Original Article
International Journal of Fuzzy Logic and Intelligent Systems
Vol. 13, No. 4, December 2013, pp. 254-268
http://dx.doi.org/10.5391/IJFIS.2013.13.4.254

ISSN(Print) 1598-2645
ISSN(Online) 2093-744X

An Overview of Unsupervised and Semi-Supervised Fuzzy Kernel Clustering

Hichem Frigui[1], Ouiem Bchir[2], and Naouel Baili[3]
[1]Multimedia Research Lab, University of Louisville, Louisville, KY, USA
[2]Computer Science Department, College of Computer and Information Systems (CCIS), King Saud University, Riyadh, Saudi Arabia
[3]Quintiles, USA

ijfis

Abstract

For real-world clustering tasks, the input data is typically not easily separable due to the highly complex data structure or when clusters vary in size, density and shape. Kernel-based clustering has proven to be an effective approach to partition such data. In this paper, we provide an overview of several fuzzy kernel clustering algorithms. We focus on methods that optimize an fuzzy C-mean-type objective function. We highlight the advantages and disadvantages of each method. In addition to the completely unsupervised algorithms, we also provide an overview of some semi-supervised fuzzy kernel clustering algorithms. These algorithms use partial supervision information to guide the optimization process and avoid local minima. We also provide an overview of the different approaches that have been used to extend kernel clustering to handle very large data sets.

Keywords: Fuzzy clustering, Kernel-based clustering, Relational Kernel clustering, Multiple Kernel clustering, Semi-supervised clustering

1. Introduction

Clustering is an essential and frequently performed task in pattern recognition and data mining. It aids in a variety of tasks related to understanding and exploring the structure of large and high dimensional data. The goal of cluster analysis is to find natural groupings in a set of objects such that objects in the same cluster are as similar as possible and objects in different clusters are as dissimilar as possible.

In most applications, categories are rarely well separated and boundaries are overlapping. Describing these real world situations by crisp sets does not allow the user to quantitatively distinguish between objects which are strongly associated with a particular category from those that have only a marginal association with multiple ones, particularly, along the overlapping boundaries. Fuzzy clustering methods are good at dealing with these situations. In fact, data elements can belong to more than one cluster with fuzzy membership degrees.

Fuzzy clustering is widely used in the machine learning field. Areas of application of fuzzy cluster analysis include data analysis [1, 2], information retrieval [3, 4], image segmentation [5], and robotics [6]. One of the most widely used fuzzy clustering algorithm is the fuzzy C-means (FCM) algorithm [7].

Typically, the data to be clustered could have an object based or a relational based represen-

Received: Dec. 12, 2013
Revised : Dec. 23, 2013
Accepted: Dec. 24, 2013

Correspondence to: Hichem Frigui
(h.frigui@louisville.edu)

tation. In object data representation, each object is represented by a feature vector, while for the relational representation only information about how two objects are related is available. Relational data representation is more general in the sense that it can be applied when only the degree of dissimilarity between objects is available, or when groups of similar objects cannot be represented efficiently by a single prototype.

Despite the large number of existing clustering methods, clustering remains a challenging task when the structure of the data does not correspond to easily separable categories, and when clusters vary in size, density, and shape. Kernel based approaches [8–14] can adapt a specific distance measure in order to make the problem easier. They have attracted great attention and have been applied in many fields such as fault diagnosis of marine diesel engine [12], bridge parameters estimation [13], watermarking [14], automatic classiffication fragments of ceramic cultural relics [15], image segmentation [16], model construction for an erythromycin fermentation process [17], oil-immersed transformer fault diagnosis [18], analyzing enterprises independent innovation capability in order to promote different strategies [19], segmentingmagnetic resonance imaging brain images [20, 21], and classification of audio signals [22].

Kernel clustering approaches rely on their ability to produce nonlinear separating hyper surfaces between clusters by performing a mapping ϕ from the input space $\mathbf{X}$ to a high dimensional feature space $\mathbf{F}$. One of the most relevant aspects in kernel applications is that it is possible to compute Euclidean distances in $\mathbf{F}$ without knowing it explicitly. This can be done using the distance kernel trick [23]:

$$\begin{aligned}\|\Phi(\mathbf{x}_i) - \Phi(\mathbf{x_j})\|^2 = K(\mathbf{x}_i, \mathbf{x}_i) + K(\mathbf{x}_j, \mathbf{x}_j) \\ - 2K(\mathbf{x}_i, \mathbf{x}_j)\end{aligned} \tag{1}$$

In Eq. (1), $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$ is the Mercer Kernel [24]. Gaussian kernels,

$$K^{(g)}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right), \qquad \sigma \in \mathbb{R} \tag{2}$$

are the most commonly used ones.

In this paper, we survey existing fuzzy kernel clustering algorithms. We provide an overview of unsupervised algorithms as well as semi-supervised algorithms that integrate partial supervision into the objective function to guide the optimization process. For most algorithms, we describe the objective function being optimized and the necessary conditions to optimize it. We also highlight the advantages and disadvantages of the different methods.

2. Unsupervised Fuzzy Kernel Based Clustering

Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ be a set of N data points to be partitioned into C clusters, and $\mathbf{R} = [\mathbf{r_{jk}}]$ is a relational matrix where $\mathbf{r}_{jk}$ represents the degree to which pairs of objects $\mathbf{x}_j$ and $\mathbf{x}_k$ are related. The matrix $\mathbf{R}$ could be given or it could be constructed from the features of the objects. Each object $\mathbf{x}_j$ belongs to every cluster i with a fuzzy membership u_{ij} that satisfies [7]:

$$0 \leq u_{ij} \leq 1,$$

and

$$\sum_{i=1}^{C} u_{ij} = 1, \ for\ i,\ j \in \{1, \ldots N\}. \tag{3}$$

The exponent $m \in (1, \infty)$ is a constant that determines the level of cluster fuzziness.

