Handling Noise and Outliers in Fuzzy Clustering

Christian Borgelt[1], Christian Braune[2], Marie-Jeanne Lesot[3] and Rudolf Kruse[2]

[1] European Centre for Soft Computing
Campus Mieres, c/ Gonzalo Gutiérrez Quirós s/n, 33600 Mieres, Spain
[2] Dept. Knowledge Processing and Language Engineering
Otto-von-Guericke-Universität Magdeburg
Universitätsplatz 2, 39106 Magdeburg, Germany
[3] Sorbonne Universités, UPMC Univ Paris 06,
UMR 7606, LIP6, F-75005, Paris, France

christian@borgelt.net,marie-jeanne.lesot@lip6.fr,
{christian.braune,rudolf.kruse}@ovgu.de

Abstract. Since it is an unsupervised data analysis approach, clustering relies solely on the location of the data points in the data space or, alternatively, on their relative distances or similarities. As a consequence, clustering can suffer from the presence of noisy data points and outliers, which can obscure the structure of the clusters in the data and thus may drive clustering algorithms to yield suboptimal or even misleading results. Fuzzy clustering is no exception in this respect, although it features an aspect of robustness, due to which outliers and generally data points that are atypical for the clusters in the data have a lesser influence on the cluster parameters. Starting from this aspect, we provide in this paper an overview of different approaches with which fuzzy clustering can be made less sensitive to noise and outliers and categorize them according to the component of standard fuzzy clustering they modify.

1 Introduction

In general, *clustering* [1–4] is a data analysis method that tries to group the records, cases or generally data points of a data set in such a way that points in the same group (or *cluster*) are as similar as possible, while points in different groups are as dissimilar as possible. There is no predefined target attribute (like a class label) that guides the analysis process and hence clustering belongs to the so-called *unsupervised* methods (in contrast to supervised methods like, for example, classifier construction): it relies solely on the location of the data points in the data space or, alternatively, on their relative distance or similarity.

Unfortunately, due to this exclusive dependence on location and/or distance information, clustering algorithms can suffer from noisy data points and outliers that are present in the data. Such data points, which we may define informally as points that do not conform (well) to the actual cluster structure of the data, can obscure the true cluster structure and thus may lead clustering algorithms to produce results that are far from optimal or even misleading.

Fuzzy clustering [5–7, 4] is no exception in this respect, although it features an aspect of robustness, due to which outliers and generally data points that are atypical for the clusters in the data have a lesser influence on the cluster parameters (like, for instance, the location of the cluster centers as well as shape and size parameters that may be present). We emphasize this aspect in the next section (Section 2), in which we briefly review standard fuzzy clustering.

Afterward we turn to methods that try to make fuzzy clustering (even more) robust w.r.t. noise and outliers. Such approaches can be roughly categorized into two classes: (1) approaches that modify the "(influence) weight" of the (atypical) data points, either by changing how membership degrees are computed from the (relative) data point distances to the clusters or by introducing and adapting an explicit data point weight, and (2) approaches that rely on other distance measures than the usually employed squared Euclidean distance or transform the distance measure before computing membership degrees.

Among the approaches in the first class are the popular noise cluster approach [8–10] (Section 3), introducing and adapting an explicit data point weight (outlier clustering) [11] (Section 4), possibilistic fuzzy clustering [12, 13] and its variants that combine it with standard fuzzy clustering [14–18] (Section 5), as well as using an alternative transformation of the membership degrees [19] (Section 6). In the second class we find approaches based on squared and particularly unsquared Minkowski distances [20–22] (Section 7) or transformed or otherwise modified distance measures [23–26] (Section 8).

2 Fuzzy Clustering

In the clustering approaches we study in this paper, the similarity of data points is formalized by a *distance measure* on the data space and the clusters are described by *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 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 [27–29]: simply sum the squared distances of the data points to the center of the cluster to which they are assigned. The c-means clustering 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. A solution to this problem consists in employing one of the different "fuzzifications" of the classical *crisp* (or *hard*) scheme (see, for instance, [5, 30, 6, 7, 4, 26]), which modify the objective function of classical c-means clustering in order to obtain graded cluster memberships.

In principle, there are two ways to do this, namely (1) by membership transformation, which maps the memberships with a convex function, and (2) by membership regularization, which adds a regularization term, usually derived from an entropy measure, to the objective function to prevent crisp assignments (see, for instance, [31] for a discussion). Here we focus on the first approach (membership transformation), because it exhibits a certain robustness property we are interested in. However, most of the approaches we study in Sections 3 to 8 can equally well be applied to fuzzy clustering by membership regularization.

Formally, we are given a data set $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ with n data points, each of which is an m-dimensional real-valued vector, that is, $\forall j; 1 \leq j \leq n : \boldsymbol{x}_j = (x_{j1}, \ldots, x_{jm}) \in \mathbb{R}^m$. These data points are to be grouped into c clusters, each of which is described by a prototype $\boldsymbol{c}_i$, $i = 1, \ldots, c$. The set of all prototypes is denoted by $\mathbf{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_c\}$. We confine ourselves here to cluster prototypes that consist only of a cluster center, that is, $\forall i; 1 \leq i \leq c : \boldsymbol{c}_i = (c_{i1}, \ldots, c_{im}) \in \mathbb{R}^m$, although (most of) the approaches we study below may just as well be applied if the cluster prototypes comprise shape and size parameters (like, for instance, in [32, 33]). 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 $\boldsymbol{x}_j$ belongs to cluster $\boldsymbol{c}_i$ ($u_{ij} = 1$) or not ($u_{ij} = 0$). In the fuzzy case, $u_{ij} \in [0, 1]$ states the degree to which $\boldsymbol{x}_j$ belongs to $\boldsymbol{c}_i$ (*degree of membership*).

