Contributions of Fuzzy Concepts to Data Clustering

Sadaaki Miyamoto

Abstract This chapter tries to answer the fundamental question of what main contributions of fuzzy clustering to the theory of cluster analysis from theoretical viewpoints. While fuzzy clustering is thought to be clearly useful by users of this technique, others think that the concept of fuzziness is not needed in clustering. Thus the usefulness of fuzzy clustering is not trivial. The discussion here is divided into two: one is on fuzzy c-means which is best-known fuzzy method of clustering. However, there is another techniques, discussed by Zadeh, in hierarchical clustering which is equivalent to the old technique of the single linkage. This chapter overviews the both techniques, beginning from basic discussion of fuzzy c-means, and introducing the fundamental concept of fuzzy classifiers and its usefulness. A concept of inductive clustering is introduced which means that a result of clustering can be extended to a partition of the whole space. Moreover hierarchical fuzzy clustering is briefly discussed where the transitive closure gives a simple algebraic form of clusters.

Keywords Fuzzy clustering · Fuzzy c-means · Fuzzy classifier · Hierarchical clustering · Inductive clustering

1 Introduction

Data clustering alias cluster analysis, which generates groups of objects from a set of data using mutual similarity (or dissimilarity) between a pair of data, has been known for a long time [8, 12, 19, 20]. According as the subjects of data mining becomes more and more popular, many different techniques of clustering have been developed.

Fuzzy c-means clustering [2–4, 9, 10, 16, 30] is now considered to be a standard technique in cluster analysis by many researchers. There is, however, another method

S. Miyamoto (✉)
University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan
e-mail: miyamoto@risk.tsukuba.ac.jp

V. Torra et al. (eds.), *Fuzzy Sets, Rough Sets, Multisets and Clustering*,
Studies in Computational Intelligence 671, DOI 10.1007/978-3-319-47557-8_2

of fuzzy clustering in hierarchical cluster analysis which uses the transitive closure of a symmetric fuzzy relation [42].

The usefulness of fuzzy c-means clustering is considered as a matter of fact by many researchers, while other researchers still think that non-fuzzy methods including statistical models [25, 34, 35] are enough for clustering data.

A notable feature of fuzzy c-means is said to be its robustness, i.e., results are stable clusters for different initial clusters. Such discussion is empirical and theoretical background for robustness is unclear.

In this paper we try to consider how the robustness of fuzzy c-means is explained from a theoretical viewpoint, for which a natural fuzzy classifier defined on the whole space is introduced. When such a partition of the whole space is naturally induced from a result of clustering, the method is called here *inductive clustering*. Kernel-based clustering is considered here, where both a non-inductive algorithm and an inductive algorithm are studied.

Moreover another method of hierarchical fuzzy clustering is discussed, which uses the transitive closure of symmetric fuzzy relations [42]. This methods has been shown to be equivalent to the well-known method of the single linkage of agglomerative hierarchical clustering [26]. Here the significance of fuzzy relation is the algebraic form of clusters instead of those generated by an algorithm.

We first consider fuzzy c-means and its variations, then hierarchical fuzzy clustering is briefly discussed.

2 Fuzzy c-Means

We begin with notations and then introduce the method of fuzzy c-means by Dunn [9, 10] and Bezdek [2, 3].

Let the set of objects for clustering be denoted by $X = \{x_1, \ldots, x_N\}$ where each object is a point of p-dimensional Euclidean space $\boldsymbol{R}^p$. Thus $x_k = (x_k^1, \ldots, x_k^p)^\top \in \boldsymbol{R}^p$, $k = 1, \ldots, N$. Clusters are denoted either by G_i or simply by i. A similarity or dissimilarity measure between two objects is assumed. For fuzzy c-means, a standard dissimilarity measure is the squared Euclidean distance:

$$D(x, y) = \|x - y\|^2 = \sum_{j=1}^{p} (x^j - y^j)^2. \tag{1}$$

In fuzzy c-means and related methods, the number of clusters denoted by c is assumed to be given beforehand. The membership of object x_k to cluster i is assumed to be given by u_{ki}. Moreover the collection of all memberships is denoted by matrix $U = (u_{ki})$. It is natural to assume that $u_{ki} \in [0, 1]$ for all $1 \leq i \leq c$ and $1 \leq k \leq N$, and moreover $\sum_{j=1}^{c} u_{kj} = 1$, for all $1 \leq k \leq N$.

The method of fuzzy c-means as well as crisp c-means uses a center for a cluster, which is denoted by $v_i = (v_i^1, \ldots, v_i^p)^\top \in \boldsymbol{R}^p$ for cluster i. For simplicity, all cluster centers are summarized into matrix $V = (v_1, \ldots, v_c)$.

Crisp c-Means Algorithm

Many studies of clustering handle K-means [24], also called crisp c-means, of which the basic algorithm is as follows [3]:
CCM: Crisp c-means algorithm.

CCM0: Generate randomly c cluster centers.
CCM1: Allocate each object $x_k (k = 1, \ldots, N)$ to the cluster of the nearest center.
CCM2: Calculate new cluster centers v_i as the centroid (alias the center of gravity).
If all cluster centers are convergent, stop. Otherwise go to **CCM1**.

End CCM.
The center of a cluster G_i is given by

$$v_i = \frac{1}{|G_i|} \sum_{x_k \in G_i} x_k, \tag{2}$$

where $|G_i|$ is the number of objects in G_i.

2.1 Fuzzy c-Means Algorithm

The fundamental idea of fuzzy c-means is an alternative optimization of an objective function, which is proposed by Dunn [9, 10] and Bezdek [2, 3]:

$$J(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ki})^m D(x_k, v_i) \quad (m \geq 1), \tag{3}$$

where $D(x_k, v_i)$ is the squared Euclidean distance (1).

Using this objective function, the following alternative optimization [3] is carried out.
FCM (fuzzy c -means) algorithm.