2.1 The Feature Space Kernel (FSK) FCM Algorithm

The FSK-FCM algorithm [25] derives a kernel version of the FCM in the feature space by minimizing

$$J^{\phi} = \sum_{i=1}^{C}\sum_{j=1}^{N} (u_{ij})^m \|\phi(\mathbf{x}_j) - \mathbf{a}_i^{\phi}\|^2. \tag{4}$$

subject to Eq. (3). In Eq. (4), $\mathbf{a}_i^{\phi}$ is the center of cluster i, in the feature space, defined as:

$$\mathbf{a}_i^{\phi} = \frac{\sum_{j=1}^{N} (u_{ij})^m \phi(\mathbf{x}_j)}{\sum_{j=1}^{N} (u_{ij})^m} \tag{5}$$

Minimization of Eq. (4) with respect to u_{ij} yields [25]

$$u_{ih}^{-1} = \sum_{j=1}^{C} \left[\frac{\begin{array}{l} k_{hh} - 2b_i \sum_{r=1}^{n} (u_{ir})^m k_{hr} \\ \qquad + b_i^2 \sum_{r=1}^{n}\sum_{s=1}^{n} (u_{ir}u_{is})^m k_{rs} \end{array}}{\begin{array}{l} k_{hh} - 2b_j \sum_{r=1}^{n} (u_{jr})^m k_{hr} \\ \qquad + b_j^2 \sum_{r=1}^{n}\sum_{s=1}^{n} (u_{jr}u_{js})^m k_{rs} \end{array}} \right]^{1/(m-1)} \tag{6}$$

where

$$b_i = \left(\frac{1}{\sum_{r=1}^{n} (u_{ir})^m}\right) \tag{7}$$

The FSK-FCM algorithm is one of the early kernel versions of the FCM. It is simple and has the flexibility of incorporating different kernels. It achieves this by simply fixing the update equation for the centers in the feature space. Thus, since this equation is not derived to optimize the objective function, there is no guarantee that the obtained centers are optimal. Morevover, in [25] the authors use Gaussian kernels with one global scale (σ) for the entire data. The selection of this parameter is not discussed in [25].

2.2 The Metric Kernel (MK) FCM Algorithm

The MK-FCM [26] is an extension of the FSK-FCM that allows the cluster centers to be optimized. It minimizes

$$J^{\phi} = \sum_{i=1}^{C} \sum_{j=1}^{N} (u_{ij})^m \| \phi(\mathbf{x}_j) - \phi(\mathbf{a}_i) \|^2 \tag{8}$$

subject to the constraint in Eq. (3). In Eq. (8), $\phi(\mathbf{a}_i)$ is the center of cluster i in the feature space F. Minimization of Eq. (8) has been proposed only for the case of a Gaussian kernel using the fact that

$$\frac{\delta K(\mathbf{x}_j, \mathbf{a}_i)}{\delta \mathbf{a}_i} = \frac{(\mathbf{x}_j - \mathbf{a}_i)}{\sigma^2} K(\mathbf{x}_j, \mathbf{a}_i). \tag{9}$$

In this case, it can be shown [26] that the update equations for the memberships and centers are

$$u_{ih} = \left[\sum_{j=1}^{C} \left(\frac{1 - K(\mathbf{x}_j, \mathbf{a}_i)}{1 - K(\mathbf{x}_j, \mathbf{a}_i)} \right)^{1/(m-1)} \right]^{-1}, \tag{10}$$

and

$$\phi(\mathbf{a}_i) = \frac{\sum_{h=1}^{n} (u_{ih})^m K(\mathbf{x}_j, \mathbf{a}_i) \mathbf{x}_j}{\sum_{h=1}^{n} (u_{ih})^m K(\mathbf{x}_j, \mathbf{a}_i)}. \tag{11}$$

Unlike the FSK-FCM, the MK-FCM learns the optimal cluster centers in the feature space. However, the update equations have been derived only for the case of Gaussian kernels with one fixed global scale.

2.3 The Kernelized Non-Euclidean Relational FCM (kNERF) Algorithm

The FSK-FCM and MK-FCM are object based algorithms and require an explicit feature representation of the data to be clustered. The kNERF algorithm [27], on the other hand, is a kernelized version of the non-Euclidean relational FCM algorithm [28], and works on relational data. kKERF does not formulate a new objective function. It simply uses a Gaussian kernel to compute a relational similarity matrix $R = [\mathbf{r}_{jk}]$ using

$$\hat{R} = 1 - exp\left(-\frac{\mathbf{r}_{jk}}{\sigma^2} \right) \tag{12}$$

Then, it uses the non-Euclidean relational fuzzy (NERF) C-means [28] to cluster $\hat{R}$.

kNERF is not a true kernel clustering algorithm. It simply uses kernels as a preprocessing step to create the similarity matrix. Since this step is not part of the optimization process, any kernel function can be used. However, this also prevents the kernel parameters from being optimized for the given data. Also, kNERF constructs a relational matrix with one global Gaussian parameter for the entire data. The selection of this parameter is discussed in [27] but there has been no attempt to devise methods to automatically select it.

2.4 The Clustering and Local Scale Learning (LSL) Algorithm

Although good results were obtained using the Gaussian kernel function, its performance depends on the selection of the scale parameter. Moreover, since one global σ is used for the entire data set, it may not be possible to find one optimal parameter when there are large variations between the distributions of the different clusters in the feature space. One way to learn optimal Gaussian parameters is through an exhaustive search of one parameter for each cluster. However, this approach is not practical since it is computationally expensive especially when the data include a large number of clusters and when the dynamic range of possible values for these parameters is large. Moreover, it is not trivial to evaluate the resulting partition in order to select the optimal parameters. To overcome this limitation, the LSL algorithm [29] has been proposed. It minimizes one objective function for both the optimal partition and for cluster dependent Gaussian parameters that reflect the intra-cluster characteristics of the data. The LSL algorithm minimizes

$$J = \sum_{i=1}^{C} \sum_{j=1}^{N} \sum_{k=1}^{N} u_{ij}^m u_{ik}^m \left(1 - exp\left(-\frac{\mathbf{r}_{jk}}{\sigma_i} \right) \right) - \sum_{i=1}^{C} \frac{K}{\sigma_i^2} \tag{13}$$

subject to the membership constraint in Eq. (3). The first term in Eq. (13) seeks compact clusters using a local relational distance, $\mathbf{D}_{jk}^i$, with respect to each cluster i. This distance is

defined as

$$\mathbf{D}_{jk}^{i} = 1 - exp\left(-\frac{\mathbf{r}_{jk}}{\sigma_i}\right) \tag{14}$$

where the scaling σ_i controls the rate of decay of $\mathbf{D}_{jk}^{i}$ as a function of the distance between $\mathbf{x}_j$ and $\mathbf{x}_k$ with respect to cluster i. The cluster dependent σ_i allows LSL to deal with the large variations, in the feature space, between the distributions and the geometric characteristics of the different clusters. The second term in Eq. (13) is a regularization term to avoid the trivial solution where all the scaling parameters σ_i are infinitely large.

