Notes on Fuzzy Clustering

Collection of Notes on the Fuzzy Clustering Literature

December 13, 2016

Arash Abadpour

Arash Abadpour (arash@abadpour.com)

This is a repository of personal notes from the fuzzy clustering literature. Parts of this document have been used in different manuscripts.

Contents

Chapter 1

Notion of Membership

Historically, the clustering problem was first tackled in communication and information-theory fields. One of the first solutions to this problem was suggested for scalar quantization and is known as Lloyd algorithm [1] or Max quantizer [2]. That method was later generalized to vector quantization [3] and the resulting algorithm is generally known as the Generalized Lloyd Algorithm (GLA).

The notion of membership is a key point of distinction between different clustering frameworks. Essentially, membership may be *Hard* or *Fuzzy*. Within the context of hard membership, each data item belongs to one cluster and it is different from all other clusters. The fuzzy membership regime, however, maintains that each data item in fact belongs to all clusters, with the stipulation that the degree of membership to different clusters is different [4]. K-means [5] and Hard C-means (HCM) [6, 7] clustering algorithms, for example, utilize hard membership values. The reader is referred to [8] and the references therein for a history of K-means clustering and other methods closely related to it. Iterative Self-Organizing Data Clustering (ISODATA) [9] is a hard clustering algorithm as well. It is worth mentioning that one of the first references to k-mean clustering was in the field of Mechanics [10].

With the introduction of Fuzzy Theory [11], many researchers incorporated this more "natural" notion into clustering algorithms [12, 13, 14]. The premise for employing a fuzzy clustering algorithm is that fuzzy membership is more applicable in practical settings, where, generally, no distinct line of separation is present between the clusters [15]. In the words of the authors of [16], a fuzzy membership regime is more applicable where"a more nuanced description of the objects affinity to the specific cluster is required". Additionally, from a practical perspective, it is observed

that hard clustering techniques are extremely more prone to falling into local minima [6] (also see [17, 18, 19]). The reader is referred to [20, 21, 22] for the wide array of fuzzy clustering methods developed in the past few decades. The reader is also referred to [23] for a recent review of the field of fuzzy clustering.

Initial work on fuzzy clustering was done in the late 60s and the early 70s [24, 25, 26, 27] and it was then generalized into Fuzzy C-means (FCM) [28, 20]. In FCM, data items, which are denoted as $x_1, \cdots, x_N$, belong to $\mathbb{R}^k$ and clusters, which are identified as $\psi_1, \cdots, \psi_C$, are represented as points in $\mathbb{R}^k$. FCM makes the assumption that the number of clusters, C, is known through a separate process or expert opinion and minimizes the following objective function,

$$\Delta = \sum_{c=1}^{C} \sum_{n=1}^{N} f_{nc}^{m} \left\| x_n - \psi_c \right\|^2 . \tag{1.1}$$

This objective function is *heuristically suggested* to result in appropriate clustering results and is constrained by,

$$\sum_{c=1}^{C} f_{nc} = 1, \forall n. \tag{1.2}$$

Here, $f_{nc} \in [0, 1]$ denotes the membership of x_n to ψ_c.

We note that the fact that membership values are binary in hard clustering approaches is not just because these approaches impose the $f_{nc} \in \{0, 1\}$ constraint [29]. In fact, merely relaxing this constraint into $f_{nc} \in [0, 1]$ does not lead to a fuzzy assignment regime between the data items and the clusters [30]. Nevertheless, it is possible to achieve fuzzy membership values through utilizing a regularization term which penalizes crisp or close-to-crisp membership settings [30]. The regularization term used in that work is derived from a maximum entropy approach and it is shown [30] to yield desirable results which resemble the possibilistic clustering framework [31]. Also, in [32], the authors utilize prior membership probabilities and a Kullback-Leibler (KL) regularization term in order to produce fuzzy membership while utilizing $m = 1$ (also see [33, 34, 35, 36]). In [37] the authors utilize Deterministic Annealing (DA) in order to produce fuzzy membership values and to avoid or reduce the possibility of entrapment in local minimums. This is carried out through modeling f_{nc} as a Gibbs distribution which transforms from fully fuzzy into fully crisp as the temperature is lowered. For a thorough review of DA in the context of clustering and other related problems refer to [38]. Utilization of regularization, annealing, and similar frameworks carries with it the need to properly configure the corresponding parameters, the values of which may be problem class or even problem instance dependent.

In (1.1), $m > 1$ is the *fuzzifier* (also called *weighing exponent* and *fuzziness*). The optimal choice for the value of the fuzzifier is a debated matter [39] and is suggested to be "an open question" [40]. It has been suggested [41] that $1 < m < 5$ is a proper range, whereas $m = 2$ is a recommended choice. The use of $m = 2$ was suggested in the early works [26] on the topic as well as later works in the late 90s [42], among others [43]. Moreover, physical evidence exists [19] for the choice of $m = 2$ and it has been suggested that the best choice for m is probably in the interval $[1.5, 2.5]$ [44]. Nevertheless, other researchers argue that the choices for the value of m are mainly empirical and lack a theoretical basis [40]. That work suggests that "a proper m depends on the data set itself" [40]. Nevertheless, it is known that larger values of m soften the boundary between the clusters [30]. The reader is referred to [29] for a review of the concept of fuzzifier and alternatives for it.

Recently, a method for determining the optimal value of m in the context of FCM has been proposed [45]. That work employs four Cluster Validity Index (CVI) models and utilizes repeated clustering for $m \in [1.1, 5]$ on four synthetic data sets as well as four real data sets adopted from the UCI Machine Learning Repository [46] (refer to [47] for a review of CVIs and [48] for coverage in the context of relational clustering). The range for m in that work is based on previous research [49] which provided lower and upper bounds on m. The investigation carried in [45] yields that $m = 2.5$ and $m = 3$ are optimal in many cases and that $m = 2$ may in fact not be appropriate for an arbitrary set of data items. This result is in line with other works which demonstrate that larger values of m provide more robustness against noise and the outliers. Nevertheless, significantly large values of m are known to push the convergence towards the sample mean, in the context of Euclidean clustering [40]. Another analysis of FCM and some of its variants in the context of robustness is carried out in [50], where in recommends $m = 4$ is recommended. In [51], the authors address the related problem of fuzzy model construction. They set up the framework using $m = 2$ and proceed to find the optimal value of m for different problem instances while maintaining $m \in [1.1, 5]$. They show that different values of m are optimal within the context of different problem classes.

A suggestion has been made [52] to generalize the concept of fuzzifier and to replace f_{nc}^m with an increasing and differentiable function $g(f_{nc})$. An alternative to the classical fuzzifier function which follows their recommendation is $g(f_{nc}) = \alpha f_{nc}^2 + (1 - \alpha) f_{nc}$, for a known $0 < \alpha < 1$ [29] (also see [53]). Another contribution to the field includes the introduction of the concept of "graded possibility" [54], which modifies a model parameter similar to the fuzzifier during the operation.

An alternative approach [55, 56, 57] suggests to modify (1.2) in favor of customized $\sum f_{nc}$ con-

straints for different values of n. That technique allows for the inclusion of *a priori* information into the clustering framework and is addressed as Conditional Fuzzy C-means (CFCM). The same modification is carried out in Credibilistic Fuzzy C-Means (CFCM) [58, 59], in which the "credibility" of data items is defined based on the distances between data items and clusters. Therefore, in that approach, (1.2) is modified in order to deflate the membership of outliers to the set of clusters (also see [60]). Customization of (1.2) is also carried out in Cluster Size Insensitive FCM (csiFCM) [61] in order to moderate the impact of data items in larger clusters on an smaller adjacent cluster. Leski [39] provides a generalized version of this approach in which $\sum \beta f_{nc}^{\alpha}$ is constrained.

FCM is the groundbreaking mathematical model and the gateway into a majority of other algorithms in the literature of fuzzy clustering. In other words, a significant number of subsequent works in the field in effect modify FCM in order to alter its behavior and to achieve desired properties. For example, in FCM, a data item which is distant from all clusters, is assigned the membership level of $\frac{1}{C}$ to every cluster [62]. That membership value may in fact be relatively high for small C values. More importantly, the fact that the level of membership of an outlier to the clusters is dependent on the number of clusters is unintuitive.

Augmentation of FCM with regularization terms and constraints is not with its inherent hazards. Many works in the literature employ such techniques and as a by-product become dependent on additional parameters which need to be set properly and carefully. Often the "proper" setting of these parameters is critical for the desired function of the corresponding algorithms. For example, in [62], the authors state that the performance of their proposed method depends on five parameters which have to be "chosen through experience". The difficulty with the proper tuning of "resolution-parameter-based techniques" is reviewed in detail in [63]. The authors of [62] provide a list of some of the affected approaches.

Chapter 2

Prototype-based Clustering

It is a common assumption that the notion of homogeneity must depend on the distances between the data items. This assumption is made implicitly when clusters are modeled as *prototypical* data items, also called *clustroids* or cluster *centroids*, as in FCM, for example. A prominent choice in these works is the use of the Euclidean distance function [64]. For example, the potential function approach considers data items as energy sources scattered in a multi-dimensional space and seeks peak values in the field [65] (also see [66, 37, 67]). We argue, however, that the *distance* between the data items may not be either defined or meaningful and that what the clustering algorithm is to accomplish is the minimization of *data item-to-cluster* distances. For example, when data items are to be clustered into certain lower-dimensional subspaces, as it is the case with Fuzzy C-Varieties (FCV) [68], the Euclidean distance between the data items is irrelevant. We note that, in fact, fuzzy clustering is sometimes equated and reduced to prototype-based clustering [30] (this reductive perspective is prevalent as of 2015 [69]).

The reader is also referred to the work on the Parallel Fuzzy C-Means (PFCM) algorithm in [70], wherein fuzzy clustering is reduced to prototype-based clustering. In fact, a person not familiar with the field may conclude based on that work that fuzzy clustering is inherently and exclusively a prototype discovery mechanism. A similar reduction of fuzzy clustering to seeking prototypes is made in the design of the Dynamic Fuzzy Clustering (DFC) technique introduced in [32]. The collaborative fuzzy clustering algorithm proposed in [71] follows a similar perspective too. The reader is referred to [54] for a unified formalism of prototype-based clustering algorithms, what that paper calls the "*CM Family*", including FCM, HCM, Deterministic Annealing (DA) [37], and Possibilistic c-Means with an entropic cost term (PCM-II) [33].

Prototype-based clustering does not necessarily require prototypes which are explicitly present. For example, in kernel-based clustering [72], it is assumed that a non-Euclidean distance can be defined between any two data items. The clustering algorithm then functions based on an FCM-style objective function and produces clustroids which are defined in the same feature space as the data items [73]. These cluster prototypes may not be explicitly represented in the data item space, but, nevertheless, they share the same mathematical model as the data items [74] (the reader is referred to a review of Kernel FCM (KFCM) and Multiple-Kernel FCM (MKFCM) in [75] and several variants of KFCM in [76]). Another example for an intrinsically prototype-based clustering approach in which the prototypes are not explicitly "visible" is the Fuzzy PCA-guided Robust k-means (FPR k-means) clustering algorithm [77] in which a centroid-less formulation [78] is adopted which, nevertheless, defines homogeneity as proximity between the data items. Fuzzy Analysis (FANNY) [79] is another algorithm in which, although there are no prototypes, but, nevertheless, homogeneity is based on the mutual distances between the data items.

Relational clustering approaches constitute another class of algorithms which are intrinsically based on the distances between the data items (for example refer to Relational FCM (RFCM) [80] and its non-Euclidean extension Nerf C-means [81]). The goal of this class of algorithms is to group the data items into *self-similar* bunches. Another algorithm in which the presence of prototypes may be less evident is Multiple Prototype Fuzzy Clustering Model (FCMP) [82], in which data items are described as a linear combination of a set of prototypes, which are, nevertheless, members of the same $\mathbb{R}^k$ as the data items are. Fuzzy clustering by Local Approximation of Memberships (FLAME) [83] and Hierarchical Agglomerative Clustering (HAC) [84, 14.3.12 Hierarchical clustering] are other clustering algorithms which inherently guide the process of clustering based on the distances between the data items. The same is applicable to Visual Assessment of cluster Tendency (VAT) [85], and its variants Automated VAT (aVAT) [86] and Improved VAT (iVAT) [87], which all function based on the distances between the data items. Additionally, some researchers utilize ℓ_r-norms, for $r \neq 2$ [88, 89, 90, 91], or other distance functions which are defined between a pair of data items [92].

The use of prototype-based clustering also leads to challenges when complex notions of homogeneity are applicable to the problem class in hand. For example, in [62] the authors argue that multiple prototypes ought to be utilized when geometries other than spherical and ellipsoidal are to be addressed. They formalize that approach into the two-level Fuzzy Convex Clustering (FCC) algorithm, which is composed of consecutive FCC Expansion (FFCE) and FCC Merging (FFCM)

stages. Through that mechanism, they employ convex polytopes, which they consider to represent "flexible prototypes". The reader is referred to [21, 93] for examples of fuzzy clustering problems which involve non-Euclidean geometries.

We argue that a successful departure from the assumption of prototypical clustering is achieved when clusters and data items have different mathematical models. For example, the Gustafson-Kessel algorithm [94] models a cluster as a pair of a point and a covariance matrix and utilizes the Mahalanobis distance between data items and clusters (also see the Gath-Geva algorithm [95], Fuzzy C-Regression Models (FCRM) [96], and the improvements given in [97]). Fuzzy shell clustering algorithms [43], which are sometimes addressed as Fuzzy C-Shells (FCS), utilize more generic geometrical structures. For example, the FCV [68] algorithm can detect lines, planes, and other hyper-planar forms (also see [98, 99, 100, 101, 102, 103, 104]), the Fuzzy C Ellipsoidal Shells (FCES) [105] algorithm searches for ellipses, ellipsoids, and hyperellipsoids, and the Fuzzy C Quadric Shells (FCQS) [43] and its variants seek quadric and hyperquadric clusters (also see Fuzzy C Plano-Quadric Shells (FCPQS) [103, 104] and [106, 107, 108]).

Chapter 3

Robustification

Researchers have argued [63] that the function of membership values in FCM and the concept of weight functions in robust statistics are related. Based on this perspective, it is argued that the classical FCM in fact provides an indirect means for attempting robustness. Nevertheless, it is known that FCM and other least square methods are highly sensitive to noise [58]. In the words of the authors of [33], "It is well known that the LS analysis is severely compromised by a single outlier in the data set". The authors of [109] in fact perform a numerical review of this topic in the context of FCM and PCM and argue that the FCM constraints are "too strong" and that PCM constraints, on the other hand, are "too weak". Hence, there has been ongoing research on the possible modifications of FCM in order to provide a (more) robust clustering algorithm [110, 111]. Refer to [63] for an extensive list of relevant works and an outline of the intrinsic similarities within a unified view (also see [112, 113]).