FCM0: Generate randomly initial fuzzy clusters. Let the solutions be $(\bar{U}, \bar{V})$
FCM1: Minimize $J(U, \bar{V})$ with respect to U. Let the optimal solution be a new $\bar{U}$.
FCM2: Minimize $J(\bar{U}, V)$ with respect to V. Let the optimal solution be a new $\bar{V}$.
FCM3: If the solution $(\bar{U}, \bar{V})$ is convergent, stop. Else go to **FCM1**.

End FCM.

A criterion for convergence is omitted here, see, e.g., [3]. Optimization with respect to U is with the constraint:

$$u_{ki} \in [0,1], \quad 1 \leq i \leq c, 1 \leq k \leq N, \quad \sum_{j=1}^{c} u_{kj} = 1, \quad 1 \leq k \leq N, \tag{4}$$

while optimization with respect to V is without any constraint.

It is well-known that, when $m = 1$, the solution U is reduced to the allocation to the cluster of the nearest center:

$$u_{ki} = 1 \iff i = \arg \min_{1 \leq j \leq c} D(x_k, v_i),$$

and the center is given by (2). Thus the algorithm is equivalent to **CCM** when $m = 1$.

Hence we assume $m > 1$ hereafter, in order to have fuzzy solutions, where the optimal solutions are as follows:

$$\bar{u}_{ki} = \left\{ \sum_{j=1}^{c} \left(\frac{D(x_k, \bar{v}_i)}{D(x_k, \bar{v}_j)} \right)^{\frac{1}{m-1}} \right\}^{-1}, \tag{5}$$

$$\bar{v}_i = \frac{\sum_{k=1}^{N} (\bar{u}_{ki})^m x_k}{\sum_{k=1}^{N} (\bar{u}_{ki})^m}. \tag{6}$$

The derivations are omitted; the readers should refer to [3] or other textbooks.

Equation (5) does not seem to work when $x_k = v_i$, In such a case (5) should be interpreted as

$$\bar{u}_{ki} = \left\{ 1 + \sum_{j \neq i} \left(\frac{D(x_k, \bar{v}_i)}{D(x_k, \bar{v}_j)} \right)^{\frac{1}{m-1}} \right\}^{-1} \tag{7}$$

by eliminating two terms which appear to have a singular point.

Sometimes we omit the bars like

$$u_{ki} = \left\{ \sum_{j=1}^{c} \left(\frac{D(x_k, v_i)}{D(x_k, v_j)} \right)^{\frac{1}{m-1}} \right\}^{-1}, \tag{8}$$

$$v_i = \frac{\sum_{k=1}^{N} (u_{ki})^m x_k}{\sum_{k=1}^{N} (u_{ki})^m}, \tag{9}$$

for simplicity and without confusions.

2.2 A Natural Classifier

Let us consider the next function defined on $\boldsymbol{R}^p$ with a given set of cluster centers V:

$$U_i(x; V) = \left\{ \sum_{j=1}^{c} \left(\frac{D(x, v_i)}{D(x, v_j)} \right)^{\frac{1}{m-1}} \right\}^{-1}, \tag{10}$$

or

$$U_i(x; V) = \left\{ 1 + \sum_{j \neq i} \left(\frac{D(x, v_i)}{D(x, v_j)} \right)^{\frac{1}{m-1}} \right\}^{-1}. \tag{11}$$

It is clear that $U_i(x; V)$ has been derived from u_{ki} simply by replacing object symbol x_k by variable x.

This replacement appears trivial and it also appears that $U_i(x; V)$ has no further information than u_{ki}. On the contrary, this function of fuzzy classifier is important if we wish to observe theoretical properties of fuzzy c-means.

We have the following proposition that shows how the solutions of fuzzy c-means classify the given space.

Proposition 1 *The function $U_i(x_k; V)$ with a given V has the following properties.*

(i) $U_i(x_k; V) = u_{ki}$, *i.e., the fuzzy classifier interpolates the membership value* u_{ki}.

(ii) *When $|x|$ tends to infinity, $U_i(x; V), i = 1, \ldots, c$, approaches the same value of $1/c$:*

$$\lim_{\|x\| \to \infty} U_i(x; V) = \frac{1}{c}.$$

(iii) *The maximum value of $U_i(x; V), i = 1, \ldots, c$, is at $x = v_i$:*

$$\max_{x \in \boldsymbol{R}^p} U_i(x; V) = U_i(v_i, V) = 1.$$

Proof Property *(i)* is trivial. Property *(ii)* is easily obtained by observing

$$\lim_{\|x\| \to \infty} \frac{D(x, v_i)^{\frac{1}{m-1}}}{D(x, v_j)^{\frac{1}{m-1}}} = 1$$

and $U_i(x; V)$ is given by (11). Finally, the third property is almost trivial since the denominator of (11) is reduced to 1 when $x = v_i$. □

Note the significance of the function $U_i(x; V)$ in these propositions. An object x_k is a fixed point, while x is a variable that can be moved toward infinity or can be a cluster center. Without such a classifier, we cannot observe theoretical properties of fuzzy c-means.

Classifier of K-Means

The crisp classifier $U_i^{\mathrm{ccm}}(x; V)$ of K-means (CCM) is obviously the nearest center allocation:

$$U_i^{\mathrm{ccm}}(x; V) = 1 \iff i = \arg\min_{1 \le j \le c} D(x; v_j). \tag{12}$$

We define

$$\mathcal{V}_i = \{x \in \boldsymbol{R}^p \ : \ U_i^{\mathrm{ccm}}(x; V) = 1 \text{ and } U_j^{\mathrm{ccm}}(x; V) = 0, \ \forall j \neq i \ \}, \tag{13}$$

We note that $\mathcal{V}_i$ is a Voronoi region [21] with center v_i and other cluster centers.

We moreover note the next proposition.