Optimization of J with respect to u_{ij} and σ_i yields the following update equations [29]:

$$u_{ij} = \frac{1}{\sum_{t=1}^{C}\left(d_{ij}^{2}/d_{tj}^{2}\right)^{\frac{1}{m-1}}}, \tag{15}$$

and

$$\sigma_i = \left(\frac{K^{\frac{2\pi^{2-\frac{p}{2}}|\mathcal{N}|}{N}}}{\sum_{j=1}^{N}\sum_{k=1}^{N} u_{ij}^{m}u_{ik}^{m}\mathbf{r}_{jk}}\right)^{\frac{2}{2+p}} \tag{16}$$

where $|\mathcal{N}|$ is the cardinality of the neighborhood of j.

In Eq. (16), σ_i is inversely proportional to the intra-cluster distances with respect to cluster i. Thus, when the intra-cluster dissimilarity is small, σ_i is large allowing the pairwise distances over the same cluster to be smaller and thus, obtain a more compact cluster. On the other hand, when the intra-cluster dissimilarity is high, σ_i is small to prevent points which are not highly similar from being mapped to the same location. According to Eq. (16), σ_i can also be seen as the average time to move between points in cluster i.

LSL has the advantages of learning cluster dependent resolution parameters and can be used to idenify clusters of various densities. However, this also makes the optimization process more complex and prone to local minima. Moreover, the partition generated by LSL depends on the choice of the constant K in Eq. (13).

2.5 The Fuzzy Clustering With Learnable Cluster Dependent Kernels (FLeCK) Algorithm

FLeCK [30] is an extension of LSL that does not assume that K is fixed. It learns the scaling parameters and K by optimizing both the intra-cluster and the inter-cluster dissimilarities. Consequently, the learned scale parameters reflect the relative density, size, and position of each cluster with respect to the other clusters. In particular, FLeCK minimizes the intra-cluster distances

$$J^{intra} = \sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{N} u_{ij}^{m}u_{ik}^{m}\left(1 - \exp\left(-\frac{\mathbf{r}_{jk}}{\sigma_i}\right)\right) + K\sum_{i=1}^{C}\sigma_i \tag{17}$$

and maximizes the inter-cluster distances

$$J^{inter} = \sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{N}\left[\begin{array}{r}\left(u_{ij}^{m}\left(1-u_{ik}^{m}\right)+u_{ik}^{m}\left(1-u_{ij}^{m}\right)\right)\\ \cdot\left(1-\exp\left(-\frac{\mathbf{r}_{jk}}{\sigma_i}\right)\right)\end{array}\right] + K\sum_{i=1}^{C}\sigma_i \tag{18}$$

The constant K provides a way of specifying the relative importance of the regularization term (second term in Eqs. (17) and 18)), compared to the sum of intra-cluster (first term in Eq. (17)) and inter-cluster distances (first term in Eq. (18)). The parameter, σ_i, controls the rate of decay of $\mathbf{D}_{jk}^{i}$. This approach learns a cluster dependent σ_i in order to deal with the large variations between the distributions and the geometric characteristics of each cluster.

Using the Lagrange multiplier technique and assuming that the values of σ_i do not change significantly from one iteration $(t-1)$ to the next one (t), it can be shown [30] that

$$\sigma_i^{(t)} = \frac{Q_1^i}{Q_2^i} \tag{19}$$

where

$$Q_1^i = \sum_{j=1}^{N}\sum_{k=1}^{N} u_{ij}^{m}u_{ik}^{m}\mathbf{r}_{jk}^{2}exp\left(-\frac{\mathbf{r}_{jk}}{\sigma_i^{(t-1)}}\right), \tag{20}$$

and

$$Q_2^i = \sum_{j=1}^{N}\sum_{k=1}^{N}\left[\begin{array}{r}\left(u_{ij}^{m}\left(1-u_{ik}^{m}\right)+u_{ik}^{m}\left(1-u_{ij}^{m}\right)\right)\\ \cdot\,\mathbf{r}_{jk}\exp\left(-\frac{\mathbf{r}_{jk}}{\sigma_i^{(t-1)}}\right)\end{array}\right] + \sum_{j=1}^{N}\sum_{k=1}^{N} u_{ij}^{m}u_{ik}^{m}\mathbf{r}_{jk}\exp\left(-\frac{\mathbf{r}_{jk}}{\sigma_i^{(t-1)}}\right). \tag{21}$$

The simultaneous optimization of Eqs. (17) and (18) with

respect to u_{ij} yields [30]:

$$u_{ij} = \frac{1}{\sum_{t=1}^{C} \left(d_{ij}^2/d_{tj}^2\right)^{\frac{1}{m-1}}}. \quad (22)$$

where

$$d_{ik}^2 = \sum_{j=1}^{N} D_{jk}^i \frac{u_{ij}^m}{\sum_{v=1}^{N} u_{iv}^m} - \frac{1}{2} \sum_{j=1}^{N} \sum_{q=1}^{N} \frac{u_{ij}^m D_{jq}^i u_{iq}^m}{\left(\sum_{v=1}^{N} u_{iv}^m\right)^2} \quad (23)$$

and $\mathbf{D}_{jk}^i$ is as defined in Eq. (14).

2.6 The Fuzzy Clustering With Multiple Kernels (FCMK) Algorithm

LSL [29] and FLeCK [30] have the advantage of learning cluster dependent scaling parameters. Thus, they can be used to identify clusters of different densities. However, these algorithms have two main limitations. First, learning σ_i for each cluster involves complex optimization and is prone to local minima. Second, data points assigned to one cluster are constrained to have one common density. An alternative approach, that involves relatively simpler optimization, and overcomes the above limitations is based on multiple kernel learning (MKL) [31–34]. Instead of using one kernel (fixed or learned) per cluster, MKL uses multiple kernels for each cluster.