Since we do not obtain graded memberships by merely allowing $u_{ij} \in [0, 1]$ (see, for example, [19, 31]), the membership degrees are transformed with a convex mapping $h : [0, 1] \to [0, 1]$. This yields an objective function of the form [19]

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

The clustering task now consists in finding for a given data set $\mathbf{X}$ and a user-specified number of clusters c, cluster prototypes $\mathbf{C}$ and a partition matrix $\mathbf{U}$ such that $J(\mathbf{X}, \mathbf{C}, \mathbf{U})$ is minimized under the constraints

$$\forall j; 1 \leq j \leq n : \sum_{i=1}^{c} u_{ij} = 1 \qquad \text{and} \qquad \forall i; 1 \leq i \leq c : \sum_{j=1}^{n} u_{ij} > 0.$$

Unfortunately, cluster prototypes $\mathbf{C}$ and a partition matrix $\mathbf{U}$ that minimize J are difficult to find by analytic means. Therefore one takes refuge to an *alternating optimization* scheme: starting from randomly chosen cluster centers (for example, sampled from the data set $\mathbf{X}$), one iterates (1) updating the partition matrix for fixed cluster prototypes and (2) updating the cluster prototypes for a fixed partition matrix until convergence. Convergence may be checked with a limit for the change of the cluster parameters (e.g. center coordinates) or a limit for the change of the membership degrees from one iteration to the next.

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 [30]. The more general form $h(u_{ij}) = u_{ij}^w$ was introduced in [6]. 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, while a crisp partition results in the limit for $w \to 1$. This leads to the commonly used objective function [6, 7, 4, 26]

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

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. (The second set of constraints, that is, $\forall i; 1 \leq i \leq c : \sum_{j=1}^{n} u_{ij} > 0$ can usually be neglected, because it is satisfied by the clustering result anyway.) This yields the Lagrange function

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

where $\varLambda = (\lambda_1, \ldots, \lambda_n)$ 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

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

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 $\forall i; 1 \leq i \leq c$: $\forall j; 1 \leq j \leq n$:

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

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.

In order to derive the update rule for the cluster centers, we need to know the (squared) distances d_{ij}^2. The most common choice is the (squared) Euclidean distance, that is, $d_{ij}^2 = (\boldsymbol{x}_j - \boldsymbol{c}_i)^\top (\boldsymbol{x}_j - \boldsymbol{c}_i)$. With this choice, we can easily derive

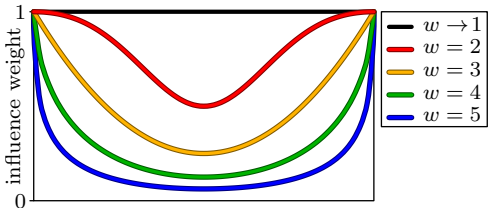

Fig. 1. "Influence weight" of a data point between two cluster centers for different values of the fuzzifier w. The two cluster centers are at the left and the right border of the diagram.

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 $\forall k; 1 \leq k \leq c:$

$$\nabla_{\boldsymbol{c}_k} J(\mathbf{X}, \mathbf{C}, \mathbf{U}) = \nabla_{\boldsymbol{c}_k} \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^w \, (\boldsymbol{x}_j - \boldsymbol{c}_i)^\top (\boldsymbol{x}_j - \boldsymbol{c}_i)$$

$$= -2 \sum_{j=1}^{n} u_{ij}^w (\boldsymbol{x}_j - \boldsymbol{c}_i) \stackrel{!}{=} 0.$$

It follows immediately $\forall i; 1 \leq i \leq c:$

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

For the topic of this paper it is important to note that this update rule draws on the *transformed* membership degrees u_{ij}^w rather than on u_{ij} directly. As a consequence the effective "influence weight" of a data point on the cluster parameters is not 1 (as one may be led to believe by the constraints $\forall j; 1 \leq j \leq n:$ $\sum_{i=1}^{c} u_{ij} = 1$), but rather $\alpha_j = \sum_{i=1}^{c} u_{ij}^w$. It is $\alpha_j = 1$ only if the data point $\boldsymbol{x}_j$ coincides with a cluster center (or if $w \to 1$); otherwise it is $\alpha_j < 1$.

As an illustration, Figure 1 shows, for $c = 2$ clusters, the influence weight of a data point lying on a straight line connecting the two cluster centers: one cluster center is at the left border of the diagram, the other at the right border. Clearly, for a fuzzifier $w > 1$ the total influence weight $\alpha_j = \sum_{i=1}^{c} u_{ij}^w$ of a data point with a less ambiguous assignment (that is, close to the left or right border of the diagram) is higher than that of a more ambiguously assigned data point (in the middle of the diagram). Also, this influence weight is the lower, the larger the fuzzifier. The minimum influence weight is always obtained for equal distances (and thus equal membership degrees $u_{ij} = 1/c$) to all c clusters. In this case the influence weight of the data point is $\alpha = \sum_{i=1}^{c} (1/c)^w = c \cdot c^{-w} = c^{1-w}$.

Note that a unit data point weight is obtained only at the cluster centers or for the limiting case of crisp clustering (that is, for $w \to 1$). This distinguishes fuzzy clustering from classical (crisp) clustering, where each data point has a unit influence (on exactly one cluster). It also distinguishes the membership transformation approach to fuzzy clustering from an approach that relies on membership regularization, since in the latter the update rule for the cluster centers refers to *untransformed* membership degrees (see, for instance, [31]), thus endowing each data point with a unit effective influence weight.

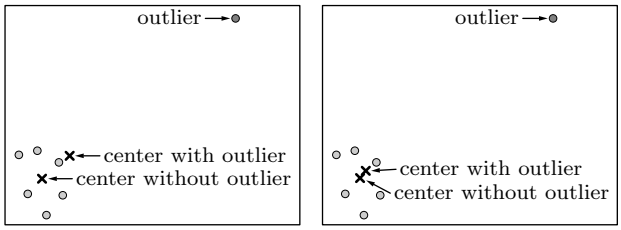

Fig. 2. Effect of an outlier on the location of a cluster center that minimizes the sum of the squared (left) and unsquared Euclidean distances (right).

Due to the reduced influence weight that ambiguously assigned data points receive in the membership transformation approach, the locations of the cluster centers depend more strongly on those data points that are "typical" for the clusters. This effect can be desirable and is very much in the spirit of, for instance, robust regression techniques, in which data points also receive a lower weight if they are not fitted well by the regression function. This connection to robust statistical methods was explored in more detail, for example, in [34, 35].