Proposition 2 *If we define*

$$\mathcal{V}_i' = \{x \in \boldsymbol{R}^p \ : \ U_i(x; V) > U_j(x; V), \ \forall j \neq i\} \tag{14}$$

for classifiers of fuzzy c-means, we have

$$\mathcal{V}_i' = \mathcal{V}_i, \quad 1 \le i \le c.$$

The proof is easy by direct calculation and omitted. This proposition means that when the result of clustering by fuzzy c-means is made crisp using the maximum membership reallocation by (14), it leads to Voronoi regions.

Note 1 The above Voronoi regions are open sets and the boundary of two or more regions are left unclassified. The problem of a point on the boundary is not essential here and it can belong to all neighboring regions or be left unclassified. Note also the next relation:

$$\bigcup_{i=1}^{c} \overline{\mathcal{V}}_i = \boldsymbol{R}^p, \qquad \mathcal{V}_i \cap \mathcal{V}_j = \emptyset \quad (i \neq j), \tag{15}$$

where $\overline{\mathcal{V}}_i$ is the closure of $\mathcal{V}_i$.

2.3 A Method of Entropy

Other objective functions of fuzzy c-means have also been proposed among which we discuss the use of entropy [23, 27]:

$$J_{\mathrm{ent}}(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} \{u_{ki} D(x_k, v_i) + \lambda^{-1} u_{ki} \log u_{ki}\}, \quad (\lambda > 0). \tag{16}$$

We easily have the solutions for alternative minimization of $J_{\text{ent}}(U, V)$:

$$u_{ki} = \frac{\exp(-\lambda D(x_k, v_i))}{\sum_{j=1}^{c} \exp(-\lambda D(x_k, v_j))}, \tag{17}$$

$$v_i = \frac{\sum_{k=1}^{N} u_{ki} x_k}{\sum_{k=1}^{N} u_{ki}}, \tag{18}$$

from which the classifier is given as follows:

$$U_i^{\text{ent}}(x; V) = \frac{\exp(-\lambda D(x, v_i))}{\sum_{j=1}^{c} \exp(-\lambda D(x, v_j))}. \tag{19}$$

These solutions are sometimes called the *entropy method* in contrast to the fuzzy c-means using (3).

We consider properties of $U_i^{\text{ent}}(x; V)$, which are more complicated than those of $U_i(x; V)$.

Proposition 3 *Define*

$$\mathcal{V}_i'' = \{x \in \boldsymbol{R}^p \ : \ U_i^{\text{ent}}(x; V) > U_j^{\text{ent}}(x; V), \ \forall j \neq i \ \} \tag{20}$$

for classifiers of fuzzy c-means using entropy term. We then have

$$\mathcal{V}_i'' = \mathcal{V}_i, \quad 1 \leq i \leq c.$$

Proposition 4 *Assume that matrix* $V = (v_1, \ldots, v_c)$ *has full rank* ($\text{rank} V = \min\{c, p\}$). *If* $\mathcal{V}_i''$ *is unbounded, then*

$$\lim_{\|x\| \to \infty; x \in \mathcal{V}_i''} U_i^{\text{ent}}(x; V) = 1,$$

whereas if $\mathcal{V}_i''$ *is bounded, then*

$$\lim_{\|x\| \to \infty} U_i^{\text{ent}}(x; V) = 0.$$

On the other hand, we have

$$0 < U_i^{\text{ent}}(x; V) < 1.$$

The proof is given in [28, 30] and omitted here.

Robustness of Fuzzy c-Means

Let us compare the solutions of the three methods of fuzzy c-means with (3), the K-means of CCM algorithm, and the entropy method using (16). For this purpose we compare the functions $U_i(x; V)$, $U_i^{\mathrm{ccm}}(x; V)$, and $U_i^{\mathrm{ent}}(x; V)$.

Suppose that x is very far away from centers $v_1, \ldots, v_c$, then the membership of x by fuzzy c-means is $U_i(x; V) \approx \frac{1}{c}$ for all $1 \leq i \leq c$, whereas $U_i^{\mathrm{ccm}}(x; V) = 1$ and $U_j^{\mathrm{ccm}}(x; V) = 0$ $(j \neq i)$ for $x \in \mathcal{V}_i'$ by CCM. The result by the entropy method is similar to CCM; $U_i^{\mathrm{ent}}(x; V) \approx 1$ and $U_j^{\mathrm{ccm}}(x; V) \approx 0$ $(j \neq i)$ for $x \in \mathcal{V}_i''$. This means that the results by the K-means (CCM) and the entropy method are strongly influenced by outliers, i.e., objects far from cluster centers.

Moreover, the function $U_i(x; V)$ has the maximum value of unity when $x = v_i$, while the entropy method does not have this property.

Thus the fuzzy c-means has the desirable properties than the K-means and the entropy method.

3 Generalization of Fuzzy c-Means

Many variations of fuzzy c-means have been studied, e.g., fuzzy c-varieties [3], fuzzy c-regressions [15], noise clustering [6], and possibilistic clustering [22]. We, however, limit ourselves to the discussion of the method of Gustafson and Kessel [14] and its extension [30] to take clusterwise covariance and another variable for cluster size into account.

In this section we introduce

$$D(x, v; S) = (x - v)^\top S^{-1}(x - v)$$

which is the squared Mahalanobis distance.

3.1 The Method of Gustafson and Kessel and Its Generalization

The method of Gustafson and Kessel incorporate clusterwise covariance variables denoted by $S_1, \ldots, S_c$. The objective function is

$$J(U, V, \boldsymbol{S}) = \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ki})^m D(x_k, v_i; S_i) \quad (m > 1), \tag{21}$$

where a simplified symbol $\boldsymbol{S} = (S_1, \ldots, S_c)$ and the clusterwise squared Mahalanobis distance $D(x_k, v_i; S_i)$ is used.