Most existing MKL methods assume that kernel weights remain constant for all data (i.e., space-level), while algorithms like Localized MKL [35] seek kernel weights that are data dependent and locally smooth (i.e., sample-level). Although sample-level non-uniform methods give the largest flexibility, in practice relaxations are typically introduced to enhance tractability. Most of the previous MKL approaches have focused on supervised [32, 33] and semi-supervised learning [35]. Recently, an extension of multiple kernels to unsupervised learning, based on maximum margin clustering, was proposed in [36]. However, this method is only designed for crisp clustering. In [37], Huang et al. designed a multiple kernel fuzzy clustering algorithm which uses one set of global kernel weights for all clusters. Therefore, their approach is not appropriate for clusters of various densities. To overcome these drawbacks, Baili and Frigui proposed the FCMK [38]. FCMK is based on the observation that data within the same cluster tend to manifest similar properties. Thus, the intra-cluster weights can be approximately uniform while kernel weights are allowed to vary across clusters.

Given a set of M embedding mappings that map the data to new feature spaces, $\Phi = \{\phi_1, \ldots, \phi_M\}$. Each mapping ϕ_k maps the p-dimensional data x as a vector $\phi_k(x)$ in its feature space whose dimensionality is L_k. Let $\{K_1, \ldots, K_M\}$ be the kernels corresponding to these implicit mapping respectively,

$$K_k(x_j, x_{j'}) = \phi_k(x_j)^T \phi_k(x_{j'}).$$

Since these implicit mappings do not necessarily have the same dimensionality, the authors in [38] construct a new set of independent mappings, $\Psi = \{\psi_1, \psi_2, \ldots, \psi_M\}$, from the original mappings Φ as

$$\psi_1(x_j) = \begin{bmatrix} \phi_1(x_j) \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \psi_2(x_j) = \begin{bmatrix} 0 \\ \phi_2(x_j) \\ \vdots \\ 0 \end{bmatrix},$$

$$\ldots, \psi_M(x_j) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \phi_M(x) \end{bmatrix}.$$

Each of these constructed mappings converts x into an L-dimensional vector, where $L = \sum_{k=1}^{M} L_k$. The linear combination of the bases in Ψ within each cluster i is defined as

$$\psi^{(i)}(x) = \sum_{k=1}^{M} w_{ik} \psi_k(x), \quad (24)$$

with

$$w_{ik} \geq 0, \forall i, \text{ and } \sum_{k=1}^{M} w_{ik} = 1. \quad (25)$$

A new cluster-dependent kernel, $K^{(i)}$, between object j and cluster i, is computed as a convex combination of M Gaussian kernels $K_1, K_2, \ldots, K_M$ with fixed scaling parameters $\sigma_1, \sigma_2, \ldots, \sigma_M$, respectively. That is,

$$K^{(i)}(x_j, a_i) = \sum_{k=1}^{M} w_{ik}^2 K_k(x_j, a_i). \quad (26)$$

In Eq. (26), $\mathbf{W} = [w_{ik}]$, where $w_{ik} \in [0, 1]$ is a resolution weight for kernel K_k with respect to cluster i. A low value of w_{ik} indicates that kernel K_k is not relevant for the density estimation of cluster i (due to its scaling parameter), and that this kernel should not have a significant impact on the creation of this cluster. Similarly, a high value of w_{ik} indicates that the bandwidth of kernel K_k is highly relevant for the density

estimation of cluster i, and that this kernel should be the main factor in the creation of this cluster.

The FCMK algorithm minimizes

$$J = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^{m}\|\psi^{(i)}(x_j) - \psi^{(i)}(a_i)\|^2 \tag{27}$$

subject to the constraints in Eqs. (3) and (25). It can be shown [38] that minimization of Eq. (27) with respect to u_{ij} yields

$$u_{ij} = \frac{1}{\sum_{q=1}^{C}\left(\frac{D_{ij}^2}{D_{qj}^2}\right)^{\frac{1}{m-1}}}, \tag{28}$$

where D_{ij} denotes the distance between data x_j and cluster center a_i, i.e., $D_{ij}^2 = \|\psi^{(i)}(x_j) - \psi^{(i)}(a_i)\|^2$. Similarly, minimization of Eq. (27) with respect to a_i yields

$$a_i = \frac{\sum_{j=1}^{N} u_{ij}^{m}\psi^{(i)}(x_j)}{\sum_{j=1}^{N} u_{ij}^{m}} = \sum_{j=1}^{N}\hat{u}_{ij}\psi^{(i)}(x_j). \tag{29}$$

where $\widehat{u}_{ij} = \frac{u_{ij}^{m}}{\sum_{j=1}^{N} u_{ij}^{m}}$ is the normalized membership.

Using Eq. (29), the distance between data x_j and cluster center a_i can be written as

$$D_{ij}^2 = \sum_{k=1}^{M}\alpha_{ijk}w_{ik}^2, \tag{30}$$

where the coefficients α_{ijk} are defined as

$$\begin{aligned}\alpha_{ijk} = K_k(x_j,x_j) - 2\sum_{h=1}^{N}\hat{u}_{ih}K_k(x_j,x_h) \\ + \sum_{h=1}^{N}\sum_{h'=1}^{N}\hat{u}_{ih}\hat{u}_{ih'}K_k(x_h,x_{h'}).\end{aligned} \tag{31}$$

Replacing Eq. (30) back in Eq. (27) and introducing a Lagrange multiplier, the optimal kernel weights need to be updated using

$$w_{ik} = \frac{\frac{1}{\beta_{ik}}}{\sum_{k'=1}^{M}\frac{1}{\beta_{ik'}}}, \tag{32}$$

where

$$\beta_{ik} = \sum_{j=1}^{N}\mu_{ij}^{m}\alpha_{ijk}. \tag{33}$$

The FCMK is based on the MK-FCM [26] and inherits the limitations of object-based clustering. First, multiple runs of the algorithm may be needed since it is susceptible to initialization problems. Second, FCMK is restricted to data for which there is a notion of center (centroid). Finally, FCMK is not suitable for all types of data. For instance, the density is estimated by counting the number of points within a specified radius, σ_k, of the cluster center. However, for data with multiple resolutions within the same cluster, density should take into account variance between pairs of points and not points to center.