The first attempt to robustifying FCM, based on one account [63], is the Ohashi Algorithm [112, 114]. That work adds a noise cluster to FCM and writes the robustified objective function as,

$$\Delta = \alpha \sum_{c=1}^{C} \sum_{n=1}^{N} f_{nc}^{m} \left\| x_n - \psi_c \right\|^2 + (1-\alpha) \sum_{n=1}^{N} \left(1 - \sum_{c=1}^{C} f_{nc} \right)^m . \tag{3.1}$$

The transformation from (1.1) to (3.1) was suggested independently by in [113, 115] where the Noise Clustering (NC) algorithm was proposed (also see Robust Fuzzy Clustering Algorithm (RFCA) [115]). The core idea in NC is that there exists one additional imaginary prototype which is at a fixed distance from all of the data items and represents noise. That approach is similar to modeling approaches which perform consecutive identification and deletion of one cluster at a time [116, 117]. Those methods, however, are expensive to carry out and require reliable cluster validity measures.

The core idea behind NC has been extended into the Possibilistic C-means (PCM) algorithm by rewriting the objective function as [118],

$$\Delta = \sum_{c=1}^{C}\sum_{n=1}^{N} t_{nc}^{m} \left\|x_n - \psi_c\right\|^2 + \sum_{c=1}^{C} \eta_c \sum_{n=1}^{N} (1 - t_{nc})^m. \tag{3.2}$$

Here, t_{nc} denotes the degree of representativeness or *typicality* of x_n to ψ_c (also addressed as a *possibilistic degree* in contrast to the *probabilistic* model utilized in FCM). As expected from the modification in the way t_{nc} is defined, compared to that of f_{nc}, PCM removes the sum of one constraint, shown in (1.2), and in effect extends the idea of one noise cluster in NC into C noise clusters. In other words, PCM could be considered as the parallel execution of C independent NC algorithms that each seek a cluster. Therefore, the value of C is somewhat arbitrary in PCM [63]. For this reason, PCM has been called a *mode-seeking* algorithm where C is the upper bound on the number of modes.

We argue that the interlocking mechanism present in FCM, i.e. (1.2), is valuable in that, not only clusters seek homogenous sets, but that they are also forced into more optimal "positions" through forces applied by competing clusters. In other words, borrowing the language used in [64], in FCM, clusters "seize" data items and it is disadvantageous for multiple clusters to claim high membership to the same data item. There is no phenomenon, however, in NC and PCM which corresponds to this internal factor. Additionally, it is likely that PCM clusters coincide and/or leave out portions of the data unclustered [119]. In fact, it is argued that the fact that at least some of the clusters generated through PCM are non-coincidental is because PCM gets trapped into local minimum [120] (also see [29, 30]). PCM is also known to be more sensitive to initialization and the exact values of the configuration parameters than other algorithms in its class [119, 64].

PCM dislodges the cluster model used in FCM and converts a set of C clusters which are interconnected through the $\sum_{c=1}^{C} f_{nc} = 1$ constraint into C independent objective functions. In other words, PCM completely deliberately releases the clusters from the pressure of their peers and, as stated by the authors of an investigation on PCM [33], while FCM is "primarily a partitioning algorithm", PCM is "primarily a mode seeking algorithm". Nevertheless, this transformation inflicts upon PCM the issue that multiple clusters may be discovered simultaneously [119] and as stated by the developers of PCM "[PCM's] weakness is that it requires a good initialization" [33]. Additionally, while FCM treats m as a non-significant parameter, PCM is dependent on the proper choice of the fuzzifier, and it is stated to malfunction in the case of $m = 2$ and require $m = 1.5$, because PCM requires a "faster decay" than FCM [33]. As stated by the authors of [33], "the value of m needs

to be chosen so that the decay rate in the membership function is meaningful". Additionally, PCM needs "proper initialization" [33] to the extent that it is recommended that an instance of FCM is executed in order to "provide a reasonable initialization" [33] for PCM. While this additional cost is a concern as the size of the problem grows, the suitability of FCM as an initialization for PCM is also stated to only be applicable when "data is not severely contaminated" [33].

It is important, however, to point out the major contribution of PCM over FCM. In fact, "[t]he power of the PCM does not lie in creating partitions but rather in finding meaningful clusters as defined by dense regions" [33]. From this perspective, the fact that PCM may generate "coincident clusters" [33] is stated to be a "a blessing in disguise" [33], because it allows PCM to function as a set of C truly independent cluster discovery mechanisms. Hence, "even if the true value of C is unknown, the outcome of the algorithm will be useful" [33]. Therefore, "the algorithm can potentially find C good clusters from a data set that may have more than C clusters" [33]. Additionally, "if the data set has less than C clusters, then the algorithm can still potentially find C good clusters, out of which some of them may be identical" [33]. These are indeed very important properties and position PCM in a superior state compared to FCM. This important aspect of PCM, however, is overshadowed by its extreme dependence on initialization and selection of the parameters. Additionally, not only PCM provides no mitigation procedure, but also it lacks any mechanism for recognizing that it has generated duplicate clusters. Therefore, we agree with the authors of [33] that "the PCM approach has the potential for solving one of the major problems with the FCM, namely, the need to know the number of clusters", but we argue that for this potential to be realized PCM needs to evolve further. One of the attempts in materializing the potential present in PCM is the progressive cluster discovery mechanism employed in [121]. In that work, data items which are classified into a discovered cluster are then removed from the set.

It has been argued that both concepts of possibilistic degrees and membership values have positive contributions to the purpose of clustering [122, 123]. As a result, FCM and PCM have been combined through rewriting the optimization function of Fuzzy Possibilistic C-Means (FPCM) as minimizing [122],

$$\Delta = \sum_{c=1}^{C} \sum_{n=1}^{N} \left(f_{nc}^{m} + t_{nc}^{\eta} \right) \left\| x_n - \psi_c \right\|^2, \tag{3.3}$$

subject to (1.2) and $\sum_{n=1}^{N} t_{nc} = 1, \forall c$. That approach was later shown to suffer from different scales for f_{nc} and t_{nc} values, especially when $N \gg C$, and, therefore, additional linear coefficients and a PCM-style term were introduced to the objective function [124]. It has been argued that

the resulting objective function employs four correlated parameters and that the optimal choice for them for a particular problem instance may not be trivial [64]. Additionally, in the new combined form, f_{nc} cannot necessarily be interpreted as a membership value [64]. The reader is also referred to Possibilistic-Fuzzy C-Means (PFCM) [109] for a related model which "hybridizes" PCM and FCM in order to "avoid various problems of PCM, FCM, and FPCM". Additionally, see [125, 126] for other variants and [31] for an added entropy-style regularization term.

The introduction of the PCM model is motivated by several factors, amongst which is to be able to relax, or somewhat circumvent, the sum-of-one constraint for the membership values. As such, through "giving up the requirement for strict partitioning" [54], the expectation is that the resulting algorithm will be able to reject outliers and to deal with data items which do not belong to any of the clusters more efficiently. As discussed here, however, the utilization of the PCM-style models has given rise to the emergence of other difficulties. In this context, the model presented in [127] is worth particular attention. That work states that the relationship between the data items and the clusters must be assessed at two levels, i.e. whether or not a data item is an outlier and, if not, which clusters it belongs to. In other words, the model developed in [127] replaces the singleton membership identifiers f_{nc} with the pair of p_n and f_{nc}. Here, p_n models the probability that x_n is an inlier and f_{nc} models the probability that it belongs to ψ_c, given that it is an inlier. An extension of this model in [128] demonstrates that the p_n variables can emerge from the f_{nc} values when the classical parallel clustering framework is converted to a serially-structured pipeline.

Weight modeling is an alternative robustification technique and is exemplified in the algorithm developed by Keller [129], in which the objective function is rewritten as,

$$\Delta = \sum_{c=1}^{C} \sum_{n=1}^{N} f_{nc}^{m} u_c \frac{1}{\omega_n^q} \left\| x_n - \psi_c \right\|^2, \tag{3.4}$$

subject to $\sum_{n=1}^{N} \omega_n = \omega$. Here, the values of ω_n are updated during the process as well.

Additionally, researchers included a robust loss function in the objective function of FCM and developed Robust C-Prototypes (RCP) [42],

$$\Delta = \sum_{c=1}^{C} \sum_{n=1}^{N} f_{nc}^{m} u_c \left(\left\| x_n - \psi_c \right\| \right). \tag{3.5}$$

Here, $u_c(\cdot)$ is the robust loss function for cluster c. They further extended RCP and developed an unsupervised version of RCP, nicknamed URCP [42]. Alternative HCM (AHCM) and Alternative FCM (AFCM) algorithms utilize this idea using $u_c(x) = 1 - e^{-\beta x^2}$ [74] (also see [130]). It is worth to mention that in the absence of a robust loss function, the transition in membership values in

the boundary between adjacent clusters is governed by the fuzzifier [29]. Hence, one is encouraged to reduce m in order to reduce the unwanted ambiguity in the interface between different clusters. Nevertheless, smaller values of m are known to correspond to hard formulation which are more sensitive to local minimums. The reader is also referred to [131, Sections 10 and 11] for the use of the Gauss-Newton method with a robust loss function.

In [111], the authors provide a taxonomy of the mitigation strategies proposed in the literature in order to address the presence of outliers in the context of fuzzy clustering of non-fuzzy data. *Possibilistic approaches* are the first robustification technique that they discuss. Those approaches utilize the Possibilistic Theory [132] and deflate the membership of outliers to the clusters [118, 31, 120, 60, 133]. The *noise approach*, on the other hand, employs one of the several formalizations of the notion of the noise cluster [113, 63, 134, 135]. Moreover, *metric approaches* utilize data item to cluster distance functions with robust properties [91, 74]. A related technique involves the use of robust loss functions and reformulates the cost function corresponding to a fuzzy clustering problem into M, R, or L estimators [136]. In essence, in the so-called *order statistics* approach, an appropriate loss function transforms the distance function into a robust entity (the reader is referred to [42, 63] for examples of this technique). Nevertheless, some of the techniques mentioned above result in a situation which the authors of [111] address as the *influence weighting approach*. In that approach, a weighting system, which may organically emerge within the model or may be imposed on it exogenously, associates lower weights to outlier data items as they take part in the recalculation of the optimal clusters (for example refer to [129, 137]). The *trimmed approach* and the *semifuzzy approach*, on the other hand, utilize heuristic techniques which are believed to result in robustification. As such, in the former approach, an outlier detection and elimination process "cleanses" the data before or during clustering [138, 139, 140, 141, 142]. In the latter approach, however, a data item which attempts to become a member of too many clusters or for which the membership to many clusters is below a certain threshold is given special treatment [143]. Utilizing this terminology, the method developed in this paper employs means which can be categorized under a combination of possibilistic approaches, the noise approach, order statistics, and the influence weighting approach.

Chapter 4

Number of Clusters

The classical FCM and PCM, and many of their variants, are based on the assumption that the number of clusters is known (an extensive review of this topic is given in [20, Chapter 4]). While PCM-style formulations may appear to relax this requirement, the corresponding modification is carried out at the cost of yielding an ill-posed optimization problem [64]. In fact, repeating the clustering for different numbers of clusters [95, 144] and *Progressive Clustering* are two of the alternative approaches to address the challenge of not requiring *a priori* knowledge about the number of clusters present in a particular data.

Among the many variants of Progressive Clustering are methods which start with a significantly large number of clusters and freeze "good" clusters [144, 145, 103, 104], approaches which combine compatible clusters [146, 144, 145, 103, 104, 42], and the technique of searching for one "good" cluster at a time until no more is found [116]. These approaches utilize one or more CVIs in order to assess the appropriateness of the clusters produced after each execution of the algorithm. For a review of CVIs in the context of relational clustering refer to [48] (also see [47]). Use of regularization terms in order to push the clustering results towards the "appropriate" number of clusters is another approach taken in the literature [147]. These regularization terms, however, generally involve additional parameters which are to be set carefully, and potentially per problem instance [122].

The conclusion of an important 1997 paper on the topic is that the solution to the general problem of robust clustering when the number of clusters is unknown is "elusive" and that the techniques available in the literature each have their limitations [63]. In this paper, we acknowledge that the problem of determining the appropriate number of clusters is hard to solve and even hard

to formalize. Additionally, we argue that this challenge is equally applicable to many clustering problems independent of the particular clustering models utilized in the algorithms. Therefore, we designate this challenge as being outside the scope of this contribution and assume that either the appropriate number of clusters is known or that an exogenous means of cluster pruning is available which can be utilized within the context of the algorithm developed in this paper (refer to [95] for an early review of this approach). The reader is referred to the category of method collectively known as Visual Assessment of cluster Tendency (VAT) [85]. More recent variants of VAT include Automated VAT (aVAT) [86] and Improved VAT (iVAT) [87].

Chapter 5

Weber Problem

In 1929, Weber published a pioneering work on finding an optimal solution to a problem which is now commonly known by his name [148]. In its modern form [149], the *Weber Problem* can be written as minimizing,

$$\Delta(\psi) = \sum_{n=1}^{N} \omega_n \phi(x_n, \psi). \tag{5.1}$$

Here, $\omega_n > 0$ are the weights and $\phi(\cdot)$ is a known function. When $\phi(\cdot)$ denotes the Euclidean distance, one of the most popular solution strategies to the Weber Problem is proposed by Weiszfeld [150] (refer to [151] for an accelerated version).

A generic approach to solving (5.1) is to confirm that $\Delta(\psi)$ is convex and then to pursue with the derivative of $\Delta(\cdot)$ relative to ψ. This process generally results in the presence of $\phi'(x_n, \psi)$ in the update equation of an iterative process, for which guaranteeing convergence is often cumbersome [152]. For example, the analysis given in [153] does not contain a proof of convergence and suggests to monitor the value of $\Delta(\psi)$ during the iterations and to restrict the rate of change in ψ. While this operation is costly, its generalization to problems in which ψ is governed by a complex mathematical model, as opposed to $\psi \in \mathbb{R}^k$, is non-trivial. Other works in the field consider ℓ_p norms and discuss local and global convergence [154, 155] as well as recommend acceleration techniques [156, 157] (also see Iteratively Reweighted Least Squares (IRLS) [158]). The reader is referred to [159, Chapter 4.5] and the references therein for further review of the topic and different options for numerical calculation.

Other approaches to solving the Weber Problem in more general settings, utilize regularization and other similar numerical techniques in order to suppress the amount of change in ψ in consecutive iterations. As such, the progression of the algorithm is maintained at a balance between a gradient

descent path with smaller sizes, which is slow but guaranteed to converge, and a Newton's method, which is faster but may diverge. The reader is referred to Tikhonov-Arsenin [160], Levenberg [161], and Marquardt [162] algorithms for mathematical details (also see [163]). A very recent review of the different incarnations of the Weber Problem and an outline of Weiszfeld's work can be found in [164].