Miyamoto et al. introduced an objective function

$$J(U, V, \boldsymbol{S}, A) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\alpha_i)^{1-m}(u_{ki})^m D(x_k, v_i; S_i) \quad (m > 1), \tag{22}$$

with an additional variable $A = (\alpha_1, \ldots, \alpha_c)$ with the constraint

$$\sum_{i=1}^{c}\alpha_i = 1, \qquad \alpha_j \geq 0, \quad 1 \leq j \leq c. \tag{23}$$

Note also that S_i is with the constraint

$$|S_i| = \rho_i \quad (\rho_i > 0) \tag{24}$$

where ρ_i is a fixed parameter and $|S_i|$ is the determinant of S_i. We assume, for simplicity, $\rho_i = 1$ [16].

The solutions are as follows:

$$u_{ki} = \left\{\sum_{j=1}^{c}\left(\frac{D(x_k, v_i; S_i)}{D(x_k, v_j; S_j)}\right)^{\frac{1}{m-1}}\right\}^{-1} \tag{25}$$

$$v_i = \frac{\sum_{k=1}^{N}(u_{ki})^m x_k}{\sum_{k=1}^{N}(u_{ki})^m} \tag{26}$$

$$S_i = \frac{1}{|\hat{S}_i|^{\frac{1}{p}}}\sum_{k=1}^{N}(u_{ki})^m (x_k - v_i)(x_k - v_i)^\top. \tag{27}$$

$$\alpha_i = \left[\sum_{j=1}^{c}\left\{\frac{\sum_{k=1}^{N}(u_{kj})^m D(x_k, v_j; S_j)}{\sum_{k=1}^{N}(u_{ki})^m D(x_k, v_i; S_i)}\right\}^m\right]^{-1} \tag{28}$$

where

$$\hat{S}_i = \sum_{k=1}^{N}(u_{ki})^m (x_k - v_i)(x_k - v_i)^\top.$$

Since four types of variables are used for the augmented method of Gustafson and Kessel, the alternative optimization iteratively calculates (25), (26), (27), and (28) until convergence.

3.2 K–L Information Method

The K–L (Kullback–Leibler) information method by Ichihashi et al. [17, 18, 30] is another generalized version of fuzzy c-means which uses the entropy method. The objective function is as follows.

$$J_{\mathrm{KL}}(U, V, S, A) = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ki} D(x_k, v_i; S_i) + \sum_{i=1}^{c}\sum_{k=1}^{N}\{\nu u_{ki} \log \frac{u_{ki}}{\alpha_i} + \log |S_i|\}. \tag{29}$$

The solutions are given by the following:

$$u_{ki} = \frac{\frac{\alpha_i}{|S_i|}\exp\left(-\frac{D(x_k, v_i; S_i)}{\nu}\right)}{\sum_{j=1}^{c}\frac{\alpha_j}{|S_j|}\exp\left(-\frac{D(x_k, v_j; S_i)}{\nu}\right)}, \tag{30}$$

$$v_i = \frac{\sum_{k=1}^{N} u_{ki} x_k}{\sum_{k=1}^{N} u_{ki}} \tag{31}$$

$$S_i = \frac{1}{\sum_{k=1}^{N} u_{ki}} \sum_{k=1}^{N} u_{ki}(x_k - v_i)(x_k - v_i)^{\top} \tag{32}$$

$$\alpha_i = \frac{1}{N}\sum_{k=1}^{N} u_{ki} \tag{33}$$

The method of K–L information is very similar to the solution of EM algorithm of the Gaussian mixture [25, 34] and moreover generalizes the latter statistical model. The G-K method, when compared with the K–L method, seems to have the robustness property discussed in the previous section.

4 Kernel Based Fuzzy c-Means

The support vector machines [38, 39] with positive definite kernel functions are now one of the most popular methods of supervised classification. Apart from support vector machines, kernel functions themselves are considered to be useful by many researchers (e.g., [13, 36]). Such positive definite kernels can be used for fuzzy c-means, as we see in this section.

The reason why we use kernels for clustering is that essentially the K-means and fuzzy c-means have linear boundaries between clusters of Voronoi regions, as we have seen above.

The introduction of the covariance variables in the last section enables the cluster boundaries to be quadratic, but more flexible nonlinear boundaries cannot be obtained.

In order to have clusters with nonlinear boundaries, we can use positive definite kernels. Kernels are introduced by using a high-dimensional mapping $\Phi : \boldsymbol{R}^p \to H$, where H is generally a Hilbert space with the inner product $\langle \cdot, \cdot \rangle_H$ and the norm $\| \cdot \|_H$.

Given objects $x_1, \ldots x_N$, we consider its images by the mapping $\Phi : \Phi(x_1), \ldots, \Phi(x_N)$. Note that the method of kernels does not assume that an explicit form of $\Phi(x_1), \ldots, \Phi(x_N)$ is known, but their inner product $\langle \Phi(x_i), \Phi(x_j) \rangle_H$ is assumed to be given using a known kernel function $K(x, y)$:

$$K(x, y) = \langle \Phi(x_i), \Phi(x_j) \rangle_H.$$

A well-known example is the Gaussian kernel:

$$K(x, y) = \exp(-C\|x - y\|^2).$$

In this case,

$$\langle \Phi(x), \Phi(y) \rangle_H = \exp(-C\|x - y\|^2).$$

We consider kernel-based fuzzy c-means [29]. The objective function uses $\Phi(x_1), \ldots, \Phi(x_N)$ and cluster centers $w_1, \ldots, w_c$ of H:

$$J(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ki})^m \|\Phi(x_k) - w_i\|_H^2 \quad (m > 1), \tag{34}$$

where $W = (w_1, \ldots, w_c)$. We have

$$u_{ki} = \left\{ \sum_{j=1}^{c} \left(\frac{\|\Phi(x_k) - w_i\|_H^2}{\|\Phi(x_k) - w_j\|_H^2} \right)^{\frac{1}{m-1}} \right\}^{-1} \tag{35}$$

$$w_i = \frac{\sum_{k=1}^{N} (u_{ki})^m \Phi(x_k)}{\sum_{k=1}^{N} (u_{ki})^m} \tag{36}$$

Note, however, that the explicit form of $\Phi(x_k)$ and hence w_i is not available.