2.7 The Relational Fuzzy Clustering With Multiple Kernels (RFCMK) Algorithm

To overcome the limitations of FCMK, the RFCMK was proposed in [39]. The RFCMK involves dissimilarity between pairs of objects instead of dissimilarity of the objects to a cluster center. It minimizes

$$J = \sum_{i=1}^{C}\frac{\sum_{j=1}^{N}\sum_{h=1}^{N} u_{ij}^{m}u_{ih}^{m}\widehat{r}_{jh}^{(i)}}{\sum_{h=1}^{N} u_{ih}^{m}}, \tag{34}$$

subject to the constraints in Eqs. (3) and (25). In Eq. (34), $\widehat{r}_{jh}^{(i)}$ is the transformed relational data between feature points x_j and x_h, with respect to cluster i define using the implicit mapping $\psi^{(i)}$:

$$\begin{aligned}\widehat{r}_{jh}^{(i)} &= \|\psi^{(i)}(x_j) - \psi^{(i)}(x_h)\|^2 \\ &= \sum_{k=1}^{M} w_{ik}^2K_k(x_j,x_j) + \sum_{k=1}^{M} w_{ik}^2K_k(x_h,x_h) \\ &\quad - 2\sum_{k=1}^{M} w_{ik}^2K_k(x_j,x_h).\end{aligned} \tag{35}$$

The distance D_{ij}^2 from feature vector x_j to the center of the i^{th} cluster, a_i, can be written in terms of the relational matrix $\widehat{R}^{(i)} = [\widehat{r}_{jh}^{(i)}]$ using [40]

$$D_{ij}^2 = \left(\widehat{R}^{(i)}v_i\right)_j - \frac{v_i^T\widehat{R}^{(i)}v_i}{2}, \tag{36}$$

where v_i is the membership vector defined by

$$v_i = \frac{(u_{i1}^{m},\ldots,u_{iN}^{m})^T}{\sum_{j=1}^{N} u_{ij}^{m}}. \tag{37}$$

It can be shown [39] that optimiation of Eq. (34) with respect to u_{ij} yields

$$\mu_{ij} = \frac{1}{\sum_{t=1}^{C}\left(\frac{D_{ij}^2}{D_{tj}^2}\right)^{\frac{1}{m-1}}}. \tag{38}$$

Rewriting the relational data Eq. (35) as

$$\widehat{r}_{jh}^{(i)} = \sum_{k=1}^{M} \alpha_{jhk} w_{ik}^2, \tag{39}$$

where

$$\alpha_{jhk} = K_k(x_j, x_j) + K_k(x_h, x_h) - 2K_k(x_j, x_h), \tag{40}$$

it can be shown [39] that the kernel weights need to be updated using

$$w_{ik} = \frac{1}{\sum_{p=1}^{M} \left(\bar{D}_{ik}/\bar{D}_{ip}\right)}, \tag{41}$$

where

$$\bar{D}_{ik} = \sum_{j=1}^{N}\sum_{h=1}^{N} u_{ij}^m u_{ih}^m \alpha_{jhk}. \tag{42}$$

In Eq. (41), the resolution weight w_{ik} is inversely proportional to α_{jhk}, which is the distance between objects x_j and x_h induced by kernel k. When, the objects are mapped close to each other, w_{ik} will be large indicating that kernel k is relevant. On the other hand, when, the objects are mapped far apart, the weight w_{ik} will be small indicating that kernel k is irrelevant.

3. Semi-Supervised Fuzzy Kernel Based Clustering

Clustering is a difficult combinatorial problem that is susceptible to local minima, especially for high dimensional data. Incorporating prior knowledge in the unsupervised learning task, in order to guide the clustering process has attracted considerable interest among researchers in the data mining and machine learning communities. Some of these methods use available side-information to learn the parameters of the Mahalanobis distance (e.g. [41–43]). The Non-linear methods have focused on the kernelization of Mahalanobis metric learning methods (e.g. [44–46]). However, these approaches are computationally expensive or even infeasible for high dimensional data.

Another approach to guide the clustering process is to use available information in the form of hints, constraints, or labels. Supervision in the form of constraints is more practical than providing class labels. This is because in many real world applications, the true class labels may not be known, and it is much easier to specify whether pairs of points should belong to the same or to different clusters. In fact, pairwise constraints occur naturally in many domains.

3.1 The Semi-Supervised Kernel FCM (SS-KFCM) Algorithm

The SS-KFCM [47] uses L labeled points and U unlabeld points. It assumes that fuzzy memberships can be assigned, using expert knowledge, to the subset of labeled data. Its objective function is defined by adding classification errors of both the labeled and the unlabeled data, i.e.,

$$J = w\sum_{j=1}^{L}\sum_{i=1}^{C}\left(u_{ij}^m - u_{ij,o}^m\right)\left(K\left(\mathbf{x}_j, \mathbf{a}_{i,0}\right) - K\left(\mathbf{x}_j, \mathbf{a}_i\right)\right) + (1-w)\sum_{j=1}^{U}\sum_{i=1}^{C} u_{ij}^m\left(2 - 2K\left(\mathbf{x}_j, \mathbf{a}_i\right)\right) \tag{43}$$

The first term in Eq. (43) is the error between the fuzzy centers calculated based on the learned clusters and the labeled information. The second term minimizes the intra-cluster distances. The optimal solutions to Eq. (43) involves learning the fuzzy memberships and the kernel parameters. It can be shown [47] that for the labeled data, the memberships need to be updated using:

$$u_{ij} = \left(\sum_{k=1}^{C}\left(\frac{K\left(\mathbf{x}_j, \mathbf{x}_j\right) - 2K\left(\mathbf{x}_j, \mathbf{a}_i\right) + K\left(\mathbf{a}_i, \mathbf{a}_i\right)}{K\left(\mathbf{x}_j, \mathbf{x}_j\right) - 2K\left(\mathbf{x}_j, \mathbf{a}_k\right) + K\left(\mathbf{a}_k, \mathbf{a}_k\right)}\right)^{\frac{1}{m-1}}\right)^{-1} \tag{44}$$

and for the unlabeled data, the fuzzy membership need to be updated using:

$$u_{ij} = \frac{u_{ij}K\left(\mathbf{x}_j, \mathbf{a}_i\right)}{\sum_{k=1}^{C} u_{ij}K\left(\mathbf{x}_j, \mathbf{a}_k\right)} \tag{45}$$