From a theoretical perspective, the Weber Problem is a special case of the problem addressed in this paper when $C = 1$. Nevertheless, the two functions $\Delta_{\mathbf{X}}(\psi)$ and $\Delta_{\mathbf{\Psi}}(x)$, which are defined and used in this work, have structural similarities to generalized Weber Problems. Therefore, some of the tools developed for solving generalized Weber Problems are applicable to the method developed in this paper.

Chapter 6

Weighted Clustering

Many fuzzy and possibilistic clustering algorithms make the assumption that the data items are equally important. Weighted fuzzy clustering, however, works on input data items which have an associated positive weight [116]. This notion can be considered as a marginal case of clustering fuzzy data [165, 166]. Other examples for this setting include clustering of a weighted set, clustering of sampled data, clustering in the presence of multiple classes of data items with different priorities [167], and a measure used in order to speed up the execution through data reduction [168, 169, 170, 171].

A formalization of this situation, wherein weights are manipulated in order to move the clustering results towards data items which are harder to include regularly is given in [172]. Similarly, researchers have utilized density motivated weights in order to reduce the impact of outliers [173] (refer to [174] for different variants of this framework). Semi Supervised FCM (ssFCM) [175] uses weight factors based on an Euclidean norm in order to balance the sizes of different hyper-spherical shaped clusters based on user intervention. The reader is referred to [176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188] and the references therein for other fuzzy clustering approaches in the context of fuzzy input data items.

Note that the extension of FCM on weighted sets has been developed under different names, including Density-Weighted FCM (WFCM) [170], Fuzzy Weighted C-means (FWCM) [189], and New Weighted FCM (NW-FCM) [190]. A more thorough review of fuzzy clustering in the context of fuzzy data can be found in [111, 191] and the references therein.

Chapter 7

Spatial context

FCM effectively ignores spatial context [192]. In other words, in FCM, PCM, and many other variants of fuzzy clustering algorithms, datums are treated as separate realizations, where there is no *a priori* relationship between x_n and x_{n+1}. Nevertheless, in the physical world, datums are always correlated. Hence, the notion of spatial context suggests that a datum is defined in its context and must be classified while the context is taken into consideration. In words of the authors of [192], "usually, one pixel is too small to represent part of an image".

It is important to emphasize that the notion of spatial context and the premise of NC are somewhat contradictory. While NC attempts to classify noisy datums into a separate cluster and to subsequently discard them, the thinking behind spatial context is that a noisy datum may be classified based on its context [192]. This is an important requirement in approaches such as image segmentation, in which a pixel identified as noise is in effect a discontinuity in the output.

While it is theoretically possible to include datum coordinates, if applicable, as additional features and then to carry out clustering [193, 144], that approach is not theoretically justifiable, because coordinate information and datum features are often inherently different and are defined in different scales. From a practical perspective, too, defining a notion of homogeneity which encompasses both datum homogeneity as well as spatial contiguity for datums is not a trivial task.

Another primitive approach to including the spatial context into the clustering process is to perform data pre-processing [194, 195, 168, 76, 192, 169]. Potential loss of details, however, is among the caveats of this technique. A marginally more appropriate approach is to perform post-processing on the membership maps [196, 197, 198, 199, 195] or to execute clustering at different scales and to fuse the results afterwards [200, 201]. The reader is referred to [202, 203, 204] for other

variations of these approaches. It has been argued [205] that the incorporation of spatial context as a pre- or post-processing stage is easy to implement but lacks proper theoretical justification.

Spatial context can also be incorporated into the optimization process as additional information. In fact, Gibbs random fields have been used in order to model spatial context within the framework of K-means clustering [206, 207, 208, 209]. Also see [210] for an iterative process which utilizes sliding windows which shrink over time. Another, more recent, example for this approach is Geometrically Guided FCM (GG-FCM) [211, 212] in which the semi-supervised framework developed in [213] is modified in order to use neighborhood information as training for image pixels (also see Spatial FCM (sFCM) [214] and Bilateral FCM (bFCM) [215]).

GG-FCM and Geometrically Guided Conditional FCM (GGC-FCM) utilize a reject class and eliminate datums which are found to be spurious [212]. That approach is not desirable in applications which require every datum to be classified into a cluster. That deficiency, however, is addressed in Spatially Guided FCM (SG-FCM) [216], in which a geometrical shape descriptor is incorporated into the objective function. Nevertheless, in those works, spatial context is either utilized as a static input [213] or as information which is dynamically recalculated through the process [216]. The latter case, however, commonly depends on engineered measures of compliance to spatial contiguity and often additionally requires the proper setting of one or more regularization coefficients. For example, the approaches outlined in [213, 216] depend on the proper setting of the value of the parameter α.

Another example for the utilization of spatial context as *a priori* information is the Improved FCM (IFCM) [171], in which a histogram-based FCM deployment produces cluster prototypes and membership values and then the resulting crisp membership information is utilized in order to produce the p_{nc} quantities for each datum and each cluster. Here, p_{nc} denotes the ratio of the neighbors of x_n which belong to ψ_c. The second stage of that algorithm then finds f_{nc} and ψ_c which minimize the following modified objective function,

$$\Delta = \sum_{c=1}^{C} \sum_{n=1}^{N} p_{nc}^{m} f_{nc}^{m} \left\| x_n - \psi_c \right\|^2 . \tag{7.1}$$

That framework effectively utilizes spatial context as static information which is injected into the objective function at some point during the process.

The aforementioned works belong to the general category of approaches which utilize engineered regularization terms. Many of the works in that category propose a superficially constructed term which penalizes excessive spatial variation. Additionally, as stated before, the regularization terms

are generally multiplied by a constant, the proper setting of the value of which is an important prerequisite for the appropriateness of the outcomes of the algorithm (as an example containing both issues refer to [217]). An early examples of that approach is the Contig-k-means [218, 219] algorithm which updates the crisp k-means clustering objective function in order to incorporate spatial contiguity into it. That approach, however, requires the proper adjustment of the value of the parameter λ. In another example, in [220], a term is devised which penalizes correlation between membership values of adjacent datums to different clusters. Another example is presented in [221], in which the authors inject a regularization term into the objective function of a Kernel FCM (KFCM) algorithm [222] and generate the Spatially constrained Kernel FCM (SKFCM) algorithm. That approach of modeling the spatial context has precedence in the literature [63, 223, 76, 224, 225, 169] and is inspired by Neighborhood EM (NEM) [226].

In NEM, the proper value for the parameter α is to be set by the user and is suggested to be dependent on the signal to noise ratio (SNR) of the input image. Dependence on additional parameters is a concern in other works as well. For example, in [192] the authors develop the Improved FCM (IFCM) algorithm wherein "neighborhood attraction" is modeled as a combination of "feature attraction" and "distance attraction". That model is a reminiscent of notions from bilateral filtering literature [227]. Nevertheless, the term which utilizes spatial context in that work depends on two parameters, the values of which are to be set by on an Artificial Neural Network (ANN) (also see [214]). Additionally, the attraction models utilized in [192] depend on datum-to-datum distances, which, as previously noted, are in fact irrelevant to a generic homogeneity model.

A more generalized approach to spatial context is utilized in Bias-Corrected FCM (BCFCM) [194], in which the following regularization term is employed,

$$\frac{\alpha}{\|\mathbf{S}_n\|}\sum_{c=1}^{C}\sum_{n=1}^{N}\left[f_{nc}^{m}\sum_{n'\in\mathbf{S}_n}\left\|x_{n'}-\psi_c\right\|^2\right]. \tag{7.2}$$

In that framework, α inversely depends on the SNR of the input image and its proper value is to be set through a separate process. An accelerated and robustified variant of that framework is given in [76] and its combination with [228] is proposed in [229]. A variant of that approach is utilized in Enhanced FCM (EnFCM) [168], in which acceleration is achieved through utilizing the image histogram. That work, as well as the Fast Generalized FCM (FGFCM) [169], require the proper setting of two parameters which govern the tradeoff between the original image and a filtered version of it. The effects of those parameters are described as "crucial" and "experience" or

"trial-and-error" are stated to be required in order for them to be properly selected [230]. Similar conditions are applicable to [76, 168], among other works.

In [231] the authors develop the Adaptive FCM (AFCM) algorithm through incorporating a multiplier field into the objective function as follows (also see [228, 232, 233]),

$$\Delta = \sum_{c=1}^{C}\sum_{n=1}^{N} f_{nc}^2 \left\| x_n - s_n\psi_c \right\|^2 + \lambda_1 \sum_{n=1}^{N} \Big[\left(\Delta_x s_n\right)^2 + \left(\Delta_y s_n\right)^2 \Big] + \lambda_2 \sum_{n=1}^{N} \Big[\left(\Delta_{xx} s_n\right)^2 + 2\left(\Delta_{xy} s_n\right)^2 + \left(\Delta_{yy} s_n\right)^2 \Big]. \tag{7.3}$$

Here, the Δ terms indicate forward difference operators and λ_is are regularization coefficients. In effect, the two regularization terms in (7.3) constrain the s_n field into a smooth surface. In addition to the difficulties of optimizing the AFCM objective function in terms of s_n, as outlined in [231], it is important to emphasize that (7.3) is essentially an Euclidean prototype-based framework and that the generalized formulation for generic datum and cluster models is not trivial.

Fuzzy Local Information C-Means (FLICM) [230] is among the state-of-the-art in the field of image segmentation. The formulation of FLICM does not depend on any particular parameters and it uses fuzzy local similarity measures which incorporate both gray level as well as spatial closeness. We argue, however, that the notion of datum-to-datum comparison which is used in FLICM, is only applicable to a certain category of problem classes which includes gray scale image segmentation but excludes color image segmentation. Moreover, the primary concern with FLICM is that it engineers a new concept, i.e. *the Fuzzy Factor G*. In fact, that work is a prime example of the introduction of a new concept based on intuition, as it is outlined in the list of "characteristics" given in [230, Section III.A]. That concept is then heuristically composed as a mathematical formula [230, (17)]. That paper then follows with verbal justification of the appropriateness of the engineered factor [230, Section III.B]. While there are important epistemological questions regarding the construction of FLICM, that framework is further extended in subsequent frameworks such as RFLICM [234] and KWFLICM [235]. An extension of FLICM is given in FCM with Edge and Local Information (FELICM) [236], in which results at image edges are improved through separate treatment of boundary pixels.

Chapter 8

The VAT Literature

The thinking process behind some of the earliest attempts at the estimation of C, is one of "find many clusters, select only a few" [237]. In this line of thinking, the clustering algorithm is allowed to generate many cluster *candidates* and significant resources have been allocated to the study of cluster validity assessment techniques. The reader is referred to a comprehensive review of this after-the-fact approach in [238] (also see Dunn's index [26], the DB index [239], and the PBM index [240]). Nevertheless, a majority of those approaches make the assumption that the validity of a cluster, or an entire clustering solution, can be measured using a scalar value. However, thorough the examination of 23 scalar measures of cluster validity, researchers have shown that "*none* of them are exceptionally reliable across a wide range of datasets" [241] (also see [242] for a Monte Carlo evaluation of 30 different validity indexes). From a theoretical perspective, too, it has been argued that scalar cluster validity indexes aggregate the entire information available in an input set of data items into one or a few metrics and that invaluable information is lost in this process. In the words of the authors of [243], "scalar measures of cluster validity are famously unreliable". The reader is referred to [244] for a review of after-the-fact approaches to cluster validity assessment and to [245, 47, 246, 247, 22] for a few representative techniques.

Nevertheless, in comparison between before-the-fact techniques, which estimate the "correct" number of clusters, and after-the-fact approaches, which validate one or a set of clusters, practical implications lean towards the former. In effect, one is inclined, if possible, to execute the clustering process with the proper settings, as opposed to moving ahead with some settings when there is the likelihood that the results are likely to need to be discarded. That situation is most drastic when it is suggested that the clustering algorithm is to be executed for a range of number of clusters in

order for the most optimal solution to be picked later.

The findings regarding the inherent deficiencies of scalar clustering validity indexes have encouraged the community to investigate alternative techniques which visualize [248, 249] or assess [250, 193] the inherent structure of a given set of data items from the vantage point of clustering. It is important to emphasize that some of those techniques utilize generic statistical methods and are sometimes expensive to carry out or require marginal probability distributions which are hard or impossible to generate in practical settings.

Visual Assessment of clustering Tendency (VAT) [85] utilizes the pairwise dissimilarity information between data items as a symmetrical matrix with non-negative elements and zero diagonals. We address this matrix as the Dissimilarity Matrix (DM) (this matrix has also been called the Dissimilarity Image (DI)). VAT provides a mechanism for reordering the rows and columns of this matrix in a way that signifies the structure in the data. In short, when VAT is successful, dense areas in the dataset yield dark squares along the diagonal of the Reordered Dissimilarity Matrix (RDM). One can, therefore, at least theoretically, count these squares and generate a reasonable estimate of the number of clusters present in the data. We note that VAT must be reviewed in the context of other visualization techniques such as trees, dendrograms, castles, and icicles [251]. More specifically, VAT belongs to the subset of techniques which utilize image-based visualizations [252, 253].

An implementation of the core technique of VAT can be found in [254] (for details refer to [255]). The VAT reordering algorithm is based on Prim's algorithm [256] for finding the MST corresponding to a weighted graph (also see [257]). Nevertheless, the intent of VAT is not to generate the tree, but to produce the order in which the vertexes are added as the tree grows. Enhanced VAT (E-VAT) [258] is a variant of VAT which applies a robust loss function on the DM before the reordering process begins in order to limit the impact of the outliers. E-VAT uses Otsu's threshold [259] for the elements of the DM as scale.

The roots of VAT can be traced back to the method known as SHADE [260] (also see [261]). SHADE uses over-striking of printed characters in order to generate a halftone display. Aside from this visualization technique, which is a direct product of the display technologies available at the time of the publication of that paper, SHADE is different from VAT in that it is essentially a cluster visualization method which is utilized after another hierarchical clustering scheme is applied on the data. Additionally, SHADE only generates the lower triangular section of the DM. A variant of SHADE, addressed as the "graphical method of shading" [262], performs quantization of the DM

prior to rearrangement. VAT is also related to the Single Linkage (SL) [263] algorithm. SL, in essence, cuts the largest edges in the MST, thus producing subsets of the data, each of which corresponds to an individual cluster. The reader is referred to [264] for a review of the different variations of SL algorithms and to [265] for a review of the "direct relation" [265] between the clusters generated by SL and the reordering prescribed by VAT. The reader is also referred to [266] for a list of practical utilizations of SL as well as a list of clustering algorithms which utilize SL. Nevertheless, it has been argued that SL "famously" [267] fails for some cluster types [21]. Moreover, it has been suggested that the technique of using a shaded matrix for assessing the structure of the data generates a "visual puzzle" which is a "crypto-graphical mystery" [251, 268].