We have two ways to handle this situation. First way is to make $\Phi(x_k)$ explicit, whereas the second way is to eliminate $\Phi(x_k)$ and express them in terms of $K(x, y)$.

Use of Gram Matrix

First way to handle $\Phi(x_k)$ is the use of the Gram matrix. Let the Gram matrix be

$$\mathcal{K} = (K(x_k, x_l)), \quad 1 \leq k, l \leq N. \tag{37}$$

Since $\mathcal{K}$ is positive semi-definite and expressed as

$$\mathcal{K} = T^\top \Lambda T,$$

where $\Lambda = diag(\lambda_1, \ldots, \lambda_N)$ is the diagonal matrix of nonnegative eigenvalues and T is the orthogonal matrix, we can define

$$\mathcal{K}^{\frac{1}{2}} = T^\top \Lambda^{\frac{1}{2}} T,$$

where $\Lambda^{\frac{1}{2}} = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_N})$. Let e_k be kth elementary vector: $e_1 = (1, 0, \ldots, 0)^\top$, $e_2 = (0, 1, 0, \ldots, 0)^\top$, and so on. Put

$$\Phi(x_k) = \mathcal{K}^{\frac{1}{2}} e_k, \quad k = 1, 2, \ldots, N. \tag{38}$$

In other words, $\Phi(x_k)$ is kth column (or row) vector of $\mathcal{K}^{\frac{1}{2}}$. Then solutions (35) and (36) are used.

Note that

$$\begin{aligned} \langle \Phi(x_k), \Phi(x_l) \rangle &= (\mathcal{K}^{\frac{1}{2}} e_k)^\top (\mathcal{K}^{\frac{1}{2}} e_l) = e_k^\top \mathcal{K}^{\frac{1}{2}} \mathcal{K}^{\frac{1}{2}} e_l \\ &= e_k^\top \mathcal{K} e_l = K(x_k, x_l), \end{aligned}$$

hence (38) is appropriate.

Note also that Φ is defined on the finite set X ($\Phi : X \to \boldsymbol{R}^p$).

Updating Dissimilarity

Second way is to eliminate w_i from the iterative calculation: updating formula of w_i by Eq. (36) is replaced by the update of dissimilarity

$$D_H(x_k, w_i) = \|\Phi(x_k) - w_i\|^2.$$

We have

$$\begin{aligned} D_H(x_k, w_i) = K(x_k, x_k) &- \frac{2}{\sum_{k=1}^N (u_{ki})^m} \sum_{j=1}^N (u_{ji})^m K(x_j, x_k) \\ &+ \frac{1}{(\sum_{k=1}^N (u_{ki})^m)^2} \sum_{j=1}^N \sum_{\ell=1}^N (u_{ji} u_{\ell i})^m K(x_j, x_\ell). \end{aligned} \tag{39}$$

Using (39), we calculate

$$u_{ki} = \left\{ \sum_{j=1}^{c} \left(\frac{D_H(x_k, w_i)}{D_H(x_k, w_j)} \right)^{\frac{1}{m-1}} \right\}^{-1} \tag{40}$$

Thus the alternative optimization of u_{ki} by (35) and w_i by (36) is replaced by the iteration of (39) and (40) until convergence.

Fuzzy classifiers of kernel-based fuzzy c-means can also be derived by substituting variable x into x_k [30]: we have

$$\begin{aligned} D(x, w_i) = K(x, x) - \frac{2}{\sum_{k=1}^{N} (u_{ki})^m} \sum_{j=1}^{N} (u_{ji})^m K(x, x_j) \\ + \frac{1}{(\sum_{k=1}^{N} (u_{ki})^m)^2} \sum_{j=1}^{N} \sum_{\ell=1}^{N} (u_{ji} u_{\ell i})^m K(x_i, x_\ell), \end{aligned} \tag{41}$$

$$U_i(x, W) = \left\{ \sum_{j=1}^{c} \left(\frac{D_H(x, w_i)}{D_H(x, w_j)} \right)^{\frac{1}{m-1}} \right\}^{-1} \tag{42}$$

Note that $\Phi(x)$ in the second method is defined for an arbitrary point $x \in \boldsymbol{R}^p$ ($\Phi : \boldsymbol{R}^p \to H$) that is different from the function in the first method.

5 Inductive Clustering Versus Non-inductive Clustering

Supervised classification method such as the standard Bayesian classification and the support vector machines provide classification rules defined on the whole space. If the space is $\boldsymbol{R}^p$ and suppose that the Bayesian rule is $P(G_i|x)$ and the SVM rule is SVM, then they are functions of $P(G_i|\cdot) : \boldsymbol{R}^p \to [0, 1]$ and $SVM : \boldsymbol{R}^p \to \{-1, +1\}$. Thus the Bayesian rule is probabilistic, while SVM rule is crisp.

Recently semi-supervised learning has been studied and accordingly the concept of transductive learning [5] has been proposed which means that classification of a finite set of new objects is derived but a classification rule of the whole space is not required. In contrast to transductive learning, a former conventional method of a classification rule of the whole space is called inductive learning.

Turning to the original topic of clustering, the author suppose that many researchers think that clustering is 'transductive' in the above sense, i.e., classification of a given set of objects is enough and nothing more is needed. The above discussed fuzzy classifiers are defined on the whole space, contrary to this general understanding. In short, we discussed inductive properties of fuzzy c-means clustering and related methods.

We now try to make the concept of inductive clustering clearer. If a method of clustering has an *intrinsic classification rule* defined on the whole space, we call the method *inductive clustering*, while the method does not have such a classification rule on the whole space and it gives a classification result on a given set of objects alone, the we call the method *non-inductive clustering* (We avoid the name of *transductive clustering*, as non-inductive property implies nothing in particular).