Despite this inherent robustness of fuzzy clustering, the influence of noisy data points and outliers on the clustering result can still be too strong to yield sufficiently good clustering results. A core reason for this is that the standard objective function is defined in terms of sums of *squared* Euclidean distances. Due to this squaring of distances, outliers can have an overly strong influence on the cluster parameters. This is illustrated in Figure 2 on the left, which shows six data points forming a cluster (light gray circles at the left bottom) and one outlier (dark gray circle at the top right). Computing the mean vector of the six data points forming the cluster—that is, computing the point that minimizes the sum of the *squared* Euclidean distances to the data points—yields a center vector that lies, as one would expect, in the middle of this group of data points (lower left cross). However, if the outlier is included in this mean computation, the cluster center is strongly pulled out of the cloud of the six data points towards the outlier. The reason is, of course, that the large distance to the outlier becomes even bigger by squaring and thus dominates the smaller (squared) distances to the other six data points, producing an undesirable result.

Summarizing our discussion, we see that we can try to tackle noise and outliers in fuzzy clustering in essentially two ways: (1) we can try to reduce the influence weight of atypical data points even further than the membership transformation already does (approaches based on this idea are studied in Sections 3 to 6) or (2) we can change or transform the distance measure in the objective function to reduce or eliminate the deteriorating effect of the squared distances (approaches based on this idea are studied in Sections 7 and 8).

3 Noise Clustering

The best known and most popular approach to handle noise and outliers in fuzzy clustering is so-called *noise clustering*, which was first proposed in [8], but received attention only after it was independently developed again in [9].

The core idea of this method is to introduce a pseudo-cluster, called the *noise cluster*, that has no specific location, but rather the same distance δ from all data points in $\mathbf{X}$. Thus data points that are far away from the actual clusters (in particular: farther away than the *noise distance* δ), receive a high degree of membership to the noise cluster. As a consequence, the influence of noisy data points and outliers on the parameters of the actual clusters is reduced, since the membership degrees to the actual clusters now sum to a value that is the smaller, the higher the degree of membership to the noise cluster.

Formally, this leads to the objective function [8, 9]

$$J(\mathbf{X}, \mathbf{U}, \mathbf{C}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{w} \, d_{ij}^{2} + \delta^{2} \sum_{j=1}^{n} u_{0j}^{w},$$

where the index $i = 0$ refers to the noise cluster. Of course, the first set of constraints now includes the noise cluster in the sum, that is, $\forall j; 1 \leq j \leq n : \sum_{i=0}^{c} u_{ij} = 1$. As a consequence, even the untransformed membership degrees to the actual clusters do not sum to 1 anymore, but only to $1 - u_{0j}$, where

$$\forall j; 1 \leq j \leq n : \quad u_{0j} = \frac{\delta^{\frac{2}{1-w}}}{\delta^{\frac{2}{1-w}} + \sum_{i=1}^{c} d_{ij}^{\frac{2}{1-w}}}$$

is the degree of membership of the data point $\boldsymbol{x}_j$ to the noise cluster. Clearly, this reduces the influence weight (in the sense of Section 2) of data points that are atypical for the actual clusters and thus renders the result much more robust.

Of course, introducing a noise clusters raises the question of how to choose the noise distance δ. If δ is (too) small, a large portion of the data set will receive a high degree of membership to the noise cluster, possibly rendering the majority of the data points noise and outliers. On the other hand, if δ is chosen (too) large, membership degrees to the noise cluster will remain small, possibly rendering its influence negligible [36]. A proper choice depends on many aspects [37]: the amount of noise present in the data set, the employed distance measure, the size of the feature space (in terms of the range of possible values for the distance measure), the number c of clusters to the found etc. In [9] it was suggested to compute the noise distance (in each iteration) from the (unweighted) average distance of the data points to the cluster prototypes as

$$\delta^{2} = \frac{\kappa}{nc} \sum_{i=1}^{c} \sum_{j=1}^{n} d_{ij}^{2},$$

where κ is user-specified factor that becomes the actual parameter.

An alternative to this basic approach consist in choosing the noise distance as the (average) "cluster radius" that is derived from the requirement that the sum of the hypervolumes of the clusters (as computed with this cluster radius) should equal the size of the feature space (derived, for example, from the extreme data points) [37]. A good value of the noise distance may also be determined

by trying multiple values for δ (starting at a large value and halving δ in each step), computing the fraction p of data points that have their highest degree of membership to the noise cluster (and thus may be considered as being assigned to the noise cluster), fitting the resulting points (δ, p) with a Pareto-curve $p = q\delta^{-s}$ and finding the point of this curve at which its slope is -1 [38]. Finally, a term may be added to the objective function that controls what fraction of the data points can be expected to be noise or outliers, thus rendering the method more robust against bad choices of the noise distance δ [36].

4 Data Point Weights

As we have seen in the preceding section, noise clustering relaxes the constraints $\forall j; 1 \leq j \leq n : \sum_{i=1}^{c} u_{ij} = 1$ somewhat by including the membership degree to the noise cluster, due to which the membership degrees to the actual clusters can sum to values less than 1. Alternatively, one may introduce an explicit *data point weight* and adapt this weight in the optimization process [11], an approach which is also referred to as *outlier clustering*. It permits that the membership degrees effectively sum to values less than 1 (namely to the data point weights) for atypical data points, while for very typical data points they may even sum to values larger than 1, endowing them with a greater influence on the clusters.

Outlier clustering is based on the objective function [11]

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

where v_j is the weight of the data point $\boldsymbol{x}_j$ and θ is a constant that acts on the data point weights in an analogous way as the fuzzifier w acts on the membership degrees. A typical choice is therefore $\theta = 2$ (in analogy to the fuzzifier w).

To avoid the trivial solution in which all data point weights go to infinity and thus the value of the objective function becomes zero, a constraint analogous to the constraints of the membership degrees is introduced, namely [11]

$$\sum_{j=1}^{n} v_j = v.$$

With the natural choice $v = n$, the total weight n of the n data points is redistributed to capture the typicality of the data points for the clusters. As an equally natural alternative, one may choose $v = n(1 - \rho)$, where ρ is a user estimate of the fraction of data points that are noise or outliers.