In a broader perspective, the core concept of VAT has been known for many decades and has been utilized in many different contexts. For two examples, the reader is referred to GENLAB [269] and Homogeneity Map [270, 271]. The former is an online tool for visualizing arbitrary DMs and the latter seeks temporal homogeneity in fMRI data. In effect, VAT belongs to the general category known as Cluster Heat Maps [272]. In that paper [272], the authors find traces of the utilization of the display mechanism employed by VAT in late 19th century [273, 274] and a diverse presence in the statistical literature in the 20th century [275]. The reader is referred to [272] for a long list of related references. In fact, Cluster Heat Map techniques are suggested to be "by far the most popular graphical representation" [276]. Moreover, Seriation approaches are among the well-known methods which reorder data items based on some notion of dissimilarity. The reader is referred to [277] for a list of loss functions used in the context of seriation and [278] for a comprehensive review of different Dissimilarity Plots.

While VAT-style algorithms utilize symmetrical square dissimilarity matrices, there are applications in which the rows and the columns of the DM refer to different mathematical objects. An extension of VAT, nicknamed coVAT [279], has been developed for this category of problems. co-VAT essentially treats the rectangular DM as the known section of a larger square DM and then *imputes* the missing values using a transitive logic. That technique is extended in further works, including Scalable co-VAT (scoVAT) [280], coVAT2 [281] and Improved co-VAT (co-iVAT) [87].

Review of the artificial samples utilized in a number of works which involve VAT (see for example [85, 243, 267, 282]) indicates that there are no outliers in these sets of input data items. In fact, VAT is fundamentally based on the assumption that the input set of data items is composed of a number of compact well-separated isotropic clusters. As stated by the authors of [86], VAT is "usually only effective at highlighting cluster tendency in data sets that contain compact well-separated

clusters". However, as also highlighted by the authors of the same paper, "[m]any practical applications involve data sets with highly irregular structure, which invalidate this assumption" [86]. We argue that the need for working with what is sometimes called "irregular geometries" [86] or "composite shape(s)" [283] is an important practical requirement. In fact, we suggest that what is often addressed as "tough cases" [87] is in fact *the* situation that needs to be addressed by a clustering algorithm. These are cases, for which, in words of the authors of [87], "VAT fails to accurately show the cluster tendency".

Another one of the many successors of VAT is the Visual Cluster Validity (VCV) approach outlined in [243]. In fact, VCV inherits from and build upon both SHADE and VAT and contributes to them the utilization of a generic cluster model. This is as opposed to the assumption by VAT that a cluster can be represented by a typical data item. Nevertheless, VCV, like SHADE, is a tool which displays the clusters that have been produced by another "outsourced" clustering algorithm. In other words, VCV transforms clusters, that are potentially defined in a hyper-dimensional geometry, into representations which can be visualized using a 2D matrix. Nevertheless, the way VCV approaches this capability is through cluster-to-cluster comparison, which is carried out based on the Euclidean distance between the parameters which define the cluster model. Hence, for example, in the context of an ellipsoidal geometry, VCV will combine the components of the mean vector with the elements of the covariance matrix. We argue that this approach mixes into a pot elements which not only accept values at very different scales, but also, and more importantly, are different from a theoretical perspective. Visual Cluster Validity (VCV2) [284] is a similar approach which uses image matching in order to compare cluster representations with the RDM.

Nevertheless, VCV utilizes the context of the data items when it estimates the dissimilarity between a pair of data items. In order to do so, VCV models the dissimilarity between two data items as the minimum of their common distance to a set of clusters. This process provides an important contribution over the simplified models which utilize explicit data item-to-data item distances as their level of dissimilarity, as, for example, is carried out by VAT and a majority of its variants. However, assessment of the performance of VCV for Euclidean and linear clusters in $\mathbb{R}^2$ in [243] shows that the results corresponding to linear clusters "are not nearly as clean as" [243] those for Euclidean clusters. Additionally, VCV employs an overestimated number of clusters in order to recognize the structure of the data and "some deterioration" [243] is observed when the clustering algorithms is executed using cluster counts which are "large" [243]. We note that the problem classes utilized in [243] are the equivalents of the 2de and 2dl models employed in this

paper. We also emphasize that the sets of data items which are utilized in [243] appear to not contain outlier data items.

Revised VAT (reVAT) [285] replaces the RDM with a set of profile graphs. That technique allows for removing data items which appear to be highly likely to belong to the same cluster from the pool of data and has been called "quasi-ordering" as well. That naming convention is in contrast with the full reordering carried out in VAT. Nevertheless, the output of reVAT is of a different form, compared to VAT, and, therefore, as stated by the developers of reVAT, "interpretation of the set of reVAT profile graphs becomes very difficult when the number of clusters is large, or there is significant overlap between groups of objects in the data" [267]. In fact, it has been argued [267] that reVAT is applicable when $C \leq 5$. This deficiency is alleviated in bigVAT [267], in which the visual form of the VAT output is re-instituted, but the DM is confined to samples which are proportionally selected from the dominant reVAT profiles.

Another sequel to VAT is Scalable VAT (sVAT) [282]. The development of sVAT is the result of the realization that although bigVAT produces a 2D visualization, but that "this image may not be as descriptive as a VAT-ordered image" [282]. Nevertheless, sVAT, too, performs sampling on the data in order to achieve a lower computational complexity. In a similar fashion to bigVAT, sVAT also utilizes an estimate of the clusters present in the input set of data items in order to produce a representative subset. However, in addition to the requirement that the full DM is provided by the user, sVAT relies on the user for providing the "desired (approximate) sample size" and an "overestimate of the true number of clusters". An extension of sVAT, nicknamed sVAT-SL, in which SL stands for Single Linkage, is suggested in [286]. sVAT-SL attempts to produce both the cluster representations as well as C. Nevertheless, sVAT-SL is in fact interrupted at the middle of the process, because "the user must choose the number of clusters ... to seek" [286]. In other words, the user of sVAT-SL is requested to observe the RDM and to decide what C ought to be. Nevertheless, while sVAT-SL is advertised as "an approximation to single-linkage clustering for big data" [286], the same paper also asserts that the clustering results generated by sVAT-SL "can be ruined by outliers" [286].

It has been argued that "a major limitation" [287] of VAT and its variants is their "inability to highlight cluster structure ... when ... [the data] contains clusters with highly complex structure" [287]. Spectral VAT (SpecVAT) [287] attempts to increase the legibility of the RDM generated by VAT algorithms through spectral decomposition of the DM prior to the reordering [288]. However, the performance of SpecVAT depends on the proper selection of the parameter k, i.e. the

number of eigenvectors used during decomposition. In fact, as demonstrated in [86], to properly select k, one ought to have a proper estimate for C, the number of clusters present in the data. The critical necessity for the existence of this *a priori* piece of information violates the premise behind VAT and SpecVAT. The reader is referred to [289] for a review of spectral clustering and to [290] for multiple variants of the SpecVAT scheme. The latter work utilizes image processing techniques on the RDM and employs a sampling framework in order to reduce the computational complexity of the resulting algorithm. It also attempts several ideas for estimating the parameters which govern the process in order to resolve some of the concerns with the original SpecVAT. The user is also referred to the graph-based variant of E-VAT, Graph-based E-VAT (GE-VAT) [291].

In spite of their differences, VAT, reVAT, bigVAT, sVAT, and many other algorithms in their class, are in fact *visual assessment* methods. In other words, these techniques are inherently reliant on the *subjective* [267] understanding of the user. This issue becomes more disconcerting when it is argued that in some practical settings, an "experienced user" [267] may have to be employed in order to perform the assessment (the reader is referred to [283] for related remarks). In fact, it has been argued that non-Euclidean geometries or overlap between clusters can give rise to VAT images for which "different viewers may deduce different numbers of clusters ..., or worse, not be able to estimate c at all" [86]. To make matters more complicated, the output of many of the algorithms in this class is a two dimensional image, which requires to be transferred and displayed *diligently* and it is argued that compression, down-scaling, and interleaving of this image "may obscure important information about potential clusters in the data" [267].

Another important detail about VAT and a majority of its affiliates is their treatment of the MST. In fact, in many of the algorithms in this category, the MST is in essence considered as a by-product which is discarded as soon as it is created. We argue, however, that this representation carries very important information about the structure of the data and the relationship between the clusters present in it. As will be shown later in this paper, we argue that the RDM is an alternative, and less useful, representation than the MST. Nevertheless, significant effort has been spent in the literature on the interpretation of the structure of the data based on the appearance of the RDM. Cluster Count Extraction (CCE) [292] is one such approach, which assesses the histogram of the DM using image processing operators such as Otsu's algorithm [259] and Fast Fourier Transform (FFT). CCE requires the proper adjustment of multiple threshold and filter radius parameters. Dark Block Extraction (DBE) [293] is another approach which counts the number of dark diagonal blocks in the RDM using thresholding and morphological operations. DBE also requires the diligent

adjustment of the threshold value used for determining the peaks of a histogram. Enhanced DBE (E-DBE) [291] incorporates a robust loss function with BDE. aVAT [86] is another approach which utilizes image processing techniques in order to interpret the RDM. That paper applies a function, which can be considered a robust loss function, on the RDM in order to increase its contrast and to facilitate the recognition process. The reader is also referred to [294] in which Dunn's cluster validity index is used in order to perform thresholding on the RDM. Other validity indexes, such as the PBM index [240], have been used for interpreting the RDM as well. Variable string length Genetic Algorithm (VGA) [295, 296] is another approach employed in order to extract the number of clusters from the RDM (also see [297]). Clustering in Ordered Dissimilarity Data (CLODD) [298] attempts to automatically produce the clusters from the RDM as well. While it is advertised that "when [the input data] has "good" clusters, CLODD will find them", the performance of CLODD depends on the values of the two "influence constant"s α and γ, the values of which are set using trial and error [298].

It is important to emphasize that, independent of the actual efficacy of the available approaches which utilize image processing tools in order to interpret the RDM, those techniques are in essence based on the assumption that the RDM is *the* representation to be processed and that the MST is *inferior* to it. This is evident in the concluding statement in [86], in which the authors advocate for the use of more efficient thresholding techniques [299] in order to enhance the interpretation of the RDM. We argue, however, that a more important consideration in this discussion is the choice of representations, and that the MST is a *more* informative representation than the RDM.

For an input set of data items $\mathbf{X}$, which contains N data items, the computational complexity of VAT is of $O(N^2)$. This is due to the fact that VAT processes the $N \times N$ matrix of dissimilarities between the data items. In addition to the fact that this model is inappropriate in the context of a generic notion of homogeneity, the quadratic complexity of VAT is prohibitive for Big Data problem instances. In fact, it has been argued that VAT "works well for relatively small data sets ($n \leq 500$)" [267]. Hence, variations of VAT have been proposed which address this issue. reVAT, bigVAT, and sVAT are three approaches which aim at lowering the computational complexity to $O(CN)$. Here, C is the inherent, and *unknown*, number of clusters which are present in $\mathbf{X}$. As commonly $C \ll N$, this transformation is greatly beneficial. Nevertheless, it has been argued that "sVAT does not asymptotically scale linearly with [the number of data items]" [286].

The $O(N^2)$ computational complexity of VAT contains the superpositions of the costs associated with two processes, both of which require $O(N^2)$ operations. In fact, not only the DM needs to be

reordered at the cost of $O(N^2)$ operations, but also this matrix needs to be calculated in the first place, and the computational complexity of that process is of $O(N^2)$ as well. In this context, the sequels of VAT, i.e. reVAT, bigVAT, and sVAT, drop the complexity of the reordering mechanism to $O(CN)$, but, nevertheless, they still require the calculation of the $N \times N$ DM. Hence, technically, the computational complexities of these algorithms are still $O(N^2)$, unless the DM is provided *a priori*. As will be discussed next, that circumstance is only applicable to a small subset of clustering problems, in which case, too, the DM must be calculated in many practical settings anyways. Hence, neither VAT nor the aforementioned variants of it are *genuinely* scalable. Here, we rely on the notion of scalability which mandates that the computational complexity of the algorithm must grow linearly as the number of input data items increases [300].

Nevertheless, a critical concern with VAT, and other models in its category, is in their core assumption about the nature of the inter-relationship between the data items. In fact, VAT is confined to situations in which it is meaningful to discuss the similarity/dissimilarity of two data items in vacuum. In this context, VAT makes the extremely limiting assumption that the data items are either given as object vectors or by numerical pairwise dissimilarity values [85]. Hence, VAT assumes that given the two data items x_1 and x_2, one can meaningfully point out the extent to which the two are likely to belong to the same cluster. This assumption is valid within the context of the Euclidean distance function as well as other ℓ_k norms in which case the clusters are in essence prototypical data items (the reader is referred to classical reviews of data item-to-data item distance metrics in [301, 302]).

In fact, in the literature, the input set of data items has been, *incorrectly* in our opinion, reduced to and assumed to be equivalent to its relational representation (for recent examples refer to [265, 86]). This limited scope is more obvious in the case of algorithms such as Relational Visual Cluster Validity (RVCV) [303] and Correlation Cluster Validity (CCV) [304], which explicitly identify their narrow scope. While that situation is applicable to relational and prototype-based clustering problem classes, it is not necessarily an inclusive framework. We argue that the notion of "dissimilarity between two data items" may not be meaningful for many problem classes which are highly relevant in practice, unless a context is defined. In other words, the question is *not* how dissimilar the two data items x_1 and x_2 are. One example to illuminate this point is clustering members of a given $\mathbb{R}^k$ into lower dimensional spaces. Under this regime, if data items are to be clustered into lines, planes, and hyperplanes, for example, the question of dissimilarity between two data items is *meaningless* unless one also provides a cluster representation. In other words,

any pair of data items x_1 and x_2 may be extremely similar or inherently dissimilar given their mutual relationship to the cluster which provides the context. Hence, we argue, the square DM representation is only applicable to a small subset of possible problem classes. The reader is referred to SpecVAT [287], as only one example of, the "natural" and implicit reduction of a set of data items to its relational representation.