In this sense, the K-means, fuzzy c-means, and the statistical model of mixture distributions [25] are all inductive clustering, and we can study theoretical properties of them, as we have seen above. Note also that kernel-based fuzzy c-means has the both versions of non-inductive clustering and inductive clustering, since the first method gives $\Phi : X \to \boldsymbol{R}^p$ which is non-inductive clustering, since we cannot use $\Phi(x)$ for $x \notin X$. In contrast, the second way is to use $\Phi : \boldsymbol{R}^p \to H$, which leads us to an inductive version, since we have $\Phi(x)$ and hence $U_i(x, W)$ for any $x \in \boldsymbol{R}^p$.

We emphasize that an advantage of inductive clustering is that we can study its theoretical properties more easily than non-inductive clustering. Indeed, methods of inductive clustering seem to have better or simpler behaviors when generating clusters, like the nearest prototype property of the K-means. On the other hand, if we give up inductiveness in clustering, we have more choice of clustering algorithms, and this attitude has been taken by researchers of clustering, since a proposal of a clustering algorithm does not lead to inductiveness in general.

In spite of this general understanding, we emphasize again the importance of inductive clustering in order to have greater progress in studies of cluster analysis.

Next subject is hierarchical clustering, where we observe again inductiveness of a method, although hierarchical clustering is generally non-inductive.

6 Hierarchical Fuzzy Clustering

Zadeh [42] discussed a fuzzy similarity relation which is reflexive, symmetric, and transitive. In other words, an arbitrary α-cut of a fuzzy similarity relation is reflexive, symmetric, and transitive as a crisp relation.

6.1 Transitive Closure of Fuzzy Relation

We assume that object set $X = \{x_1, \ldots, x_N\}$ are not necessarily in an Euclidean space. Rather, a relation $S(x, y)$ of X satisfying reflexivity and symmetry

$$S(x, x) = 1, \quad \forall x \in X, \tag{43}$$

$$S(x, y) = S(y, x), \quad \forall x, y \in X \tag{44}$$

is assumed, where a larger value of $S(x, y)$ means that x and y are more similar and a smaller value of it implies they are less similar.

Fuzzy transitivity means that

$$S(x, z) \geq \min\{S(x, y), S(y, z)\}, \quad \forall y \in X, \tag{45}$$

but we do not assume this property of transitivity for a given S, since transitivity is a too strong condition in real applications.

Moreover we do not use the term of 'similarity relation' as Zadeh used, but similarity is a general term to show two objects are similar, in order to keep compatibility of terms between the two fields of fuzzy systems and statistical data analysis. We use the term of *fuzzy equivalence relation* instead of similarity relation when a fuzzy relation is reflexive, symmetric, and transitive.

An α-cut $[S]_\alpha$ of S is a crisp relation:

$$[S]_\alpha(x, y) = \begin{cases} 1 & (S(x, y) \geq \alpha), \\ 0 & (S(x, y) < \alpha). \end{cases}$$

Note the next proposition.

Proposition 5 *If a fuzzy relation S is reflexive, symmetric, and transitive, then every α-cut of it is a crisp equivalence relation:*

$$[S]_\alpha(x, x) = 1, \quad \forall x \in X, \tag{46}$$

$$[S]_\alpha(x, y) = [S]_\alpha(y, x), \quad \forall x, y \in X, \tag{47}$$

$$[S]_\alpha(x, y) = 1,\ [S]_\alpha(y, z) = 1 \ \Rightarrow\ [S]_\alpha(x, z) = 1. \tag{48}$$

Proof The first two equations are trivial. The third relation is also easily proved by observing that if $S(x, y) \geq \alpha$ and $S(y, z) \geq \alpha$, then (48) follows from (45). □

Note that we do not assume the transitivity. In order to have a transitive relation from a reflexive and symmetric fuzzy relation, we calculate the transitive closure. For this purpose we introduce the max-min composition of fuzzy relations:

$$(S \circ T)(x, z) = \max_{y \in X} \min\{S(x, y), T(y, z)\},$$

where S and T are fuzzy relations of X. Using the max-min composition, we can define the transitive closure S^* of S:

$$S^*(x, y) = \max\{S(x, y), S^2(x, y), S^3(x, y), \ldots\},$$

where $S^2 = S \circ S$ and $S^k = S \circ S^{k-1}$. It also is not difficult to see $S^* = S^{N-1}$ when S is reflexive and symmetric.

When S is reflexive and symmetric, the transitive closure S^* is also reflexive and symmetric, and moreover transitive. The proof that S^* is transitive is omitted here. Readers can refer to, e.g., [26].

Note that Proposition 5 holds for S^*. Then each α-cut of S^* induces an equivalence class of X, and moreover if α decreases, the equivalence class becomes coarser, and when it increases, the equivalence class becomes finer. Thus S^* defines hierarchical clusters.

6.2 Single Linkage and Transitive Closure

We describe general algorithm of agglomerative hierarchical clustering as follows:
AHC (Algorithm of Agglomerative Hierarchical Clustering).

AHC1: Let initial clusters be individual objects: $G_i = \{x_i\}, i = 1, \ldots, N$. $S(G_i, G_j) = S(x_i, x_j), 1 \leq i, j \leq N$, and put $K = N$.
AHC2: Find pair of clusters of maximum similarity:

$$(G_p, G_q) = \arg\max_{i,j} S(G_i, G_j). \tag{49}$$

Merge $G_r = G_p \cup G_q$. $K = K - 1$ and if $K = 1$, stop.
AHC3: Update $S(G_r, G')$ for all other clusters G'. Go to **AHC1**.

End AHC.

The updating step of **AHC3** admits different choices of similarity between clusters, among which the single linkage, the complete linkage, and the average linkage use the followings:

Single Linkage:

$$S(G_r, G') = \max_{x \in G_r, y \in G'} S(x, y) \tag{50}$$

$$= \max\{S(G_p, G'), S(G_q, G')\} \tag{51}$$

Complete Linkage:

$$S(G_r, G') = \min_{x \in G_r, y \in G'} S(x, y) \tag{52}$$

$$= \min\{S(G_p, G'), S(G_q, G')\} \tag{53}$$

Average Linkage:

$$S(G_r, G') = \frac{\sum_{x \in G_r, y \in G'} S(x, y)}{|G_r||G'|} \tag{54}$$

$$= \frac{|G_p|}{|G_r|} S(G_p, G') + \frac{|G_q|}{|G_r|} S(G_q, G') \tag{55}$$

Discussion in this section is mostly focused upon the single linkage.