Note that the objective function contains (a function of) the reciprocal values $1/v_j$ of the data point weights, which produces exactly the desired effect: in order to minimize the objective function, large membership degrees will have to be combined with large data point weights and small membership degrees with small data point weights. Note also that with this approach the optimization scheme has three steps: (1) optimize the data point weights for fixed membership

degrees and cluster prototypes, (2) optimize the membership degrees for fixed data point weights and cluster prototypes, and finally (3) optimize the cluster prototypes for fixed data point weights and membership degrees. Finally, note that for the last step the membership degrees u_{ij} and the data point weights v_j can be combined into membership degrees $\tilde{u}_{ij}^m = u_{ij}^m / v_j^\theta$, since both values are fixed in this step. As a consequence, the update rules for the cluster parameters are not affected by using outlier clustering and hence it can also be used, for example, with shape and size parameters for the clusters (like, for instance, in [32, 33]) or other modifications of the cluster prototypes.

In order to derive the update rule for the data point weights v_j, the same approach is employed as it was demonstrated in Section 2 for the membership degrees. The constraint $\sum_{j=1}^{n} v_j = v$ is incorporated into the objective function with the help of a Lagrange multiplier. Then the fact is exploited that at the minimum of the objective function the partial derivatives w.r.t. the data point weights v_j must vanish. In this way we easily obtain [11] $\forall j; i \leq j \leq n$:

$$v_j = v \cdot \frac{\left(\sum_{i=1}^{c} u_{ij}^w \, d_{ij}^2\right)^{\frac{1}{\theta+1}}}{\sum_{k=1}^{n} \left(\sum_{i=1}^{c} u_{ik}^w \, d_{ik}^2\right)^{\frac{1}{\theta+1}}},$$

which vanishes only if all clusters collapse to a single point. Using a threshold for the data point weights v_j one may finally identify data points as outliers.

5 Possibilistic Clustering

While the two approaches studied in Sections 3 and 4 merely relax the constraints $\forall j; 1 \leq j \leq n : \sum_{i=1}^{c} u_{ij} = 1$, by (implicitly or explicitly) allowing the membership degrees to sum to values less than 1 (because the membership degree to the noise cluster is deducted or an adaptable data point weight is introduced), *possibilistic (fuzzy) clustering* [12, 13] is more radical and abandons these constraints altogether, allowing the membership degrees to sum to arbitrary values. However, this permits the trivial solution $\forall i; 1 \leq i \leq c : \forall j; 1 \leq j \leq n : u_{ij} = 0$, which obviously minimizes the objective function $J(\mathbf{X}, \mathbf{U}, \mathbf{C}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^w \, d_{ij}^2$, but, as is equally obvious, is entirely useless.

To fix this problem, a term is added to the objective function that drives the membership degrees away from zero, leading to [12]

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

Here the η_i are suitable positive numbers (one per cluster $\boldsymbol{c}_i$, $1 \leq i \leq c$) that determine the (squared) 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

$$\eta_i = \frac{\sum_{j=1}^{n} u_{ij}^w d_{ij}^2}{\sum_{j=1}^{n} u_{ij}^w}$$

and may or may not be updated in each iteration of the optimization process [12]. The membership degrees are then computed as

$$u_{ij} = \left(1 + \left(d_{ij}^2/\eta_i\right)^{\frac{1}{w-1}}\right)^{-1}.$$

It should be noted that the above objective function is truly optimized only if all clusters are identical [39], because the missing constraints decouple the clusters (as can be seen from the computation of the membership degrees). Possibilistic clustering thus actually *requires* that the optimization process gets stuck in a local optimum in order to yield useful results, which is a somewhat strange property. Although the missing constraints certainly help dealing with outliers, this property limits the usefulness of a pure possibilistic approach, although the problem may be mitigated by introducing cluster repulsion [39].

A first solution to this problem was suggested in [14], which combined possibilistic and standard fuzzy clustering, where the latter is sometimes also called *probabilistic fuzzy clustering*, because of the formal resemblance of the membership degrees of a data point to probabilities, due to the constraints $\forall j; 1 \leq j \leq n : \sum_{i=1}^{c} u_{ij} = 1$. This approach works with the objective function

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

with the usual constraint for the membership degrees u_{ij}, but the constraints $\forall i; 1 \leq i \leq c : \sum_{j=1}^{n} v_{ij} = 1$ for the possibilistic *typicality values* v_{ij}. However, it turns out that the membership degrees dominate this approach and since the typicality values depend on the number n of data points, they become very small for large data sets. As an improvement, in [15, 16] the objective function

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

was proposed, which contains a second term that is characteristic for possibilistic clustering. This leads to the usual (probabilistic) update rule for the membership degrees u_{ij}, while the possibilistic typicality values are updated with

$$v_{ij} = \left(1 + \left(b d_{ij}^2/\eta_i\right)^{\frac{1}{\kappa-1}}\right)^{-1},$$

that is, like the membership degrees in possibilistic fuzzy clustering.

A fundamentally different solution is the graded possibilistic approach presented in [17], which allows for a smooth transition between possibilistic and probabilistic fuzzy clustering. By drawing on an adequately relaxed form of the constraints $\forall j; 1 \leq j \leq n : \sum_{i=1}^{c} u_{ij} = 1$, data points can have a lower total influence weight (in the sense of Section 2), but the cluster prototypes are still coupled and (thus) the trivial solution (that is, $\forall i, j : u_{ij} = 0$) is avoided.

The class of constraints suggested in [17] is $\forall j; 1 \leq j \leq n : \sum_{i=1}^{c} u_{ij}^{[\xi]} = 1$, where $[\xi] = [\xi_*, \xi^*]$ is an interval variable, with the natural restrictions $0 \leq \xi_* \leq 1$

and $1 \leq \xi^*$. These generalized constraints are satisfied if for each j there exists a value $\xi_j \in [\xi]$ such that $\sum_{i=1}^{c} u_{ij}^{\xi_j} = 1$. Note that standard probabilistic fuzzy clustering results as a special case of this scheme for $[\xi] = [1, 1]$ and possibilistic fuzzy clustering for $[\xi] = [0, \infty]$. Note also that we may choose $\xi_* = \alpha$ and $\xi^* = \frac{1}{\alpha}$ with a single parameter $\alpha \in [0, 1]$ as a natural simplification.