VAT approaches have been augmented with path-based distance models in order to alleviate the limiting scope mandated by the assumption of the relational model. For example, Improved VAT (iVAT) [86] prescribes that it is not the direct data item-to-data item distance which must be stated in the DM but that x_1 and x_2 are "similar" if there is a sequence of data items, with x_1 and x_2 at the two ends of the sequence, each of which are at close distances to each other [305]. Sample experimental results provided in [86] suggests that iVAT is capable of recognizing arbitrary sequences of data items in $\mathbb{R}^2$. The authors of that paper also propose the Automatic VAT (aVAT) technique which applies a function, that can be considered a robust loss function, on the RDM in order to facilitate the semi-unsupervised extraction of the diagonal blocks from it. The reader is also referred to a variant of the iVAT algorithm, nicknamed Efficient iVAT (efiVAT), which utilizes dynamic programming [87]. A related algorithm, named clusiVAT [266], samples the input set of data items in order to generate the cluster representations using SL. clusiVAT then extends the classification results to the entire set of data items. The reader is referred to [306] for variants of clusiVAT and iVAT, nicknamed clusiVAT+ and iVAT+, which utilize additional "efficient thresholding schemes". While these contributions provide some relief, especially when the clusters are isolated and separated, the core problem is still far from being addressed by these techniques. One scenario in which the path-based model falls short of resolving the challenge is the clustering of members of $\mathbb{R}^k$ into lower-dimensional spaces. In this scenario, any intersection between a pair of clusters is a real threat that may cause the two clusters to "leak" into each other. This case is closely related to the "zigzagging" phenomenon which the original implementation of VAT was prone to and was alleviated using a clever initialization procedure for the reordering mechanism [85].

Moreover, even the mere assumption that the data items can necessarily be reduced to members of $\mathbb{R}^k$ is in essence a dangerous reduction. As will be shown later in this paper, the data items merely need to be *mathematical objects* for which *distance to a given cluster* can be defined. This is a major step forward compared to the majority of the works in the VAT literature, where the assumption that $x_n \in \mathbb{R}^k$ is made out of convenience, without any reference to the implications of that reduction.

There is a significant overlap between the group of individuals who worked on VAT and Fuzzy C-Means (FCM) [20]. While, historically, FCM precedes VAT, VAT is in fact an attempt to set one of the key inputs required by FCM, i.e. the value of C. Additionally, as stated above, there have been multiple attempt to forge VAT into an alternative to FCM, i.e. to generate the cluster representations, and not just their count, by VAT. This convoluted relationship between VAT and FCM, however, to the best of our knowledge, has never led to a marriage of the two ideas. In other words, VAT either feeds into FCM or replaces it, but never intertwines with it. In this paper, we take up this task and demonstrate that a robustified variant of FCM, nicknamed *Connie* [127], can be used as a single-cluster clustering mechanism, i.e. by assuming $C = 1$, in order to scan the cluster space and to produce cluster representations which are aggregated using a VAT reordering process in order to produce a MST, the sub-trees of which yield the sought for clusters in the input set of data items.

Chapter 9

Other

Not cited here [307, 308, 309, 310, 311, 312, 313, 314, 315].

Bibliography

[1] S. P. Lloyd, Least squares quantization in PCM, IEEE Transactions on Information Theory 28 (2) (1982) 129–137.

[2] J. Max, Quantizing for minimum distortion, IRE Transactions on Information Theory 6 (1) (1960) 7–12.

[3] Y. Linde, A. Buzo, R. M. Gray, An algorithm for vector quantizer design, IEEE Transactions on Communications 28 (1) (1980) 84–95.

[4] J. V. de Oliveira, W. Pedrycz, Advances in Fuzzy Clustering and its Applications, Wiley, 2007.

[5] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, 1967, pp. 281–297.

[6] R. Duda, P. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.

[7] R. Gray, Y. Linde, Vector quantizers and predictive quantizers for Gauss-Markov sources, IEEE Transactions on Communications 30 (2) (1982) 381–389.

[8] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, A. Y. Wu, An efficient k-means clustering algorithm: Analysis and implementation, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (7) (2002) 881–892.

[9] G. H. Ball, D. J. Hall, A clustering technique for summarizing multivariate data, Behavioral Science 12 (2) (1967) 153–155.

[10] H. Steinhaus, Sur la division des corp materiels en parties, Bulletin de l'Academie Polonaise des Sciences 1 (1956) 801–804.

[11] L. A. Zadeh, Fuzzy sets, Information Control 8 (1965) 338–353.

[12] M.-S. Yang, A survey of fuzzy clustering, Mathematical and Computer Modelling 18 (11) (1993) 1–16.

[13] A. Baraldi, P. Blonda, A survey of fuzzy clustering algorithms for pattern recognition. I, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 29 (6) (1999) 778–785.

[14] A. Baraldi, P. Blonda, A survey of fuzzy clustering algorithms for pattern recognition. II, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 29 (6) (1999) 786–801.

[15] H. Cheng, J.-R. Chen, J. Li, Threshold selection based on fuzzy c-partition entropy approach, Pattern Recognition 31 (7) (1998) 857–870.

[16] B. Abu-Jamous, R. Fa, A. K. Nandi, Fuzzy Clustering, John Wiley & Sons, Ltd, 2015, pp. 167–179.

[17] J. C. Bezdek, Numerical taxonomy with fuzzy sets, Journal of Mathematical Biology 1 (1) (1974) 57–71.

[18] J. C. Bezdek, J. Dunn, Optimal fuzzy partitions: A heuristic for estimating the parameters in a mixture of normal distributions, IEEE Transactions on Computers C-24 (8) (1975) 835–838.

[19] J. C. Bezdek, A physical interpretation of fuzzy ISODATA, IEEE Transactions on Systems, Man and Cybernetics SMC-6 (5) (1976) 387–389.

[20] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[21] J. C. Bezdek, J. M. Keller, R. Krishnapuram, N. R. Pal, Fuzzy Models and Algorithms for Pattern Recognition and Image Processing, Kluwer Academic Publishers, Boston, 1999.

[22] R. Xu, D. Wunsch, Survey of clustering algorithms, IEEE Transactions on Neural Networks 16 (3) (2005) 645–678.

[23] S. Miyamoto, Fuzzy clustering - Basic ideas and overview, in: J. Kacprzyk, W. Pedrycz (Eds.), Springer Handbook of Computational Intelligence, Springer Berlin Heidelberg, 2015, pp. 239–248.

[24] E. H. Ruspini, A new approach to clustering, Information & Control 15 (1) (1969) 22–32.

[25] E. H. Ruspini, Numerical methods for fuzzy clustering, Information Sciences 2 (3) (1970) 319–350.

[26] J. C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Journal of Cybernetics 3 (3) (1973) 32–57.

[27] J. C. Dunn, Well separated clusters and optimal fuzzy partitions, Journal of Cybernetics 4 (1974) 95–104.

[28] J. C. Bezdek, Fuzzy mathematics in pattern classification, Ph.D. thesis, Cornell University (1973).

[29] F. Klawonn, F. Hoppner, What is fuzzy about fuzzy clustering? Understanding and improving the concept of the fuzzifier, in: M. R. Berthold, H.-J. Lenz, E. Bradley, R. Kruse, C. Borgelt (Eds.), Advances in Intelligent Data Analysis V, Vol. 2810 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2003, pp. 254–264.

[30] C. Borgelt, Objective functions for fuzzy clustering, in: C. Moewes, A. Nurnberger (Eds.), Computational Intelligence in Intelligent Data Analysis, Vol. 445 of Studies in Computational Intelligence, Springer Berlin Heidelberg, 2013, pp. 3–16.

[31] R. Krishnapuram, J. M. Keller, The possibilistic C-means algorithms: Insights and recommendations, IEEE Transactions on Fuzzy Systems 4 (3) (1996) 385–393.

[32] T. M. Nguyen, Q. Wu, Dynamic fuzzy clustering and its application in motion segmentation, IEEE Transactions on Fuzzy Systems 21 (6) (2013) 1019–1031.

[33] R. Krishnapuram, J. M. Keller, The possibilistic C-means algorithm: insights and recommendations, IEEE Transactions on Fuzzy Systems 4 (3) (1996) 385–393.

[34] H. Ichihashi, K. Miyagishi, K. Honda, Fuzzy c-means clustering with regularization by K-L information, in: Proceedings of The 10th IEEE International Conference on Fuzzy Systems, Vol. 2, 2001, pp. 924–927.

[35] S. Chatzis, T. Varvarigou, A fuzzy clustering approach toward Hidden Markov random field models for enhanced spatially constrained image segmentation, Fuzzy Systems, IEEE Transactions on 16 (5) (2008) 1351–1361.

[36] S. Chatzis, T. Varvarigou, Factor analysis latent subspace modeling and robust fuzzy clustering using t-distributions, IEEE Transactions on Fuzzy Systems 17 (3) (2009) 505–517.

[37] K. Rose, E. Gurewitz, G. Fox, A deterministic annealing approach to clustering, Pattern Recognition Letters 11 (9) (1990) 589–594.

[38] K. Rose, Deterministic annealing for clustering, compression, classification, regression, and related optimization problems, Proceedings of the IEEE 86 (11) (1998) 2210–2239.

[39] J. M. Leski, Generalized weighted conditional fuzzy clustering, IEEE Transactions on Fuzzy Systems 11 (6) (2003) 709–715.

[40] J. Yu, Q. Cheng, H. Huang, Analysis of the weighting exponent in the FCM, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 34 (1) (2004) 634–639.

[41] M. Trivedi, J. C. Bezdek, Low-level segmentation of aerial images with fuzzy clustering, IEEE Transactions on Systems, Man, and Cybernetics 16 (4) (1986) 589–598.

[42] H. Frigui, R. Krishnapuram, A robust algorithm for automatic extraction of an unknown number of clusters from noisy data, Pattern Recognition Letters 17 (12) (1996) 1223–1232.

[43] F. Klawonn, R. Kruse, H. Timm, Fuzzy shell cluster analysis, in: G. della Riccia, H. Lenz, R. Kruse (Eds.), Learning, networks and statistics, Springer, 1997, pp. 105–120.

[44] N. R. Pal, J. C. Bezdek, On cluster validity for the fuzzy C-means model, IEEE Transactions on Fuzzy Systems 3 (3) (1995) 370–379.

[45] K. Zhou, C. Fu, S. L. Yang, Fuzziness parameter selection in fuzzy c-means: The perspective of cluster validation, Science China Information Sciences 57 (11) (2014) 1–8.

[46] M. Lichman, UCI machine learning repository (2013).
URL `http://archive.ics.uci.edu/ml`

[47] J. C. Bezdek, N. R. Pal, Some new indexes of cluster validity, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 28 (3) (1998) 301–315.

[48] I. Sledge, J. C. Bezdek, T. C. Havens, J. M. Keller, Relational generalizations of cluster validity indices, IEEE Transactions on Fuzzy Systems 18 (4) (2010) 771–786.

[49] I. Ozkan, I. Turksen, Upper and lower values for the level of fuzziness in FCM, in: P. P. Wang, D. Ruan, E. E. Kerre (Eds.), Fuzzy Logic, Vol. 215 of Studies in Fuzziness and Soft Computing, Springer Berlin Heidelberg, 2007, pp. 99–112.

[50] K.-L. Wu, Analysis of parameter selections for fuzzy c-means, Pattern Recognition 45 (1) (2012) 407–415.

[51] W. Pedrycz, H. Izakian, Cluster-centric fuzzy modeling, IEEE Transactions on Fuzzy Systems 22 (6) (2014) 1585–1597.

[52] F. Klawonn, Fuzzy clustering: Insights and a new approach, Mathware and soft computing 11 (2004) 125–142.

[53] P. J. Rousseeuw, E. Trauwaert, L. Kaufman, Fuzzy clustering with high contrast, Journal of Computational and Applied Mathematics 64 (1-2) (1995) 81–90.

[54] F. Masulli, S. Rovetta, Soft transition from probabilistic to possibilistic fuzzy clustering, IEEE Transactions on Fuzzy Systems 14 (4) (2006) 516–527.

[55] W. Pedrycz, Conditional fuzzy C-means, Pattern Recognition Letters 17 (6) (1996) 625–631.

[56] W. Pedrycz, Fuzzy set technology in knowledge discovery, Fuzzy Sets and Systems 98 (3) (1998) 279–290.

[57] W. Pedrycz, Conditional fuzzy clustering in the design of radial basis function neural networks, IEEE Transactions on Neural Networks 9 (4) (1998) 601–612.

[58] K. K. Chintalapudi, M. Kam, The credibilistic fuzzy C-means clustering algorithm, in: IEEE International Conference on Systems, Man, and Cybernetics (SMC 1998), Vol. 2, 1998, pp. 2034–2039.

[59] K. K. Chintalapudi, M. Kam, A noise-resistant fuzzy C means algorithm for clustering, in: Proceedings of IEEE World Congress on Computational Intelligence, Vol. 2, 1998, pp. 1458–1463.

[60] M.-S. Yang, K.-L. Wu, Unsupervised possibilistic clustering, Pattern Recognition 39 (1) (2006) 5–21.

[61] J. Noordam, W. van den Broek, L. Buydens, Multivariate image segmentation with cluster size insensitive fuzzy C-means, Chemometrics and Intelligent Laboratory Systems 64 (1) (2002) 65–78.

[62] I. H. Suh, J.-H. Kim, F. Chung-Hoon Rhee, Convex-set-based fuzzy clustering, IEEE Transactions on Fuzzy Systems 7 (3) (1999) 271–285.

[63] R. N. Dave, R. Krishnapuram, Robust clustering methods: A unified view, IEEE Transactions on Fuzzy Systems 5 (2) (1997) 270–293.

[64] R. Kruse, C. Doring, M.-J. Lesot, Fundamentals of fuzzy clustering, in: J. V. de Oliveira, W. Pedrycz (Eds.), Advances in Fuzzy Clustering and its Applications, Wiley, England, 2007, pp. 3–29.

[65] R. Yager, D. Filev, Approximate clustering via the mountain method, IEEE Transactions on Systems, Man and Cybernetics 24 (8) (1994) 1279–1284.

[66] G. Beni, X. Liu, A least biased fuzzy clustering method, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (9) (1994) 954–960.

[67] K. Rose, E. Gurewitz, G. Fox, Constrained clustering as an optimization method, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (8) (1993) 785–794.

[68] J. M. Leski, Fuzzy c-varieties/elliptotypes clustering in reproducing kernel Hilbert space, Fuzzy Sets and Systems 141 (2) (2004) 259–280.

[69] C. Borgelt, C. Braune, M.-J. Lesot, R. Kruse, Handling noise and outliers in fuzzy clustering, in: D. E. Tamir, N. D. Rishe, A. Kandel (Eds.), Fifty Years of Fuzzy Logic and its Applications, Vol. 326 of Studies in Fuzziness and Soft Computing, Springer International Publishing, 2015, pp. 315–335.

[70] T. Kwok, K. Smith, S. Lozano, D. Taniar, Parallel fuzzy c-means clustering for large data sets, in: B. Monien, R. Feldmann (Eds.), Euro-Par 2002 Parallel Processing, Vol. 2400 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2002, pp. 365–374.