We have the proposition of equivalence between the transitive closure and the single linkage [26].

Proposition 6 *Given a set of objects* $X = \{x_1, \ldots, x_N\}$ *and a similarity measure* $S(x, y)$ *for all* $x, y \in X$*, the following three methods give the same hierarchical clusters:*

1. *clusters by the single linkage;*
2. *clusters by the transitive closure* S^**;*
3. *clusters as vertices of connected components of fuzzy graph with vertices* X *and edges* $X \times X$ *with membership values* $S(x, y)$*.*

The connected components of a fuzzy graph is the essential part in this proposition, which means the family of those connected components of all α-cuts of the fuzzy graph. Since connected components grow with decreasing α, those sets of vertices form hierarchical clusters. The proof of this proposition is given in [26] and omitted here, but the idea of the proof is to reduce both the transitive closure and the single linkage clusters to the connected components. Thus fuzzy graph is fundamental in this proposition.

The significance of fuzzy relation and its transitive closure is the algebraic expression of a method of agglomerative clustering in contrast to the general understanding that a method of clustering is essentially a proposal of an algorithm.

Seemingly no new results are included in this theorem. However, Miyamoto [31] showed that ideas in other methods of DBSCAN [11] and Wishart's mode analysis [41] are captured into the above results of equivalence. Concretely, the transitive closure $[S \wedge (\boldsymbol{a}\boldsymbol{a}^\top)]^*$ is proposed in [31], where $\boldsymbol{a}$ is a fuzzy set of X; $\boldsymbol{a}$ is the abstraction of dense points in [41] and core points in [11].

Inductive Property of Hierarchical Clustering

Agglomerative hierarchical clustering in general is non-inductive, as in the assumption that a given X is not in a metric space. When space $\boldsymbol{R}^p$ is given and $X \subset \boldsymbol{R}^p$ is, e.g., with a Euclidean metric, the single linkage can be regarded as an inductive method [33] where the nearest neighbor allocation is used. Roughly, a point x in the Euclidean space can be allocated to the cluster i if a point nearest to x exists in that cluster.

A question arises whether or not this result can be extended to the complete linkage and the average linkage: the furthest neighbor allocation and the average distance allocation can be used, respectively, in these methods. No good answer exists to this question, since it is doubtful that such furthest and/or average allocation methods are as useful as nearest neighbor allocation in the single linkage. Note also that an algebraic expression like the transitive closure is unavailable for the complete linkage or the average linkage.

7 Conclusion

We studied fuzzy clustering and its significance. The method of fuzzy c-means is known to have robustness, and robustness property has been discussed from a theoretical viewpoint using a natural fuzzy classifier, which is derived from substituting an object symbol by a variable. Such function of classification is useful in considering theoretical properties of a clustering method and leads to the concept of inductive clustering, while the original idea of clustering is non-inductive. Kernel fuzzy c-means have been considered in which both non-inductive and inductive algorithms are derived.

Entropy methods including K–L information fuzzy c-means are also discussed which are more closely related to the Gaussian mixture of distributions. They are less robust when compared their theoretical properties with those of the fuzzy c-means including the Gustafson-Kessel method and its extension.

Hierarchical fuzzy clustering was also considered where the transitive closure of a symmetric fuzzy relation is proved to be equivalent to the single linkage method. Thus the transitive closure is an algebraic expression of the well-known agglomerative hierarchical algorithm. Although the result appears purely theoretical, the equivalence leads to a new method of hierarchical clustering [31]. An α-cut of the transitive closure will produce a crisp classifier for the whole space if the problem is given in an Euclidean space, but a fuzzy classifier is difficult to be obtained.

We omitted derivations of solutions of which readers should refer to [3, 16, 30].

An important issue which we omitted here is cluster validity whereby the number of clusters can be decided, which is discussed in [3, 7]. Another topic of recent interest is semi-supervised classification [5, 32, 43] including constrained clustering [1, 37, 40]. Fuzzy classifiers will be useful also in semi-supervised classification, which will be studied in near future.

References

1. S. Basu, I. Davidson, K.L. Wagstaff, *Constrained Clustering*, CRC Press, Boca Raton, 2009.
2. J.C. Bezdek, *Fuzzy Mathematics in Pattern Classification*, Ph.D. Thesis, Cornell Univ., Ithaca, NY, 1973.
3. J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, 1981.
4. J.C. Bezdek, J. Keller, R. Krishnapuram, N.R. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Kluwer, Boston, 1999.
5. O. Chapelle, B. Schölkopf, A. Zien, eds., *Semi-Supervised Learning*, MIT Press, Cambridge, Massachusetts, 2006.
6. R.N. Davé, R. Krishnapuram, Robust clustering methods: a unified view, *IEEE Trans. on Fuzzy Systems*, Vol. 5, pp. 270–293, 1997.
7. D. Dumitrescu, B. Lazzerini, L.C. Jain, *Fuzzy Sets and Their Application to Clustering and Training*, CRC Press, Boca Raton, Florida, 2000.
8. R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.