With this approach the membership degrees are computed as $u_{ij} = u_{ij,\circ}/\kappa_j$, where $u_{ij,\circ}$ is a "free" or "raw" or unnormalized membership degree, as it results from standard possibilistic fuzzy clustering (see above) and [17]

$$\kappa_j = \begin{cases} \left(\sum_{i=1}^{c} u_{ij,\circ}^{1/\alpha}\right)^{\alpha} & \text{if } \sum_{i=1}^{c} u_{ij,\circ}^{1/\alpha} > 1, \\ \left(\sum_{i=1}^{c} u_{ij,\circ}^{\alpha}\right)^{1/\alpha} & \text{if } \sum_{i=1}^{c} u_{ij,\circ}^{\alpha} < 1, \\ 1 & \text{otherwise.} \end{cases}$$

An extensive discussion of several formulations of this soft transition or graded possibilistic approach to fuzzy clustering can be found in [18].

6 Alternative Transformation

A disadvantage of the standard membership transformation approach to fuzzy clustering, which relies on $h(u_{ij}) = u_{ij}^w$ (see Section 2), is that it *always* produces membership degrees. That is, regardless of how far away a data point is from a cluster center, its membership degree never vanishes. This is one of the core reasons for the negative influence of noise and outliers on fuzzy clustering results.

In order to allow some membership degrees to be zero, an alternative membership transformation was suggested in [19]: $h(u_{ij}) = \alpha u_{ij}^2 + (1-\alpha)u_{ij}$, $\alpha \in (0, 1]$, or, with a more easily interpretable parametrization, $h(u_{ij}) = \frac{1-\beta}{1+\beta} u_{ij}^2 + \frac{2\beta}{1+\beta} u_{ij}$, $\beta \in [0, 1)$. It relies on the standard transformation $h(u_{ij}) = u_{ij}^2$ 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 [19]. 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 we omit here, can be found in [19, 26]. 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 $\forall i; 1 \leq i < c : d_{\varsigma(i)j} \leq d_{\varsigma(i+1)j}$ (that is, ς sorts the distances) 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 $\boldsymbol{x}_j$ 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.

7 Unsquared Distances

Up to now we considered how fuzzy clustering can be made more robust by changing the way in which data points are assigned to the clusters. Now we turn to the more fundamental approach of changing how distances between the data points and the clusters are measured.[1] As explained in Section 2, one of the core reasons for outliers having a strong influence on the cluster parameters is the use of *squared* Euclidean distances. If we used *unsquared* Euclidean distances instead, the clustering algorithm would become much more robust w.r.t. noise and outliers. This can be seen clearly in the right diagram of Figure 2: the outlier in the top right of the diagram has a much weaker influence on the cluster center if it is computed as the point that minimizes the sum of the *unsquared* Euclidean distances to the data points. Although the center moves if the outlier is included in the computations, it stays much closer to the center computed without the outlier and remains inside the group of data points forming the cluster.

However, a disadvantage of unsquared Euclidean distances is that the standard approach of finding the cluster update rules as it was reviewed in Section 2 becomes problematic, since the square is essential for obtaining (simple) derivatives. Several solutions have been suggested to solve or circumvent this problem. In the first place, one may rely on a scheme as it was introduced for hard clustering with the c-medoids algorithm [40]: instead of computing c cluster centers in the data space that minimize the sum of the distances, one selects those c data points that have this property. This is achieved by starting with a random selection of c data points as the initial cluster centers and assigning, as in c-means clustering, each data point to the center that is closest to it. Then it is tried to improve each cluster center in turn by replacing it with a data point that is not currently a cluster center. The best replacement is chosen and then another replacement is sought for improvement. The process stops if no replacement of a cluster center reduces the sum of unsquared distances to the data points.

This c-medoids approach has been transferred to fuzzy clustering, for example, in [41] under the name "relational fuzzy c-means clustering" (RFCM) and in [3] under the name FANNY (Fuzzy Analysis). The difference between the two approaches consists merely in the fuzzifier used, which is fixed to 2 in FANNY, but can take any value greater than 1 in RFCM. An efficient version for large data sets was proposed in [42]. A combination of this scheme with the noise clustering approach studied in Section 3 was presented in [43].

The restriction that in the c-medoids approach only data points can become cluster centers can be removed by using so-called c-medians clustering [2]. Again,

[1] Note that this approach is not restricted to fuzzy clustering, but can be applied for any clustering scheme, including classical c-means clustering.

however, the problem consists in finding the c (geometric) medians that minimize the sum of the distances to the data points. This is easy only if instead of the Euclidean distance another member of the Minkowski family of distance functions, namely the L_1 distance, is used: $d_{ij} = \sum_{k=1}^{m} |c_{ik} - x_{jk}|$. In this case the medians can be determined separately in each of the m dimensions of the data space, reducing the problem to trivial statistics in one dimension.[2]

For any other member L_p, $p \geq 1$, of the Minkowski family, an iterative majorization scheme was suggested in [22]. This extends the core idea of [21], which introduced an iterative majorization scheme for squared Minkowski distances with $1 \leq p \leq 2$, after the special cases of the L_1 distance and the L_∞ distance had been studied, for example, in [20] and [44]. The approach in [22] is even more general than merely allowing unsquared distances from the Minkowski family. Rather it defines the objective function as [22]

$$J(\mathbf{X}, \mathbf{U}, \mathbf{C}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{w} \, d_{ij,p}^{2\lambda} \qquad \text{with} \qquad d_{ij,p}^{2\lambda} = \Big(\sum_{k=1}^{m} |c_{ik} - x_{jk}|^p \Big)^{\frac{2\lambda}{p}},$$

where $p \geq 1$ is the parameter that selects the member of the Minkowski family of distance functions and the parameter λ, $0 \leq \lambda \leq 1$, allows to make the clustering algorithm robust by choosing a small value for λ. For example, $p = 2$ and $\lambda = \frac{1}{2}$ specify the most interesting case of unsquared Euclidean distances.

Intuitively, the iterative majorization procedure consists in finding, for the current state of the cluster prototypes, a sufficiently simple auxiliary function majorizing the actual objective function. That is, this auxiliary function touches the objective function at the current cluster prototypes and is nowhere smaller than the objective function. Furthermore, it should be easy to find the optimum of this majorizing function, so that one can jump to this optimum in a single step, obtaining new cluster prototypes. Then a new majorizing function is constructed for the new prototypes and the process is iterated until convergence. Presenting mathematical details of this scheme is beyond of the scope of this paper, though. An interested reader is referred to [22], which provides an extensive treatment.