[71] J. Zhou, C. P. Chen, L. Chen, H.-X. Li, A collaborative fuzzy clustering algorithm in distributed network environments, IEEE Transactions on Fuzzy Systems 22 (6) (2014) 1443–1456.

[72] S. Miyamoto, D. Suizu, Fuzzy c-means clustering using kernel functions in support vector machines, Journal of Advanced Computational Intelligence and Intelligent Informatics 7 (1) (2003) 25–30.

[73] D.-M. Tsai, C.-C. Lin, Fuzzy C-means based clustering for linearly and nonlinearly separable data, Pattern Recognition 44 (8) (2011) 1750–1760.

[74] K.-L. Wu, M.-S. Yang, Alternative C-means clustering algorithms, Pattern Recognition 35 (10) (2002) 2267–2278.

[75] L. Chen, C. Chen, M. Lu, A multiple-kernel fuzzy C-means algorithm for image segmentation, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 41 (5) (2011) 1263–1274.

[76] S. Chen, D. Zhang, Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 34 (4) (2004) 1907–1916.

[77] K. Honda, N. Sugiura, H. Ichihashi, Fuzzy PCA-guided robust k-means clustering, IEEE Transactions on Fuzzy Systems 18 (1) (2010) 67–79.

[78] H. Zha, C. Ding, M. Gu, X. He, H. Simon, Spectral relaxation for K-means clustering, in: Proceedings of Advances in Neural Information Processing Systems, 2002, pp. 1057–1064.

[79] L. Kaufman, P. J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons Inc, New York, 1990.

[80] R. J. Hathaway, J. W. Davenport, J. C. Bezdek, Relational duals of the C-means clustering algorithms, Pattern Recognition 22 (2) (1989) 205–212.

[81] R. J. Hathaway, J. C. Bezdek, NERF C-means: Non-Euclidean relational fuzzy clustering, Pattern Recognition 27 (3) (1994) 429–437.

[82] S. Nascimento, B. Mirkin, F. Moura-Pires, Multiple prototype model for fuzzy clustering, in: D. J. Hand, J. N. Kok, M. R. Berthold (Eds.), Advances in Intelligent Data Analysis, Vol. 1642 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 1999, pp. 269–279.

[83] L. Fu, E. Medico, FLAME, A novel fuzzy clustering method for the analysis of DNA microarray data, BMC Bioinformatics 8 (3).

[84] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, New York, 2009.

[85] J. C. Bezdek, R. J. Hathaway, VAT: A tool for visual assessment of (cluster) tendency, in: Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN 2002), Vol. 3, 2002, pp. 2225–2230.

[86] L. Wang, U. T. Nguyen, J. C. Bezdek, C. A. Leckie, K. Ramamohanarao, iVAT and aVAT: Enhanced visual analysis for cluster tendency assessment, in: M. J. Zaki, J. X. Yu, B. Ravindran, V. Pudi (Eds.), Advances in Knowledge Discovery and Data Mining, Vol. 6118 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2010, pp. 16–27.

[87] T. C. Havens, J. C. Bezdek, An efficient formulation of the improved visual assessment of cluster tendency (iVAT) algorithm, IEEE Transactions on Knowledge and Data Engineering 24 (5) (2012) 813–822.

[88] K. Jajuga, L_1-norm based fuzzy clustering, Fuzzy Sets and Systems 39 (1) (1991) 43–50.

[89] L. Bobrowski, J. C. Bezdek, C-means clustering with the ℓ_1 and ℓ_∞ norms, IEEE Transactions on Systems, Man, and Cybernetics 21 (3) (1991) 545–554.

[90] R. J. Hathaway, J. C. Bezdek, Optimization of clustering criteria by reformulation, IEEE Transactions on Fuzzy Systems 3 (1995) 241–246.

[91] R. J. Hathaway, J. C. Bezdek, Y. Hu, Generalized fuzzy C-means clustering strategies using L_p norm distances, IEEE Transactions on Fuzzy Systems 8 (5) (2000) 576–582.

[92] N. B. Karayiannisa, M. M. Randolph-Gips, Non-Euclidean C-means clustering algorithms, Intelligent Data Analysis 7 (2003) 405–425.

[93] F. Hoppner, F. Klawonn, R. Kruse, T. Runkler, Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition, Wiley & Sons, Ltd., Chichester, United Kingdom, 1999.

[94] D. E. Gustafson, W. C. Kessel, Fuzzy clustering with a fuzzy covariance matrix, in: IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes, Vol. 17, San Diego, CA, 1979, pp. 761–766.

[95] I. Gath, A. Geva, Unsupervised optimal fuzzy clustering, IEEE Transaction on Pattern Analysis Machine Intelligence 11 (7) (1989) 773–781.

[96] R. J. Hathaway, J. C. Bezdek, Switching regression models and fuzzy clustering, IEEE Transactions on Fuzzy Systems 1 (3) (1993) 195–204.

[97] R. Babuska, P. van der Veen, U. Kaymak, Improved covariance estimation for Gustafson-Kessel clustering, in: Proceedings of the 2002 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2002), Vol. 2, 2002, pp. 1081–1085.

[98] J. C. Bezdek, C. Coray, R. Gunderson, J. Watson, Detection and characterization of cluster substructure I. linear structure: Fuzzy c-lines, SIAM Journal on Applied Mathematics 40 (2) (1981) 339–357.

[99] J. C. Bezdek, C. Coray, R. Gunderson, J. Watson, Detection and characterization of cluster substructure II. Fuzzy c-varieties and convex combinations thereof, SIAM Journal on Applied Mathematics 40 (2) (1981) 358–372.

[100] I. Anderson, J. C. Bezdek, R. Dave, Polygonal shape description of plane boundaries, Systems science and science 1 (1982) 295–301.

[101] R. W. Gunderson, An adaptive FCV clustering algorithm, International Journal of Man-Machine Studies 19 (1) (1983) 97–104.

[102] R. Krishnapuram, H. Frigui, O. Nasraoui, Quadratic shell clustering algorithms and the detection of second-degree curves, Pattern Recognition Letters 14 (1993) 545–552.

[103] R. Krishnapuram, H. Frigui, O. Nasraoui, Fuzzy and possibilistic shell clustering algorithms and their application to boundary detection and surface approximation - part I, IEEE Transaction on Fuzzy Systems 3 (1) (1995) 29–43.

[104] R. Krishnapuram, H. Frigui, O. Nasraoui, Fuzzy and possibilistic shell clustering algorithms and their application to boundary detection and surface approximation - Part II, IEEE Transaction on Fuzzy Systems 3 (1) (1995) 44–60.

[105] H. Frigui, R. Krishnapuram, A comparison of fuzzy shell-clustering methods for the detection of ellipses, IEEE Transactions on Fuzzy Systems 4 (2) (1996) 193–199.

[106] R. N. Dave, Fuzzy shell-clustering and applications to circle detection in digital images, International Journal of General Systems 16 (4) (1990) 343–355.

[107] R. N. Dave, Generalized fuzzy c-shells clustering and detection of circular and elliptical boundaries, Pattern Recognition 25 (7) (1992) 713–721.

[108] R. N. Dave, K. Bhaswan, Adaptive fuzzy C-shells clustering and detection of ellipses, IEEE Transactions on Neural Networks 3 (5) (1992) 643–662.

[109] N. R. Pal, K. Pal, J. M. Keller, J. C. Bezdek, A possibilistic fuzzy c-means clustering algorithm, IEEE Transactions on Fuzzy Systems 13 (4) (2005) 517–530.

[110] J. Leski, Towards a robust fuzzy clustering, Fuzzy Sets and Systems 137 (2) (2003) 215–233.

[111] P. D'Urso, L. D. Giovanni, Robust clustering of imprecise data, Chemometrics and Intelligent Laboratory Systems 136 (2014) 58–80.

[112] J. J. D. Gruijter, A. B. McBratney, A modified fuzzy K-means method for predictive classification, in: H. H. Bock (Ed.), Classification and Related Methods of Data Analysis, Elsevier, Amsterdam, The Netherlands, 1988, pp. 97–104.

[113] R. N. Dave, Characterization and detection of noise in clustering, Pattern Recognition Letters 12 (11) (1991) 657–664.

[114] Y. Ohashi, Fuzzy clustering and robust estimation, Presented at the 9th SAS Users Group International (SUGI) Meeting at Hollywood Beach, Florida. (1984).

[115] R. N. Dave, Robust fuzzy clustering algorithms, in: Second IEEE International Conference on Fuzzy Systems, Vol. 2, 1993, pp. 1281–1286.

[116] J. M. Jolion, P. Meer, S. Bataouche, Robust clustering with applications in computer vision, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (8) (1991) 791–802.

[117] S. Zhuang, T. Wang, P. Zhang, A highly robust estimator through partially likelihood function modeling and its application in computer vision, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (1) (1992) 19–35.

[118] R. Krishnapuram, J. M. Keller, A possibilistic approach to clustering, IEEE Transactions on Fuzzy Systems 1 (2) (1993) 98–110.

[119] M. Barni, V. Cappellini, A. Mecocci, Comments on "A possibilistic approach to clustering", IEEE Transactions on Fuzzy Systems 4 (3) (1996) 393–396.

[120] H. Timm, C. Borgelt, C. Doring, R. Kruse, An extension to possibilistic fuzzy cluster analysis, Fuzzy Sets and Systems 147 (1) (2004) 3–16.

[121] R. Krishnapuram, Generation of membership functions via possibilistic clustering, in: IEEE World Congress on Computational Intelligence, 1994, pp. 902–908 vol.2.

[122] N. R. Pal, K. Pal, J. C. Bezdek, A mixed c-means clustering model, in: Proceedings of the Sixth IEEE International Conference on Fuzzy Systems, Vol. 1, 1997, pp. 11–21.

[123] R. Dave, S. Sen, On generalising the noise clustering algorithms, in: Proceedings of the 7th IFSA World Congress (IFSA 1997), 1997, pp. 205–210.

[124] N. R. Pal, K. Pal, J. M. Keller, J. C. Bezdek, A new hybrid C-means clustering model, in: Proceedings of the 2004 IEEE International Conference on Fuzzy Systems, Vol. 1, 2004, pp. 179–184.

[125] J.-S. Zhang, Y.-W. Leung, Improved possibilistic C-means clustering algorithms, IEEE Transactions on Fuzzy Systems 12 (2) (2004) 209–217.

[126] X.-Y. Wang, J. M. Garibaldi, Simulated annealing fuzzy clustering in cancer diagnosis, Informatica 29 (1) (2005) 61–70.

[127] A. Abadpour, Rederivation of the fuzzypossibilistic clustering objective function through Bayesian inference, Fuzzy Sets and Systems 305 (2016) 29–53.

[128] A. Abadpour, A sequential bayesian alternative to the classical parallel fuzzy clustering model, Information Sciences 318 (2015) 28–47.

[129] A. Keller, Fuzzy clustering with outliers, in: Proceesings of the 19th International Conference of the North American Fuzzy Information Processing Society (NAFIPS 2000), 2000, pp. 143–147.

[130] D.-Q. Zhang, S.-C. Chen, A comment on "Alternative C-means clustering algorithms", Pattern Recognition 37 (2) (2004) 173–174.

[131] B. Triggs, P. F. McLauchlan, R. I. Hartley, A. W. Fitzgibbon, Bundle adjustment - A modern synthesis, in: B. Triggs, A. Zisserman, R. Szeliski (Eds.), Vision Algorithms: Theory and Practice, Vol. 1883 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2000, pp. 298–372.

[132] D. Dubois, H. Prade, Possibility theory, Plenum Press, New York, 1988.

[133] R. W. Frank Rehm, R. Kruse, Fuzzy clustering with repulsive prototypes, in: A. Laurent, M.-J. Lesot (Eds.), Scalable Fuzzy Algorithms for Data Management and Analysis: Methods and Design, 2010, pp. 332–336.

[134] R. N. Dave, S. Sen, Robust fuzzy clustering of relational data, Fuzzy Systems, IEEE Transactions on 10 (6) (2002) 713–727.

[135] F. Rehm, F. Klawonn, R. Kruse, A novel approach to noise clustering for outlier detection, Soft Computing 11 (5) (2007) 489–494.

[136] P. J. Huber, E. Ronchetti, Robust Statistics, Wiley, New York, 2009.

[137] P. D'Urso, Fuzzy clustering for data time arrays with inlier and outlier time trajectories, IEEE Transactions on Fuzzy Systems 13 (5) (2005) 583–604.

[138] L. A. Garcia-Escudero, A. Gordaliza, Robustness properties of k-means and trimmed k-means, Journal of the American Statistical Association 94 (447) (1999) 956–969.

[139] L. A. Garcia-Escudero, A. Gordaliza, C. Matran, A central limit theorem for multivariate generalized trimmed k-means, Annals of Statistics 27 (3) (1999) 1061–1079.

[140] R. Krishnapuram, A. Joshi, O. Nasraoui, L. Yi, Low-complexity fuzzy relational clustering algorithms for web mining, IEEE Transactions on Fuzzy Systems 9 (4) (2001) 595–607.

[141] L. A. Garcia-Escudero, A. Gordaliza, C. Matran, Trimming tools in exploratory data analysis, Journal of Computational and Graphical Statistics 12 (2) (2003) 434–449.

[142] L. A. Garcia-Escudero, A. Gordaliza, A proposal for robust curve clustering, Journal of Classification 22 (2) (2005) 185–201.

[143] S. Z. Selim, M. Ismail, Soft clustering of multidimensional data: a semi-fuzzy approach, Pattern Recognition 17 (5) (1984) 559–568.

[144] R. Krishnapuram, C.-P. Freg, Fitting an unknown number of lines and planes to image data through compatible cluster merging, Pattern Recognition 25 (4) (1992) 385–400.

[145] R. Krishnapuram, O. Nasraoui, H. Frigui, The fuzzy C-spherical shells algorithm: a new approach, IEEE Transactions on Neural Networks 3 (5) (1992) 663–671.

[146] R. N. Dave, T. Fu, Robust shape detection using fuzzy clustering: Practical applications, Fuzzy Sets and Systems 65 (2-3) (1994) 161–185.

[147] H. Frigui, R. Krishnapuram, Clustering by competitive agglomeration, Pattern Recognition 30 (7) (1997) 1109–1119.

[148] C. J. Friedrich, Alfred Weber's Theory of the Location of Industries, Chicago University Press, Chicago, 1929, translated from the original title "Uber den Standort der Industrien" by Alfred Weber.

[149] G. Wesolowski, The Weber problem: History and perspective, Location Science 1 (1993) 5–23.

[150] E. Weiszfeld, Sur le point pour lequel la somme des distances de n points donnes est minimum, Tohoku Mathematical Journal 43 (1937) 355–386.

[151] Z. Drezner, A note on accelerating the Weiszfeld procedure, Location Science 3 (1995) 275–279.