9. J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *J. of Cybernetics*, Vol. 3, pp. 32–57, 1974.
10. J.C. Dunn, Well-separated clusters and optimal fuzzy partitions, *J. of Cybernetics*, Vol. 4, pp. 95–104, 1974.
11. M. Ester, H.-P. Kriegel, J. Sander, X.W. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, Proc. of 2nd Intern. Conf. on Knowledge Discovery and Data Mining (KDD-96), AAAI Press, pp. 226–231, 1996.
12. B.S. Everitt, *Cluster Analysis, 3rd Ed.*, Arnold, London, 1993.
13. M. Girolami, Mercer kernel based clustering in feature space, *IEEE Trans. on Neural Networks*, Vol. 13, No. 3, pp. 780–784, 2002.
14. E.E. Gustafson, W.C. Kessel, Fuzzy clustering with a fuzzy covariance matrix, IEEE CDC, San Diego, California, pp. 761–766, 1979.
15. R.J. Hathaway, J.C. Bezdek, Switching regression models and fuzzy clustering, *IEEE Trans. on Fuzzy Systems*, Vol. 1, No. 3, pp. 195–204, 1993.
16. F. Höppner, F. Klawonn, R. Kruse, T. Runkler, *Fuzzy Cluster Analysis*, Jhon Wiley & Sons, 1999
17. H. Ichihashi, K. Honda, N. Tani, Gaussian mixture PDF approximation and fuzzy c-means clustering with entropy regularization, Proc. of Fourth Asian Fuzzy Systems Symposium, Vol. 1, pp. 217–221, 2000.
18. H. Ichihashi, K. Miyagishi, K. Honda, Fuzzy c-means clustering with regularization by K-L information, Proc. of 10th IEEE International Conference on Fuzzy Systems, Vol. 2, pp. 924–927, 2001.
19. A.K. Jain, R.C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, NJ, 1988.
20. L. Kaufman, P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, 1990.
21. T. Kohonen, *Self-Organizing Maps, 2nd Ed.*, Springer, Berlin, 1997.
22. R. Krishnapuram, J. M. Keller, A possibilistic approach to clustering, *IEEE Trans. on Fuzzy Systems*, Vol. 1, pp. 98–110, 1993.
23. R.-P. Li and M. Mukaidono, A maximum entropy approach to fuzzy clustering, Proc. of the 4th IEEE Intern. Conf. on Fuzzy Systems (FUZZ-IEEE/IFES'95), Yokohama, Japan, March 20–24, 1995, pp. 2227–2232, 1995.
24. J.B. MacQueen, Some methods of classification and analysis of multivariate observations, Proc. of 5th Berkeley Symposium on Math. Stat. and Prob., pp. 281–297, 1967.
25. G. McLachlan, D. Peel, *Finite Mixture Models*, Wiley, New York, 2000.
26. S. Miyamoto, *Fuzzy Sets in Information Retrieval and Cluster Analysis*, Kluwer, Dordrecht, 1990.
27. S. Miyamoto, M. Mukaidono, Fuzzy *c*-means as a regularization and maximum entropy approach, Proc. of the 7th International Fuzzy Systems Association World Congress (IFSA'97), June 25–30, 1997, Prague, Czech, Vol. II, pp. 86–92, 1997.
28. S. Miyamoto, *Introduction to Cluster Analysis*, Morikita-Shuppan, Tokyo, 1999 (in Japanese).
29. S. Miyamoto, D. Suizu, Fuzzy *c*-means clustering using kernel functions in support vector machines, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 7, No. 1, pp. 25–30, 2003.
30. S. Miyamoto, H. Ichihashi, K. Honda, *Algorithms for Fuzzy Clustering*, Springer, Berlin, 2008.
31. S. Miyamoto, Statistical and non-statistical models in clustering: an introduction and recent topics, A. Okada, D. Vicari, G. Ragozini, Eds., Analysis and Modelling of Complex Data in Behavioural and Social Sciences, JCS-CLADAG 12, Anacapri, Italy, Sept. 3–4, 2012, Cleup, Padova, ISBN 978-88-6129-916-0, pp. 3–6 (Web and USB Proc.) 2012.
32. S. Miyamoto, An Overview of Hierarchical and Non-hierarchical Algorithms of Clustering for Semi-supervised Classification, V. Torra et al. (Eds.): MDAI 2012, LNAI 7647, pp. 1–10, 2012.
33. S. Miyamoto, Inductive and Non-inductive Methods of Clustering, Proc. of 2012 IEEE International Conference on Granular Computing, Aug. 11–12, Hangzhou, China, pp. 12–17, 2012.

34. R.A. Redner, H.F. Walker, Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review*, Vol. 26, No. 2, pp. 195–239, 1984.
35. K. Rose, E. Gurewitz, and G. Fox, "A deterministic annealing approach to clustering," *Pattern Recognition Letters*, Vol. 11, pp. 589–594, 1990.
36. B. Schölkopf, A.J. Smola, *Learning with Kernels*, the MIT Press, 2002.
37. N. Shental, A. Bar-Hillel, T. Hertz, D. Weinshall, Computing Gaussian mixture models with EM using equivalence constraints, In: Advances in Neural Information Processing Systems, Vol. 16, 2004.
38. V.N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.
39. V.N. Vapnik, *The Nature of Statistical Learning Theory: 2nd Ed.*, Springer, New York, 2000.
40. N. Wang, X. Li, X. Luo, Semi-supervised Kernel-based Fuzzy *c*-Means with Pairwise Constraints, Proc. of WCCI 2008, pp. 1099–1103, 2008.
41. Wishart, D.: Mode analysis: a generalization of nearest neighbour which reduces chaining effects, In: A.J. Cole, ed., Numerical Taxonomy, Proc. Colloq., in Numerical Taxonomy, Univ. of St. Andrews, pp. 283–311, 1968.
42. L.A. Zadeh, Similarity relations and fuzzy orderings, *Information Sciences*, Vol. 3, pp. 177–200, 1971.
43. X. Zhu, A.B. Goldberg, *Introduction to Semi-Supervised Learning*, Morgan and Claypool, 2009.

http://www.springer.com/978-3-319-47556-1

Fuzzy Sets, Rough Sets, Multisets and Clustering
Torra, V.; Dahlbom, A.; Narukawa, Y. (Eds.)
2017, X, 347 p. 40 illus., 15 illus. in color., Hardcover
ISBN: 978-3-319-47556-1