8 Transformed Distances

Instead of using one of the approaches discussed in the preceding section, one may also stick with the (squared or unsquared) Euclidean distance and modify the distance computation or transform the distance measure before computing the membership degrees to increase robustness. One of the most straightforward approaches in this direction is to use an ε-insensitive distance function [24]. It contains the c-medians approach that was mentioned in the preceding section as

[2] Note that computing the membership degrees remains unchanged, regardless of the distance measure and whether it is squared or not, because for this computation the cluster prototypes are fixed and thus the distances are effectively constants.

a special case (for $\varepsilon = 0$), because it employs the objective function [24]

$$J(\mathbf{X}, \mathbf{U}, \mathbf{C}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{w} \, d_{ij,\varepsilon} \qquad \text{with} \qquad d_{ij,\varepsilon} = \sum_{k=1}^{p} \max\{0, |x_{jk} - c_{ik}| - \varepsilon\},$$

where ε is the user-specified insensitivity parameter. The update rules for the membership degrees and the cluster centers can be derived in a fairly standard fashion from this objective function (using Lagrange multipliers to incorporate the constraints and partial derivatives), but as the result is mathematically somewhat involved, we do not reproduce it here, but refer an interested reader to [24].

Note that the idea of an ε-insensitive distance function is essentially to give a larger weight to typical data points, since the points in the ε-vicinity of a cluster center are assigned crisply (i.e. $u_{ij} = 1$) to this cluster center, unless such a data point has a vanishing ε-insensitive distance from multiple cluster centers, in which case each equal membership degrees to all of these clusters are chosen. Together with the employed unsquared Manhattan distance, this considerably increases the robustness of the algorithm w.r.t. noise and outliers. This effect is particularly pronounced if a larger fuzzifier is employed (compare Figure 1, even though this figure refers to squared Euclidean distances).

A more general alternative consist in exploiting the idea of robust estimators (especially M-estimators, cf. [45]) as in [25], which uses the objective function

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

where the ρ_i, $1 \leq i \leq c$, are robust symmetric positive definite functions having their minimum at 0 (with $\rho_i(d_{ij}) = d_{ij}^{w}$ as a special case). In [25] the same function ρ is used for all clusters, which is derived from Tukey's bisquare function [45]. This leads to update rules for the membership degrees, in which merely the distances are replaced by $\rho(d_{ij})$, while the cluster centers are updated with

$$\boldsymbol{c}_i = \frac{\sum_{j=1}^{n} u_{ij}^{w} f_{ij} \boldsymbol{x}_j}{\sum_{j=1}^{n} u_{ij}^{w} f_{ij}} \qquad \text{where} \qquad f_{ij} = \frac{\mathrm{d}\rho(d_{ij})}{\mathrm{d}\, d_{ij}}.$$

An even more general, but closely related approach is the *alternating cluster estimation (ACE)* scheme that was proposed in [23] (see also [4]). The idea of this approach is to abandon the requirement of an objective function that is to be optimized and from which the update rules can be derived. Rather the alternating optimization scheme is taken as the core algorithmic component, for which plausible update rules are chosen for the two steps of recomputing the membership degrees and recomputing the cluster parameters.

In its most common form, such an approach first transform the distances d_{ij} with a radial function $r : \mathbb{R} \to [0, 1]$ to obtain "free" or "raw" membership degrees $r(d_{ij})$ to the clusters. These raw membership degrees may then be normalized using, for instance, the constraints $\forall j; 1 \leq j \leq n : \sum_{i=1}^{c} u_{ij} = 1$. Typical choices for the radial functions (the name of which stems from the fact that they

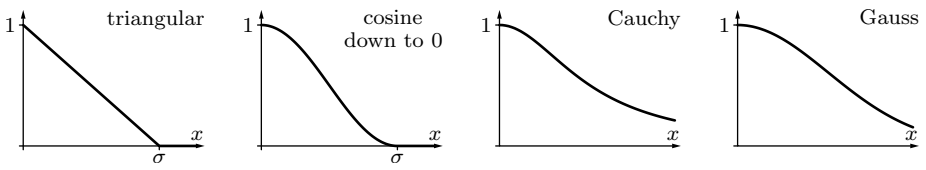

Fig. 3. Radial functions that may be used in alternating cluster estimation (ACE) [23].

are defined on a ray—latin: *radius*—from the cluster center), are shown in Figure 3. Especially those radial functions that have a finite support (that is, for which exists $x_0 \in \mathbb{R}_+$ with $\forall x > x_0 : r(x_0) = 0$) are well suited for handling noise and outliers, because data points with a distance outside the support of the radial function have a vanishing influence on the corresponding cluster.

The update rules for this scheme are simply (assuming merely cluster centers)

$$u_{ij} = \frac{r(d_{ij})}{\sum_{k=1}^{c} r(d_{kj})} \qquad \text{and} \qquad \boldsymbol{c}_i = \frac{\sum_{j=1}^{n} u_{ij}^w \boldsymbol{x}_j}{\sum_{j=1}^{n} u_{ij}^w}.$$

Generally, these update rules cannot be derived from an objective function (as shown in Section 2 for the standard case), but are merely transferred from the standard approach. It should be noted, though, that for certain radial functions, for example, the (generalized) Gaussian and the Cauchy function

$$r_{\text{Gauss}}(x) = e^{-\frac{1}{2}r^a} \qquad \text{and} \qquad r_{\text{Cauchy}}(x) = \frac{1}{x^a + b},$$

where a and b are parameters to be specified by a user, a formulation with the help of an objective function is possible, so that the needed update rules can be obtained in the usual way (using Lagrange multipliers to incorporate the constraints and setting partial derivatives equal to 0, see [26] for details).

A noteworthy alternative, which also relies on an ACE scheme instead of deriving the update equations for the membership degrees and cluster parameters from an objective function, is to compute the membership degrees as $u_{ij} = ((\max_k d_{ik}) - d_{ij}) / \max_k d_{ik}$ [46]. In this way the degree of membership of the data point that is farthest from a cluster center always vanishes, which has the additional advantage that it renders the membership degrees independent of the scale of the data set. Note that it is closely related to a possibilistic approach, because it is usually not $\forall j; 1 \leq j \leq n : \sum_{i=1}^{c} u_{ij} = 1$.