[152] H. Curry, The method of steepest descent for nonlinear minimization problems, Quarterly of Applied Mathematics 2 (1944) 258–261.

[153] H. W. Kuhn, R. E. Kuenne, An efficient algorithm for the numerical solution of the generalized Weber problem in the spatial economics, Journal of Regional Science 4 (1962) 21–33.

[154] J. Brimberg, R. F. Love, Local convergence in a generalized Fermat-Weber problem, Annals of Operations Research 40 (1) (1992) 33–66.

[155] J. Brimberg, R. F. Love, Global convergence of a generalized iterative procedure for the minisum location problem with ℓ_p distances, Operations Research 41 (6) (1993) 1153–1163.

[156] J. Brimberg, R. Chen, D. Chen, Accelerating convergence in the Fermat-Weber location problem, Operations Research Letters 22 (4-5) (1998) 151–157.

[157] A. Cohen, Stepsize analysis for descent methods, Journal of Optimization Theory and Applications 33 (2) (1981) 187–205.

[158] M. R. Osborne, Finite Algorithms in Optimization and Data Analysis, John Wiley, New York, 1985.

[159] A. Bjorck, Numerical Methods for Least Squares Problems, Society for Industrial and Applied Mathematics, 1996.

[160] A. N. Tikhonov, V. Y. Arsenin, Solutions of ill posed problems, Mathematics of Computation 32 (144) (1977) 1320–1322.

[161] K. Levenberg, A method for the solution of certain non-linear problems in least squares, Quarterly Journal of Applied Mathmatics II (2) (1944) 164–168.

[162] D. W. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, Journal of the Society for Industrial and Applied Mathematics 11 (2) (1963) 431–441.

[163] J. J. More, The Levenberg-Marquardt algorithm: Implementation and theory, in: G. Watson (Ed.), Numerical Analysis, Vol. 630 of Lecture Notes in Mathematics, Springer Berlin Heidelberg, 1978, pp. 105–116.

[164] A. Beck, S. Sabach, Weiszfelds method: Old and new results, Journal of Optimization Theory and Applications (2014) 1–40.

[165] P. D'Urso, Fuzzy clustering of fuzzy data, in: J. V. de Oliveira, W. Pedrycz (Eds.), Advances in Fuzzy Clustering and its Applications, Wiley, England, 2007, pp. 155–192.

[166] C. Hwang, F.-H. Rhee, Uncertain fuzzy clustering: Interval type-2 fuzzy approach to C-means, IEEE Transactions on Fuzzy Systems 15 (1) (2007) 107–120.

[167] A. Abadpour, A. S. Alfa, J. Diamond, Video-on-demand network design and maintenance using fuzzy optimization, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 38 (2) (2008) 404–420.

[168] L. Szilagyi, Z. Benyo, S. Szilagyi, H. S. Adam, MR brain image segmentation using an enhanced fuzzy C-means algorithm, in: Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2003), Vol. 1, 2003, pp. 724–726.

[169] W. Cai, S. Chen, D. Zhang, Fast and robust fuzzy C-means clustering algorithms incorporating local information for image segmentation, Pattern Recognition 40 (3) (2007) 825–838.

[170] R. J. Hathaway, Y. Hu, Density-weighted fuzzy C-means clustering, IEEE Transactions on Fuzzy Systems 17 (1) (2009) 243–252.

[171] Y. Yang, Image segmentation based on fuzzy clustering with neighborhood information, Optica Applicata 39 (1) (2009) 135–147.

[172] R. Nock, F. Nielsen, On weighting clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (8) (2006) 1223–1235.

[173] J.-L. Chen, J.-H. Wang, A new robust clustering algorithm-density-weighted fuzzy C-means, in: Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC 1999), Vol. 3, 1999, pp. 90–94.

[174] A. H. Hadjahmadi, M. M. Homayounpour, S. M. Ahadi, Bilateral weighted fuzzy C-means clustering, Iranian Journal of Electrical & Electronic Engineering 8 (2012) 108–121.

[175] A. M. Bensaid, L. O. Hall, J. C. Bezdek, L. P. Clarke, Partially supervised clustering for image segmentation, Pattern Recognition 29 (5) (1996) 859–871.

[176] M. Sato, Y. Sato, Fuzzy clustering model for fuzzy data, in: Joint Proceedings of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium Fuzzy Systems, Vol. 4, 1995, pp. 2123–2128.

[177] R. J. Hathaway, J. C. Bezdek, W. Pedrycz, A parametric model for fusing heterogeneous fuzzy data, IEEE Transactions on Fuzzy Systems 4 (3) (1996) 270–281.

[178] M.-S. Yang, C.-H. Ko, On a class of fuzzy c-numbers clustering procedures for fuzzy data, Fuzzy Sets and Systems 84 (1) (1996) 49–60.

[179] W. Pedrycz, J. C. Bezdek, R. J. Hathaway, G. Rogers, Two nonparametric models for fusing heterogeneous fuzzy data, IEEE Transactions on Fuzzy Systems 6 (3) (1998) 411–425.

[180] M.-S. Yang, H.-H. Liu, Fuzzy clustering procedures for conical fuzzy vector data, Fuzzy Sets and Systems 106 (2) (1999) 189–200.

[181] O. Takata, S. Miyamoto, K. Umayahara, Fuzzy clustering of data with uncertainties using minimum and maximum distances based on L_1 metric, in: Proceedings of IFSA World Congress and 20th NAFIPS International Conference, Vol. 5, 2001, pp. 2511–2516.

[182] S. Auephanwiriyakul, J. M. Keller, Analysis and efficient implementation of a linguistic fuzzy c-means, IEEE Transactions on Fuzzy Systems 10 (5) (2002) 563–582.

[183] M.-S. Yang, P.-Y. Hwang, D.-H. Chen, Fuzzy clustering algorithms for mixed feature variables, Fuzzy Sets and Systems 141 (2) (2004) 301–317.

[184] W.-L. Hung, M.-S. Yang, Fuzzy clustering on LR-type fuzzy numbers with an application in Taiwanese tea evaluation, Fuzzy Sets and Systems 150 (3) (2005) 561–577.

[185] P. DUrso, P. Giordani, A weighted fuzzy c-means clustering model for fuzzy data, Computational Statistics & Data Analysis 50 (6) (2006) 1496–1523.

[186] P. DUrso, A. Santoro, Fuzzy clusterwise linear regression analysis with symmetrical fuzzy output variable, Computational Statistics & Data Analysis 51 (1) (2006) 287–313.

[187] P. DUrso, R. Massari, A. Santoro, A class of fuzzy clusterwise regression models, Information Sciences 180 (24) (2010) 4737–4762.

[188] R. Coppi, P. D'Urso, P. Giordani, Fuzzy and possibilistic clustering for fuzzy data, Computational Statistics & Data Analysis 56 (4) (2012) 915–927.

[189] C.-H. Li, W.-C. Huang, B.-C. Kuo, C.-C. Hung, A novel fuzzy weighted C-means method for image classification, International Journal of Fuzzy Systems 10 (3) (2008) 168–173.

[190] C.-C. Hung, S. Kulkarni, B.-C. Kuo, A new weighted fuzzy C-means clustering algorithm for remotely sensed image classification, IEEE Journal of Selected Topics in Signal Processing 5 (3) (2011) 543–553.

[191] P. D'Urso, L. D. Giovanni, R. Massari, Trimmed fuzzy clustering for interval-valued data, Advances in Data Analysis and Classification 9 (1) (2015) 21–40.

[192] S. Shen, W. Sandham, M. Granat, A. Sterr, MRI fuzzy segmentation of brain tissue using neighborhood attraction with neural-network optimization, IEEE Transactions on Information Technology in Biomedicine 9 (3) (2005) 459–467.

[193] A. K. Jain, R. C. Dubes, Algorithms for Clustering Data, Prentice-Hall, 1981.

[194] M. N. Ahmed, S. M. Yamany, N. Mohamed, A. A. Farag, T. Moriarty, A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data, IEEE Transactions on Medical Imaging 21 (3) (2002) 193–199.

[195] J.-H. Xue, A. Pizurica, W. Philips, E. Kerre, R. V. D. Walle, I. Lemahieu, An integrated method of adaptive enhancement for unsupervised segmentation of MRI brain images, Pattern Recognition Letters 24 (15) (2003) 2549–2560.

[196] A.-E.-O. Boudraa, Automated detection of the left ventricular region in magnetic resonance images by fuzzy C-means model, The International Journal of Cardiac Imaging 13 (4) (1997) 347–355.

[197] Y. Tolias, S. Panas, On applying spatial constraints in fuzzy image clustering using a fuzzy rule-based system, IEEE Signal Processing Letters 5 (10) (1998) 245–247.

[198] M. R. Rezaee, Application of fuzzy techniques in image segmentation, Ph.D. thesis, University of Leiden, Leiden, Netherlands (1998).

[199] S. H. Park, I. D. Yun, S. U. Lee, Color image segmentation based on 3-D clustering: Morphological approach, Pattern Recognition 31 (8) (1998) 1061–1076.

[200] Y. Tolias, S. Panas, Image segmentation by a fuzzy clustering algorithm using adaptive spatially constrained membership functions, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 28 (3) (1998) 359–369.

[201] S. T. Acton, D. P. Mukherjee, Scale space classification using area morphology, IEEE Transactions on Image Processing 9 (4) (2000) 623–635.

[202] J. O. Eklundh, H. Yamamoto, A. Rosenfeld, A relaxation method for multispectral pixel classification, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-2 (1) (1980) 72–75.

[203] J. Y. Hsiao, A. A. Sawchuk, Supervised textured image segmentation using feature smoothing and probabilistic relaxation techniques, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (12) (1989) 1279–1292.

[204] J. Y. Hsiao, A. A. Sawchuk, Unsupervised textured image segmentation using feature smoothing and probabilistic relaxation techniques, Computer Vision, Graphics, and Image Processing 48 (1) (1989) 1–21.

[205] A. W.-C. Liew, S. H. Leung, W. H. Lau, Fuzzy image clustering incorporating spatial continuity, IEE Proceedings on Vision, Image and Signal Processing 147 (2) (2000) 185–192.

[206] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6 (6) (1984) 721–741.

[207] J. Besag, On the statistical analysis of ditty pictures, Journal of the Royal Statistical Society. Series B (Methodological) 48 (3) (1986) 259–302.

[208] H. Derin, H. Elliott, Modeling and segmentation of noisy and textured images using Gibbs random fields, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-9 (1) (1987) 39–55.

[209] S. Lakshmanan, H. Derin, Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (8) (1989) 799–813.

[210] T. N. Pappas, An adaptive clustering algorithm for image segmentation, IEEE Transactions on Signal Processing 40 (4) (1992) 901–914.

[211] J. Noordam, W. van den Broek, L. Buydens, Geometrically guided fuzzy C-means clustering for multivariate image segmentation, in: Proceedings of 15th International Conference on Pattern Recognition, Vol. 1, 2000, pp. 462–465.

[212] J. Noordam, W. H. A. M. van den Broek, Multivariate image segmentation based on geometrically guided fuzzy C-means clustering, Journal of Chemometrics 16 (1) (2002) 1–11.

[213] W. Pedrycz, J. Waletzky, Fuzzy clustering with partial supervision, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 27 (5) (1997) 787–795.

[214] K.-S. Chuang, H.-L. Tzeng, S. Chen, J. Wu, T.-J. Chen, Fuzzy C-means clustering with spatial information for image segmentation, Computerized Medical Imaging and Graphics 30 (1) (2006) 9–15.

[215] Y. Liu, K. Xiao, A. Liang, H. Guan, Fuzzy C-means clustering with bilateral filtering for medical image segmentation, in: E. Corchado, V. Snasel, A. Abraham, M. Wozniak, M. Grana, S.-B. Cho (Eds.), Hybrid Artificial Intelligent Systems, Vol. 7208 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2012, pp. 221–230.

[216] J. Noordam, W. H. A. M. van den Broek, L. M. C. Buydens, Unsupervised segmentation of predefined shapes in multivariate images, Journal of Chemometrics 17 (4) (2003) 216–224.

[217] I. Despotovic, E. Vansteenkiste, W. Philips, Spatially coherent fuzzy clustering for accurate and noise-robust image segmentation, IEEE Signal Processing Letters 20 (4) (2013) 295–298.

[218] S. Roberts, G. R. Gisler, J. P. Theiler, Spatio-spectral image analysis using classical and neural algorithms, in: C. H. Dagli, M. Akay, C. L. P. Chen, B. R. Fernaaandez, J. Ghosh (Eds.), Smart Engineering Systems: Neural Networks, Fuzzy Logic, and Evolutionary Programming, Vol. 6 of Intelligent Engineering Systems Through Artificial Neural Networks, ASME Press, New York, 1996, pp. 425–430.

[219] J. P. Theiler, G. R. Gisler, A contiguity-enhanced k-means clustering algorithm for unsupervised multispectral image segmentation, Proceedings of SPIE (1997) 108–118.

[220] D. L. Pham, Spatial models for fuzzy clustering, Computer Vision and Image Understanding 84 (2) (2001) 285–297.

[221] D. Zhang, S. Chen, A novel kernelized fuzzy C-means algorithm with application in medical image segmentation, Artificial Intelligence in Medicine 32 (1) (2004) 37–50.

[222] M. Girolami, Mercer kernel-based clustering in feature space, IEEE Transactions on Neural Networks 13 (3) (2002) 780–784.

[223] D. L. Pham, Fuzzy clustering with spatial constraints, in: Proceedings of the International Conference on Image Processing (ICIP 2002), Vol. 2, 2002, pp. II–65–II–68.

[224] Y. Yang, C. Zheng, P. Lin, Fuzzy C-means clustering algorithm with a novel penalty term for image segmentation, Opto-Electronics Review 13 (4) (2005) 309–315.

[225] Y. Yang, S. Huang, Image segmentation by fuzzy C-means clustering algorithm with a novel penalty term, Computing and Informatics 26 (1) (2007) 17–31.

[226] C. Ambroise, G. Govaert, Convergence of an EM-type algorithm for spatial clustering, Pattern Recognition Letters 19 (10) (1998) 919–927.

[227] S. Paris, P. Kornprobst, J. Tumblin, F. Durand, Bilateral filtering: Theory and applications, Foundations and Trends in Computer Graphics and Vision 4 (1) (2009) 1–73.

[228] D. L. Pham, J. L. Prince, Adaptive fuzzy segmentation of magnetic resonance images, IEEE Transactions on Medical Imaging 18 (9) (1999) 737–752.

[229] X. Li, L. Li, H. Lu, D. Chen, Z. Liang, Inhomogeneity correction for magnetic resonance images with fuzzy C-mean algorithm, in: Proceedings of SPIE, Vol. 5032, 2003, pp. 995–1005.

