cluster Analysis. *Fuzzy Sets and Systems* 147:3–16. Elsevier Science, Amsterdam, Netherlands
40. Wei C and Fahn C (2002) The Multisynapse Neural Network and Its Application to Fuzzy Clustering. *IEEE Trans. Neural Networks* 13(3):600–618. IEEE Press, Piscataway, NJ, USA
41. Yang MS (1993) On a Class of Fuzzy Classification Maximum Likelihood Procedures. *Fuzzy Sets and Systems* 57:365–375. Elsevier, Amsterdam, Netherlands
42. Yasuda M, Furuhashi T, Matsuzaki M, and Okuma S (2001) Fuzzy Clustering using Deterministic Annealing Method and Its Statistical Mechanical Characteristics. *Proc. 10th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE 2001, Melbourne, Australia)*, 2:797–800. IEEE Press, Piscataway, NJ, USA
43. Yu J and Yang MS (2007) A Generalized Fuzzy Clustering Regularization Model With Optimality Tests and Model Complexity Analysis. *IEEE Trans. Fuzzy Systems* 15(5):904–915. IEEE Press, Piscatway, NJ, USA