Optimization of Eq. (43) with respect to σ does not lead to a closed form expression. Instead, σ is updated iteratively using:

$$\sigma = \sigma - \eta_0 \frac{\delta J\left(L, U\right)}{\delta \sigma} \tag{46}$$

The SS-KFCM algorithm assumes that a subset of the data is labeled with fuzzy membership degrees. However, for real applications, this information may not be available and can be tedious to generate using expert knowledge. An alternative approach uses pairwise constraints. For the following semi-supervised algorithms, we assume that pairwise *"Should-Link"* constraints (pairs of points that should belong to the same cluster) and *"Should not-Link"* constraints (pairs of points that should belong to different clusters) are provided with the input. let $\boldsymbol{Sl}$ be the indicator matrix for the set of *"Should-Link"* pairs

of constraints such that $\boldsymbol{Sl}(j,k) = 1$ means that $\mathbf{x}_j$ and $\mathbf{x}_k$ should be assigned to the same cluster and 0 otherwise. Similarly, let $\boldsymbol{SNl}$ be the indicator matrix for the set of *"Should not-Link"* pairs such that $\boldsymbol{SNl}(j,k) = 1$ means that $\mathbf{x}_j$ and $\mathbf{x}_k$ should not be assigned to the same cluster and 0 otherwise.

3.2 Semi-Supervised Relational Clustering With Local Scaling

The semi-supervised local scaling learning (SS-LSL) [48] minimizes

$$J = \sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{N} u_{ij}^m u_{ik}^m \left(1 - \exp\left(-\frac{\mathbf{r}_{jk}}{\sigma_i}\right)\right) - \sum_{i=1}^{C}\frac{K_1}{\sigma_i^2} - w\sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{N} u_{ij}^m u_{ik}^m \boldsymbol{Sl}(j,k) + w\sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{N} u_{ij}^m u_{ik}^m \boldsymbol{SNl}(j,k) \tag{47}$$

subject to the constraint in Eq. (3). SS-LSL is an extension of the LSL algorithm [29] that incorporates partial supervision. The second term in Eq. (47), is a reward term for satisfying *"Should-Link"* constraints and a penalty term for violating *"Should not-Link"* constraints.

In Eq. (47), the weight $w \in (0,1)$ provides a way of specifying the relative importance of the *"Should-Link"* and *"Should not-Link"* constraints compared to the sum of inter-cluster distances. In [48], the authors recommend fixing it as the ratio of the number of constraints to the total number of points.

Setting the derivative of J with respect to σ_i gives the same update equation for σ_i as the LSL algorithm [29] (Refer to Eq. (16)).

The objective function in Eq. (47) can be rewritten as

$$J = \sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{N} u_{ij}^m u_{ik}^m \hat{D}_{jk}^i - \sum_{i=1}^{C}\frac{K_1}{\sigma_i^2} \tag{48}$$

where

$$\hat{\mathbf{D}}_{jk}^i = \mathbf{D}_{jk}^i - w\boldsymbol{Sl}(j,k) + w\boldsymbol{SNl}(j,k) \tag{49}$$

can be regarded as the "effective distance" that takes into account the satisfaction and violation of the constraints.

It can be shown [48] that optimization of J with respect to u_{ij} yields

$$u_{ij} = \frac{1}{\sum_{t=1}^{C}\left(\hat{d}_{ij}^2/\hat{d}_{tj}^2\right)^{\frac{1}{m-1}}}. \tag{50}$$

where

$$\hat{d}_{ik}^2 = \sum_{j=1}^{N}\hat{D}_{jk}^i\frac{u_{ij}^m}{\sum_{v=1}^{N}u_{iv}^m} - \frac{1}{2}\sum_{j=1}^{N}\sum_{q=1}^{N}\frac{u_{ij}^m\hat{D}_{jq}^i u_{iq}^m}{\left(\sum_{v=1}^{N}u_{iv}^m\right)^2} \tag{51}$$

3.3 The SS-FLeCK Algorithm

SS-FLeCK [49] is an extension of FLeCK [30] that incorporates partial supervision information. It attempts to satisfy a set of *"Should-Link"* and *"Should-Not Link"* constraints by minimizing

$$J^{intra} = \sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{N} u_{ij}^m u_{ik}^m \left(1 - \exp\left(-\frac{\mathbf{r}_{jk}}{\sigma_i}\right)\right) + \sum_{i=1}^{C}K\sigma_i - w\sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{N} u_{ij}^m u_{ik}^m \boldsymbol{Sl}(j,k) + w\sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{N} u_{ij}^m u_{ik}^m \boldsymbol{SNl}(j,k)$$

The optimal K is determined by simultaneously minimizing the intra-cluster distances Eq. (52) and maximizing the inter-cluster distances in Eq. (18).

It can be shown [49] that optimization of Eq. (52) and Eq. (18) with respect to σ_i yields the same update equation for σ_i as in FLeCK (i.e., Eq. (19)). Similarly, optimization of Eq. (52) with respect to the memberships yields the same update equation as SS-LSL (i.e. Eq. (50)).

3.4 The SS-FCMK Algorithm

The SS-FCMK [50] uses the same partial supervision information to extend FCMK [38] by minimizing:

$$J = \sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{M} u_{ij}^2\frac{w_{ik}^2}{\sigma_k}\left(1 - \exp\left(-\frac{\|x_j - a_i\|^2}{2\sigma_k^2}\right)\right) + \gamma\left(\sum_{\substack{SL(j,h)\\=1}}\sum_{i=1}^{C}\sum_{\substack{s=1,\\s\neq i}}^{C} u_{ij}u_{sh} + \sum_{\substack{SNL(j,h)\\=1}}\sum_{i=1}^{C} u_{ij}u_{ih}\right), \tag{52, 53}$$

subject to Eqs. (3) and (25).