[230] S. Krinidis, V. Chatzis, A robust fuzzy local information C-means clustering algorithm, IEEE Transactions on Image Processing 19 (5) (2010) 1328–1337.

[231] D. L. Pham, J. L. Prince, An adaptive fuzzy C-means algorithm for image segmentation in the presence of intensity inhomogeneities, Pattern Recognition Letters 20 (1) (1999) 57–68.

[232] W. M. Wells, W. Grimson, R. Kikinis, F. A. Jolesz, Adaptive segmentation of MRI data, IEEE Transactions on Medical Imaging 15 (4) (1996) 429–442.

[233] A. W.-C. Liew, H. Yan, An adaptive spatial fuzzy clustering algorithm for 3-D MR image segmentation, IEEE Transactions on Medical Imaging 22 (9) (2003) 1063–1075.

[234] M. Gong, Z. Zhou, J. Ma, Change detection in synthetic aperture radar images based on image fusion and fuzzy clustering, IEEE Transactions on Image Processing 21 (4) (2012) 2141–2151.

[235] M. Gong, Y. Liang, J. Shi, W. Ma, J. Ma, Fuzzy C-means clustering with local information and kernel metric for image segmentation, Image Processing, IEEE Transactions on 22 (2) (2013) 573–584.

[236] N. Li, H. Huo, Y. ming Zhao, X. Chen, T. Fang, A spatial clustering method with edge weighting for image segmentation, IEEE Geoscience and Remote Sensing Letters 10 (5) (2013) 1124–1128.

[237] J. C. Dunn, Indices of partition fuzziness and the detection of clusters in large data sets, in: B. G. M.M. Gupta, G.N. Saridis (Ed.), Automata and Decision Process, North-Holland Publishers, Amsterdam, 1977, pp. 271–283.

[238] L. Hubert, P. Arabie, Comparing partitions, Journal of Classification 2 (1) (1985) 193–218.

[239] D. L. Davies, D. W. Bouldin, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1 (2) (1979) 224–227.

[240] M. K. Pakhira, S. Bandyopadhyay, U. Maulik, Validity index for crisp and fuzzy clusters, Pattern Recognition 37 (3) (2004) 487–501.

[241] J. C. Bezdek, Q. W. Li, Y. Attikiouzel, M. Windham, A geometric approach to cluster validity for normal mixtures, Soft Computing 1 (4) (1997) 166–179.

[242] G. W. Milligan, M. C. Cooper, An examination of procedures for determining the number of clusters in a data set, Psychometrika 50 (2) (1985) 159–179.

[243] R. J. Hathaway, J. C. Bezdek, Visual cluster validity for prototype generator clustering models, Pattern Recognition Letters 24 (9–10) (2003) 1563–1569.

[244] U. Maulik, S. Bandyopadhyay, Performance evaluation of some clustering algorithms and validity indices, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (12) (2002) 1650–1654.

[245] T. Calinski, J. Harabasz, A dendrite method for cluster analysis, Communications in Statistics 3 (1) (1974) 1–27.

[246] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (2) (2001) 411–423.

[247] X. Hu, L. Xu, Proceedings of the 4th International Conference Intelligent Data Engineering and Automated Learning (IDEAL 2003), Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, Ch. A Comparative Study of Several Cluster Number Selection Criteria, pp. 195–202.

[248] J. W. Tukey, Exploratory Data Analysis, Addison-Wesley, Reading, MA, 1977.

[249] W. S. Cleveland, Visualizing Data, Hobart Press, Summit. NJ, 1993.

[250] B. S. Everitt, Graphical Techniques for Multivariate Data, North Holland, New York, 1978.

[251] L. Freeman, Displaying hierarchical clusters, Connections 17 (2) (1994) 46–52.

[252] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65.

[253] T. Tran-Luu, Mathematical concepts and novel heuristic methods for data clustering and visualization, Ph.D. thesis, University of Maryland, College Park, MD (1996).

[254] S. Gaddam, Clustertendency repository on github, https://github.com/venganesh/ClusterTendency (Accessed in December 2015).

[255] R. J. Hathaway, J. C. Bezdek, Extending fuzzy and probabilistic clustering to very large data sets, Computational Statistics & Data Analysis 51 (1) (2006) 215–234.

[256] R. C. Prim, Shortest connection networks and some generalizations, Bell System Technical Journal 36 (1957) 1389–1401.

[257] K. Rosen, Discrete Mathematics and Its Applications, McGraw-Hill, New York, 1999.

[258] P. Prabhu, K. Duraiswamy, Enhanced VAT for cluster quality assessment in unlabeled datasets, Journal of Circuits, Systems and Computers 21 (01) (2012) 1–19.

[259] N. Otsu, A threshold selection method from gray level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1) (1979) 62–66.

[260] R. L. Ling, A computer generated aid for cluster analysis, Communications of the ACM 16 (6) (1973) 355–361.

[261] N. Gale, W. C. Halperin, C. M. Costanzo, Unclassed matrix shading and optimal ordering in hierarchical cluster analysis, Journal of Classification 1 (1) (1984) 75–92.

[262] R. Johnson, D. W. Wichern, Applied Multivariate Statistical Analysis, Prentice-Hall, Englewood Cliffs, 1992.

[263] R. Sibson, SLINK: An optimally efficient algorithm for the single-link cluster method, The Computer Journal 16 (1) (1973) 30–34.

[264] M. S. Aldenderfer, R. K. Blashfield, Computer programs for performing hierarchical cluster analysis, Applied Psychological Measurement 2 (3) (1978) 403–411.

[265] T. C. Havens, J. C. Bezdek, J. M. Keller, M. Popescu, J. M. Huband, Is VAT really single linkage in disguise?, Annals of Mathematics and Artificial Intelligence 55 (3) (2009) 237–251.

[266] D. Kumar, J. C. Bezdek, M. Palaniswami, S. Rajasegarar, C. Leckie, T. C. Havens, A hybrid approach to clustering in big data, IEEE Transactions on Cybernetics PP (99) (2015) 1–1. doi:10.1109/TCYB.2015.2477416.

[267] J. M. Huband, J. C. Bezdek, R. J. Hathaway, bigVAT: Visual assessment of cluster tendency for large data sets, Pattern Recognition 38 (11) (2005) 1875–1886.

[268] E. R. Tufte, The Visual Display of Quantitative Information, Graphics Press, 1983.

[269] GENLAB, http://www-it.et.tudelft.nl/2002).

[270] R. Baumgartner, R. Somorjai, R. Summers, W. Richter, L. Ryner, Correlator beware: Correlation has limited selectivity for fMRI data analysis, NeuroImage 12 (2) (2000) 240–243.

[271] R. Baumgartner, R. Somorjai, R. Summers, W. Richter, Ranking fMRI time courses by minimum spanning trees: Assessing coactivation in fMRI, NeuroImage 13 (4) (2001) 734–742.

[272] L. Wilkinson, M. Friendly, The history of the cluster heat map, The American Statistician 63 (2) (2009) 179–184.

[273] T. Loua, Atlas statistique de la population de Paris, Dejey, Paris, 1873.

[274] W. M. F. Petrie, Sequences in prehistoric remains, The Journal of the Anthropological Institute of Great Britain and Ireland 29 (3/4) (1899) 295–301.

[275] W. C. Brinton, Graphic Methods for Presenting Facts, The Engineering Magazine Company, New York, 1914.

[276] J. N. Weinstein, A postgenomic visual icon, Science 319 (5871) (2008) 1772–1773.

[277] M. Hahsler, K. Hornik, C. Buchta, Getting things in order: An introduction to the R package seriation, Journal of Statistical Software 25 (1).

[278] M. Hahsler, K. Hornik, Dissimilarity plots: A visual exploration tool for partitional clustering, Journal of Computational and Graphical Statistics 20 (2) (2011) 335–354.

[279] J. C. Bezdek, R. J. Hathaway, J. M. Huband, Visual assessment of clustering tendency for rectangular dissimilarity matrices, IEEE Transactions on Fuzzy Systems 15 (5) (2007) 890–903.

[280] L. A. F. Park, J. C. Bezdek, C. A. Leckie, Visualization of clusters in very large rectangular dissimilarity data, in: Proceedings of 4th International Conference on Autonomous Robots and Agents (ICARA 2009), 2009, pp. 251–256.

[281] T. C. Havens, J. C. Bezdek, A new formulation of the coVAT algorithm for visual assessment of clustering tendency in rectangular data, International Journal of Intelligent Systems 27 (6) (2012) 590–612.

[282] R. J. Hathaway, J. C. Bezdek, J. M. Huband, Scalable visual assessment of cluster tendency for large data sets, Pattern Recognition 39 (7) (2006) 1315–1324.

[283] P. Prabhu, K. Duraiswamy, An efficient visual analysis method for cluster tendency evaluation, data partitioning and internal cluster validation, Computing and Informatics 32 (5) (2013) 1013–1037.

[284] J. M. Huband, J. C. Bezdek, Computational intelligence: Research frontiers, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, Ch. VCV2 - Visual Cluster Validity, pp. 293–308.

[285] J. M. Huband, J. C. Bezdek, R. J. Hathaway, Revised visual assessment of (cluster) tendency (reVAT), in: Proceedings of IEEE Annual Meeting of the Fuzzy Information (NAFIPS 2004), Vol. 1, 2004, pp. 101–104.

[286] T. C. Havens, J. C. Bezdek, M. Palaniswami, Scalable single linkage hierarchical clustering for big data, in: Proceedings of 2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, 2013, pp. 396–401.

[287] L. Wang, X. Geng, J. C. Bezdek, C. Leckie, R. Kotagiri, SpecVAT: Enhanced visual cluster analysis, in: Proceedings of 2008 Eighth IEEE International Conference on Data Mining, 2008, pp. 638–647.

[288] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: T. G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14, 2001, pp. 585–591.

[289] A. Y. Ng, M. I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: T. G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems, Vol. 14, 2001, pp. 849–856.

[290] L. Wang, X. Geng, J. C. Bezdek, C. Leckie, R. Kotagiri, Enhanced visual analysis for cluster tendency assessment and data partitioning, IEEE Transactions on Knowledge and Data Engineering 22 (10) (2010) 1401–1414.

[291] P. Prabhu, K. Duraiswamy, Enhanced dark block extraction method performed automatically to determine the number of clusters in unlabeled data sets, International Journal of Computers, Communications and Control 8 (2) (2013) 275–293.

[292] I. J. Sledge, T. C. Havens, J. M. Huband, J. C. Bezdek, J. M. Keller, Finding the number of clusters in ordered dissimilarities, Soft Computing 13 (12) (2009) 1125–1142.

[293] L. Wang, C. Leckie, K. Ramamohanarao, J. C. Bezdek, Automatically determining the number of clusters in unlabeled data sets, IEEE Transactions on Knowledge and Data Engineering 21 (3) (2009) 335–350.

[294] T. C. Havens, J. C. Bezdek, J. M. Keller, M. Popescu, Dunn's cluster validity index as a contrast measure of VAT images, in: Proceedings of the 19th International Conference on Pattern Recognition (ICPR 2008), 2008, pp. 1–4.

[295] S. Bandyopadhyay, S. K. Pal, Pixel classification using variable string genetic algorithms with chromosome differentiation, IEEE Transactions on Geoscience and Remote Sensing 39 (2) (2001) 303–308.

[296] M. K. Pakhira, S. Bandyopadhyay, U. Maulik, A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification, Fuzzy Sets and Systems 155 (2) (2005) 191–214.

[297] M. K. Pakhira, Finding number of clusters before finding clusters, Procedia Technology 4 (2012) 27–37, 2nd International Conference on Computer, Communication, Control and Information Technology(C3IT-2012) on February 25 - 26, 2012.

[298] T. C. Havens, J. C. Bezdek, J. M. Keller, M. Popescu, Clustering in ordered dissimilarity data, International Journal of Intelligent Systems 24 (5) (2009) 504–528.

[299] M. Sezgin, B. Sankur, Survey over image thresholding techniques and quantitative performance evaluation, Journal of Electronic Imaging 13 (1) (2004) 146–168.

[300] V. Ganti, J. Gehrke, R. Ramakrishnan, Mining very large databases, Computer 32 (8) (1999) 38–45.

[301] I. Borg, J. Lingoes, Multidimensional Analysis, Springer-Verlag, New York, 1987.

[302] M. Kendall, J. Gibbons, Rank Correlation Methods, Oxford University Press, New York, 1990.

[303] Y. Ding, R. F. Harrison, Relational visual cluster validity (RVCV), Pattern Recognition Letters 28 (15) (2007) 2071–2079.

[304] M. Popescu, J. C. Bezdek, J. M. Keller, T. C. Havens, J. M. Huband, A new cluster validity measure for bioinformatics relational datasets, in: Proceedings of IEEE World Congress on Computational Intelligence Fuzzy Systems (FUZZ-IEEE 2008), 2008, pp. 726–731.

[305] B. Fischer, T. Zoller, J. M. Buhmann, Proceedings of the third international workshop on energy minimization methods in computer vision and pattern recognition (emmcvpr 2001), Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, Ch. Path Based Pairwise Data Clustering with Application to Texture Segmentation, pp. 235–250.

[306] D. Kumar, J. C. Bezdek, S. Rajasegarar, C. Leckie, M. Palaniswami, A visual-numeric approach to clustering and anomaly detection for trajectory data, The Visual Computer (2015) 1–17.

[307] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD 1996), 1996, pp. 226–231.

[308] F. Klawonn, Exploring data sets for clusters and validating single clusters, Procedia Computer Science 96 (2016) 1381–1390.

[309] P. D'Urso, M. Disegna, R. Massari, L. Osti, Fuzzy segmentation of postmodern tourists, Tourism Management 55 (2016) 297–308.

[310] A. Guillon, M.-J. Lesot, C. Marsala, N. R. Pal, Proximal Optimization for Fuzzy Subspace Clustering, Springer International Publishing, 2016, pp. 675–686.

[311] L. Himmelspach, S. Conrad, Fuzzy c-Means Clustering of Incomplete Data Using Dimension-Wise Fuzzy Variances of Clusters, Springer International Publishing, 2016, pp. 699–710.

[312] S. Askari, N. Montazerin, M. F. Zarandi, E. Hakimi, Generalized entropy based possibilistic fuzzy c-means for clustering noisy data and its convergence proof, Neurocomputing.

[313] T. Denaux, S. Sriboonchitta, O. Kanjanatarakul, Evidential clustering of large dissimilarity data, Knowledge-Based Systems 106 (2016) 179 – 195.

[314] C. Borgelt, R. Kruse, Agglomerative Fuzzy Clustering, Springer International Publishing, 2017, pp. 69–77.

[315] F. Dotto, A. Farcomeni, L. A. García-Escudero, A. Mayo-Iscar, Robust Fuzzy Clustering via Trimming and Constraints, Springer International Publishing, 2017, pp. 197–204.