9 Summary

In this paper we reviewed several approaches to make fuzzy clustering (even) more robust against noise and outliers. The studied approaches fall into two categories: (1) reduce the "influence weight" of atypical data points and outliers on the cluster parameters by changing how membership degrees are computed

from the distances, and (2) change the distance function or transform it before the membership computation in order to reduce the degrees of memberships of atypical data points and outliers. Approaches in the former category are usually easier to handle, because in them the update rules are fairly easily obtained from an objective function using standard tools. Changing the distance measure causes more problems in this respect and thus often either a majorization approach has to be called upon or the rooting in an objective function is abandoned as in alternating cluster estimation (ACE). However, all of these approaches have the desired effect of making fuzzy clustering (even) more robust. Our personal favorites are noise clustering (see Section 3) and using an alternative membership transformation (see Section 6). However, this should not be interpreted as a recommendation against any of the other approaches.

References

1. B.S. Everitt. *Cluster Analysis.* Heinemann, London, United Kingdom 1981
2. A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, NJ, USA 1988
3. L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* J. Wiley & Sons, New York, NY, USA 1990
4. F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis.* J. Wiley & Sons, Chichester, United Kingdom 1999
5. E.H. Ruspini. A New Approach to Clustering. *Information and Control* 15(1):22–32. Academic Press, San Diego, CA, USA 1969. Reprinted in [47], 63–70
6. J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, New York, NY, USA 1981
7. J.C. Bezdek, J. Keller, R. Krishnapuram, and N. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing.* Kluwer, Dordrecht, Netherlands 1999
8. Y. Ohashi. Fuzzy Clustering and Robust Estimation. *Proc. 9th Meeting SAS Users Group Int.* Hollywood Beach, FL, USA 1984
9. R.N. Davé. Characterization and Detection of Noise in Clustering. *Pattern Recognition Letters* 12:657–664. Elsevier Science, Amsterdam, Netherlands 1991
10. R.N. Davé and S. Sen. On Generalizing the Noise Clustering Algorithms. *Proc. 7th Int. Fuzzy Systems Association World Congress (IFSA'97)*, 3:205–210. Academia, Prague, Czech Republic 1997
11. A. Keller. Fuzzy Clustering with Outliers. *Proc. 19th Conf. North American Fuzzy Information Processing Society (NAFIPS'00, Atlanta, Canada)*, 143–147. IEEE Press, Piscataway, NJ, USA 2000
12. R. Krishnapuram and J.M. Keller. A Possibilistic Approach to Clustering. *IEEE Trans. on Fuzzy Systems* 1(2):98–110. IEEE Press, Piscataway, NJ, USA 1993
13. R. Krishnapuram and J.M. Keller. The Possibilistic c-Means Algorithm: Insights and Recommendations. *IEEE Trans. on Fuzzy Systems* 4(3):385–393. IEEE Press, Piscataway, NJ, USA 1996
14. N.R. Pal, K. Pal, and J.C. Bezdek. A Mixed C-means Clustering Model. *Proc. 6th IEEE Int. Conf. on Fuzzy Systems (FUZZIEEE'97, Barcelona, Spain)*, 11–21. IEEE Press, Piscataway, NJ, USA 1997

15. N.R. Pal, K. Pal, J.M. Keller, and J.C. Bezdek. A New Hybrid C-means Clustering Model. *Proc. 13th IEEE Int. Conf. on Fuzzy Systems (FUZZIEEE'04, Budapest, Hungary)*, 179–184. IEEE Press, Piscataway, NJ, USA 2004
16. N.R. Pal, K. Pal, J.M. Keller, and J.C. Bezdek. A Possibilistic Fuzzy C-Means Clustering Algorithm. *IEEE Trans. on Fuzzy Systems* 13(4): 517–530. IEEE Press, Piscataway, NJ, USA 2005
17. F. Masulli and S. Rosetta. Soft Transition from Probabilistic to Possibilistic Fuzzy Clustering. *IEEE Trans. on Fuzzy Systems* 14(4):516–527. IEEE Press, Piscataway, NJ, USA 2006
18. K. Honda, H. Ichihashi, A. Notsu, F. Masulli and S. Rovetta. Several Formulations for Graded Possibilistic Approach to Fuzzy Clustering. *Proc. 5th Int. Conf. Rough Sets and Current Trends in Computing (RSCTC 2006, Kobe, Japan)*, 939–948. Springer-Verlag, Berlin/Heidelberg, Germany 2006
19. F. Klawonn and F. Höppner. 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 2003
20. K. Jajuga. L_1-norm Based Fuzzy Clustering. *Fuzzy Sets and Systems* 39(1):43–50. Elsevier Science, Amsterdam, Netherlands 1991
21. P.J.F. Groenen and K. Jajuga. Fuzzy Clustering with Squared Minkowski Distances. *Fuzzy Sets and Systems* 120:227–237. Elsevier Science, Amsterdam, Netherlands 2001
22. P.J.F. Groenen, U. Kaymak, and J. van Rosmalen. Fuzzy Clustering with Minkowski Distance Functions. Chapter 3 of: J. Valente de Oliveira and W. Pedrycz (eds.) *Advances in Fuzzy Clustering and Its Applications*. J. Wiley & Sons, Chichester, United Kingdom 2007.
23. T.A. Runkler and J.C. Bezdek. Alternating Cluster Estimation: A New Tool for Clustering and Function Approximation. *IEEE Trans. on Fuzzy Systems* 7(4):377–393. IEEE Press, Piscataway, NJ, USA 1999
24. J. Łęski. An ε-insensitive Approach to Fuzzy Clustering. *Int. J. Appl. Math. & Comp. Science* 11(4):993–1007. University of Zielona Góra, Poland 2001
25. H. Frigui and R. Krishnapuram. A Robust Algorithm for Automatic Extraction of an Unknown Number of Clusters from Noisy Data. *Pattern Recognition Letters* 17:1223–1232. Elsevier Science, Amsterdam, Netherlands 1996
26. C. Borgelt. *Prototype-based Classification and Clustering*. Habilitationsschrift, Otto-von-Guericke-University of Magdeburg, Germany 2005
27. G.H. Ball and D.J. Hall. A Clustering Technique for Summarizing Multivariate Data. *Behavioral Science* 12(2):153–155. J. Wiley & Sons, Chichester, United Kingdom 1967
28. J.A. Hartigan and M.A. Wong. A k-Means Clustering Algorithm. *Applied Statistics* 28:100–108. Blackwell, Oxford, United Kingdom 1979
29. S. Lloyd. Least Squares Quantization in PCM. *IEEE Trans. on Information Theory* 28:129–137. IEEE Press, Piscataway, NJ, USA 1982
30. J.C. Dunn. 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 1973 Reprinted in [47], 82–101
31. C. Borgelt. Objective Functions for Fuzzy Clustering. In: C. Moewes and A. Nürnberger (eds.) *Computational Intelligence in Intelligent Data Analysis*, 3–16. Springer-Verlag, Berlin/Heidelberg, Germany 2012