In Eq. (53), the weight $\gamma \in (0,1)$ provides a way of specifying the relative importance of the *should-link* and *should*

not-link constraints compared to the sum of inter-cluster distances. It can be shown [50] that the update equation for the membership of point x_j to cluster i is

$$u_{ij} = u_{ij}^{FCMK} + u_{ij}^{Constraints} \tag{54}$$

where u_{ij}^{FCMK} is given by Eq. (28), and

$$u_{ij}^{Constraints} = \frac{\gamma}{2D_{ij}^2}\left(\overline{C}_j - C_{ij}\right). \tag{55}$$

In Eq. (55), C_{ij} and $\overline{C}_j$ are defined as

$$C_{ij} = \sum_{SL(j,h)=1} \sum_{s=1,s\neq i}^{C} u_{sh} + \sum_{SNL(j,h)=1} u_{ih}, \tag{56}$$

and

$$\overline{C}_j = \frac{\sum_{i=1}^{C} \frac{\left(\sum_{SL(j,h)=1} \sum_{s=1,s\neq i}^{C} u_{sh} + \sum_{SNL(j,h)=1} u_{ih}\right)}{D_{ij}^2}}{\sum_{i=1}^{C} \frac{1}{D_{ij}^2}}. \tag{57}$$

The first term in Eq. (54) is the membership term in the FCMK algorithm and only focuses on distances between feature points and prototypes. The second term takes into account the available supervision: memberships are reinforced or reduced according to the pairwise constraints provided by the user. C_{ij} is the constraint violation cost associated with feature point x_j if it is assigned to cluster i, while $\overline{C}_j$ is the weighted average, over all the clusters, of the violation costs associated with feature point x_j. If the violated constraints do not involve feature point x_j, then $u_{ij}^{Constraints} = 0$.

Since the constraints in Eq. (53) do not depend on $\mathbf{W}$ explicitly, optimization of Eq. (53) with respect to $\mathbf{W}$ yields the same update Eq. (32) as in FCMK.

3.5 The SS-RFCMK Algorithm

The SS-RFCMK [50] used partial supervision to extend RFCMK [39]. It minimizes

$$J = \sum_{i=1}^{C} \frac{\sum_{j=1}^{N}\sum_{h=1}^{N} u_{ij}^2 u_{ik}^2 \widehat{r}_{jh}^{(i)}}{2\sum_{h=1}^{N} \mu_{ih}^2} + \gamma\left(\sum_{SL(j,h)=1} \sum_{i=1}^{C} \sum_{s=1,s\neq i}^{C} u_{ij}u_{sh} + \sum_{SNL(j,h)=1} \sum_{i=1}^{C} u_{ij}u_{ih}\right), \tag{58}$$

subject to Eqs. (3) and (25).

Using the Lagrange multiplier technique, it can be shown [50] that the memberships need to be updated using

$$u_{ij} = u_{ij}^{RFCMK} + u_{ij}^{Constraints} \tag{59}$$

where u_{ij}^{RFCMK} is as defined in Eq. (38)

$$\mu_{ij}^{Constraints} = \frac{\gamma}{2D_{ij}^2}\left(\overline{C_j} - C_{ij}\right). \tag{60}$$

In Eq. (60), C_{ij} and $\overline{C_j}$ are as defined in Eqs. (56) and (57).

Since the constraints in Eq. (58) do not depend on w_{ik} explicitly, optimization of Eq. (58) with respect to $\mathbf{W}$ yields the same update Eq. (41) as in RFCMK.

4. Other Fuzzy Kernel Clustering Algorithms

In this paper, we have focused on kernel clustering algorithms that are based on the FCM objective function. There are several other fuzzy kernel clustering approaches. For instance, the multisphere support vectors (MSV) clustering [51] extends the SV clustering approach [52]. It defines a cluster as a sphere in the feature space. This yields kernel-based mapping between the original space and the feature space. The kernel possibilistic C-means (KPCM) algorithm [53] applies the kernel approach to the possibilistic C-means (PCM) algorithm [54]. The weighted kernel fuzzy clustering algorithm (WKFCA) [55] is a kernelized version of the SCAD algorithm [56]. It performs feature weighting along with fuzzy kernel clustering. In [57], the authors propose a kernel fuzzy clustering model that extends the additive fuzzy clustering model to the case of a nonlinear model. More specifically, it has been shown that the additive clustering [58] is special case of fuzzy kernel clustering. The similarity structure is then captured and mapped to higher dimensional space by using kernel functions. The kernel fuzzy clustering methods based on local adaptive distances [59] performs feature weighting and fuzzy kernel clustering simultaneously. The sum of the weights of the variables with respect to each cluster is not equal to one as in [55]. However, the product of the weights of the variables for each cluster is constrained be equal to one. The genetic multiple kernel interval type 2 FCM clustering [60] combines heuristic method based on genetic algorithm (GA) and MK-FCM. It automatically determines the optimal number of clusters and the initial centroids in the first step. Then, it adjusts the coefficients of the kernels and combines them in the feature space to produce a new kernel. Other kernel clustering

algorithms, based on type 2 fuzzy sets, include [61, 62]. A kernel intuitionistic FCM clustering algorithm (KIFCM) was proposed in [63]. EKIFCM has two main phases. The first one is KIFCM and the second phase is parameters selection of KIFCM with GA. KIFCM is a combination of Atanassov's intuitionistic fuzzy sets (IFSs) [64] with kernel-based FCM (KFCM) [47].

5. Fuzzy Kernel Based Clustering for Very Large Data

All of the kernel clustering algorithms that we have outlined do not scale to very large (VL) data sets. VL data or big data are any data that cannot be loaded into the computer's working memory. Since clustering is one of the primary tasks used in the pattern recognition and data mining communities to search VL databases in various applications, it is desirable to have clustering algorithms that scale well to VL data.