32. E.E. Gustafson and W.C. Kessel. 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 1979. Reprinted in [47], 117–122
33. I. Gath and A.B. Geva. Unsupervised Optimal Fuzzy Clustering. *IEEE on Trans. Pattern Analysis and Machine Intelligence (PAMI)* 11:773–781. IEEE Press, Piscataway, NJ, USA 1989. Reprinted in [47], 211–218
34. R.N. Davé and R. Krishnapuram. Robust Clustering Methods: A Unified View. *IEEE Trans. on Fuzzy Systems* 5:270–293. IEEE Press, Piscataway, NJ, USA 1997
35. R.N. Davé and S. Sumit. Generalized Noise Clustering as a Robust Fuzzy C-M-Estimators Model. *Proc. 17th Ann. Conf. North American Fuzzy Information Processing Society (NAFIPS'98, Pensacola Beach, Florida)*, 256–260. IEEE Press, Piscataway, NJ, USA 1998
36. F. Klawonn. Noise Clustering with a Fixed Fraction of Noise. In: A. Lotfi and M. Garibaldi (eds.) *Applications and Science in Soft Computing*, 133–138. Springer-Verlag, Berlin/Heidelberg, Germany 2004
37. F. Rehm, F. Klawonn, and R. Kruse. A Novel Approach to Noise Clustering for Outlier Detection. *Soft Computing* 11(5):489–494. Springer-Verlag, Berlin/Heidelberg, Germany 2007
38. M.G.C.A. Cimino, G. Frosini, B. Lazzerini, and F. Marcelloni. On the Noise Distance in Robust Fuzzy C-Means. *Proc. Int. Conf. on Computational Intelligence (ICCI 2004, Istanbul, Turkey)*, 361–364. Int. Comp. Intelligence Society 2004
39. H. Timm, C. Borgelt, C. Döring, and R. Kruse. An Extension to Possibilistic Fuzzy Cluster Analysis. *Fuzzy Sets and Systems* 147:3–16. Elsevier Science, Amsterdam, Netherlands 2004
40. P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. J. Wiley & Sons, Chichester, United Kingdom 1987
41. R.J. Hathaway, J.W. Devenport, and J.C. Bezdek. Relational Dual of the C-Means Clustering Algorithm. *Pattern Recognition* 22(2):205–212. Elsevier, Amsterdam, Netherlands 1989
42. R. Krishnapuram, A. Joshi, and L. Yi. A Fuzzy Relative of the K-Medoids Algorithm with Application to Document and Snippet Clustering. *Proc. 8th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE'99, Seoul, Korea)*, 3:1281–1286. IEEE Press, Piscataway, NJ, USA 1999
43. S. Sen and R.N. Dave. Clustering of Relational Data containing Noise and Outliers. *Proc. 7th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE'98, Anchorage, Alaska)*, 3:1411–1416. IEEE Press, Piscataway, NJ, USA 1998
44. L. Bobrowski and J.C. Bezdek. C-means Clustering with the L_1 and L_∞ Norms. *IEEE Trans. on Systems, Man, and Cybernetics* 21(3):545–554. IEEE Press, Piscataway, NJ, USA 1991
45. F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. J. Wiley & Sons, New York, NY, USA 1986
46. T. Binu and G. Raju. A Novel Fuzzy Clustering Method for Outlier Detection in Data Mining. *Int. J. of Recent Trends in Engineering* 1(2):161–165. Academy Publisher, British Virgin Islands, United Kingdom 2009
47. J.C. Bezdek and N.R. Pal. *Fuzzy Models for Pattern Recognition*. IEEE Press, New York, NY, USA 1992

Christian Borgelt obtained his PhD in 2000 from the University of Magdeburg, Germany, and the *venia legendi* for computer science in 2006. Since 2006 he is a principal researcher at the European Center for Soft Computing in Mieres (Asturias), Spain, where he leads the Intelligent Data Analysis and Graphical Models Research Unit. His research interests include frequent pattern mining, association rules, clustering algorithms, decision and regression trees, graphical models, neural networks as well as other soft computing, computational intelligence and intelligent data analysis methods.

Christian Braune obtained his masters degree in 2012 from the Otto-von-Guericke University, Magdeburg, Germany. Since then he is working on his PhD in the Computational Intelligence group with Rudolf Kruse. His research interests are the analysis of parallel point processes, clustering and data analysis in general.

Marie-Jeanne Lesot obtained her PhD in 2005 from the University Pierre et Marie Curie in Paris, France. Since 2006 she is an associate professor in the department of Computer Science Lab of Paris 6 (LIP6), France, and member of the Learning and Fuzzy Intelligent systems (LFI) group. Her research interests focus on fuzzy machine learning with an objective of data interpretation and semantics integration and, in particular, to model and manage subjective information; they include similarity measures, fuzzy clustering, linguistic summaries, affective computing and information scoring.

Rudolf Kruse obtained his PhD in 1980 from the University of Braunschweig, Germany, and the *venia legendi* for mathematics in 1984. Following a short stay at the Fraunhofer Gesellschaft, he joined the University of Braunschweig as a professor for computer science in 1986. Since 1996 he is a full professor for computer science at the University of Magdeburg, Germany, where he leads the computational intelligence research group. He has carried out research and projects in statistics, artificial intelligence, expert systems, fuzzy control, fuzzy data analysis, computational intelligence, and information mining. He is a fellow of the International Fuzzy Systems Association (IFSA), fellow of the European Coordinating Committee for Artificial Intelligence (ECCAI) and fellow of the Institute of Electrical and Electronics Engineers (IEEE).