The scalability issue has been studied by many researchers. In general, there are three main approaches to clustering for VL data: sampling-based methods, incremental algorithms and data approximation algorithms. The sample and extend algorithms apply clustering to a sample of the dataset found by either progressive [65] or random sampling [66]. Then, the sample results are noniteratively extended to approximate the partition for the rest of the data. Representative algorithms include random sample and extend kernel FCM (rseKFCM) algorithm [67] and the random and extend RFCMK (rseRFCMK) [68]. The rseRFCMK is an extension of the RFCMK algorithm to VL data based on sampling followed by non-iterative extension. The main problem with sampling-based methods is the choice of the sample. For instance, if the sample is not representative of the full dataset, then sample and extend methods cannot accurately approximate the clustering solution.

On the other hand, the incremental algorithms are designed to operate on the full dataset by separating it into multiple chunks. First, these algorithms sequentially load manageable chunks of the data. Then, they cluster each chunk in a single pass. They construct the final partition as a combination of the results from each chunk. In [66, 67], a single pass kernel fuzzy C-means (spKFCM) algorithm was proposed. The spKFCM algorithm runs weighted KFCM (wKFCM) on sequential chunks of the data, passing the clustering solution from each chunk onto the next. spKFCM is scalable as its space complexity is only based on the size of the sample. The single pass RFCMK (spRFCMK) [68] is an extension of the RFCMK algorithm to VL data. The spRFCMK is an incremental technique that makes one sequential pass through subsets of the data.

In [66], an online algorithm for kernel FCM (oKFCM) was proposed in which data is assumed to arrive in chunks. Each chunk is clustered and the memory is freed by summarizing the clustering result by weighted centroids. In contrast to spKFCM, oKFCM is not truly scalable and is not recommended for VL data. This is because, rather than passing the clustering results from one chunk to the next, oKFCM clusters the weighted centroids in one final run to obtain the clustering for the entire stream.

6. Conclusion

Fuzzy kernel clustering has proven to be an effective approach to partition data when the clusters are not well-defined. Several algorithms have been proposed in the past few years and were described in this paper. Some algorithms, such as FSK-FCM, are simple and can incorporate any kernel function. However, these methods impose an intuitive equation to update the centers in the feature space. Thus, there is no guarantee that the optimized objective function of these algorithms correspond to the optimal partition.

Other kernel algorithms can solve for the optimal centers by restricting the kernel to be Gaussian. Some of these algorithms, such as FSK-FCM, use one global scale (σ) for all clusters. These are relatively simpler algorithms that are not very sensitive to the initialization. However, they require the user to specify the scale, and may not perform well when the data exhibit large variations between the distributions of the different clusters. Other algorithms, such as FLeCK, use more complex objective functions to learn cluster-dependent kernel resolution parameters. However, because the search space of these methods include many more parameters, they are more prone to local minima. Clustering with multiple kernels is a good compromize between methods that use one fixed global scale and methods that learn one scale for each cluster. These algorithms, such as FCMK, use a set of kernels with different, but fixed, scales. Then, they learn relevance weights for each kernel within each cluster.

Clustering, in general, is a difficult combinatorial problem that is susceptible to local minima. This problem is more acute in Kernel based clustering as they solve the partitioning problem in a much higher feature space. As a result, several semi-supervised kernel clustering methods have emerged. These algorithms incorporate prior knowledge in order to guide the

optimization process. This prior knowledge is usually available in the form of a small set of constraints that specify which pairs of points should be assigned to the same cluster, and which ones should be assigned to different clusters.

Scalability to very large data is another desirable feature to have in a clustering algorithm. In fact, many applications involve very large data that cannot be loaded into the computer's memory. In this case, algorithm scalability becomes a necessary condition. We have outlined three main approaches that have been used with kernel clustering methods. These include sampling-based methods, incremental algorithms, and data approximation algorithms.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

[1] W. Pedrycz, A. Amato, V. Di Lecce, and V. Piuri, "Fuzzy clustering with partial supervision in organization and classification of digital images," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 4, pp. 1008-1026, Aug. 2008. http://dx.doi.org/10.1109/TFUZZ.2008.917287

[2] V. Loia, W. Pedrycz, and S. Senatore, "Semantic web content analysis: a study in proximity-based collaborative clustering," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 6, pp. 1294-1312, Dec. 2007. http://dx.doi.org/10.1109/TFUZZ.2006.889970

[3] H. Frigui and C. Hwang, "Fuzzy clustering and aggregation of relational data with instance-level constraints," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 6, pp. 1565-1581, Dec. 2008. http://dx.doi.org/10.1109/TFUZZ.2008.2005692

[4] Y. J. Horng, S. M. Chen, Y. C. Chang, and C. H. Lee, "A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 2, pp. 216-228, Apr. 2005. http://dx.doi.org/10.1109/TFUZZ.2004.840134

[5] S. P. Chatzis and T. A. Varvarigou, "A fuzzy clustering approach toward Hidden Markov random field models for enhanced spatially constrained image segmentation," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 5, pp. 1351-1361, Oct. 2008. http://dx.doi.org/10.1109/TFUZZ.2008.2005008

[6] P. X. Liu and M. Q. H. Meng, "Online data-driven fuzzy clustering with applications to real-time robotic tracking," *IEEE Transactions on Fuzzy Systems*, vol. 12, no. 4, pp. 516-523, Aug. 2004. http://dx.doi.org/10.1109/TFUZZ.2004.832521

[7] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*, New York, NY: Plenum Press, 1981.

[8] M. Girolami, "Mercer kernel-based clustering in feature space," *IEEE Transactions on Neural Networks*, vol. 13, no. 3, pp. 780-784, May 2002. http://dx.doi.org/10.1109/TNN.2002.1000150

[9] R. Inokuchi and S. Miyamoto, "LVQ clustering and SOM using a kernel function," in *Proceedings of the IEEE International Conference on Fuzzy Systems*, Budapest, Hungary, July 25-29, 2004, pp. 1497-1500. http://dx.doi.org/10.1109/FUZZY.2004.1375395

[10] A. K. Qin and P. N. Suganthan, "Kernel neural gas algorithms with application to cluster analysis," in *Proceedings of the 17th International Conference on Pattern Recognition*, Cambridge, United Kingdom, August 23-26, 2004, pp. 617-620. http://dx.doi.org/10.1109/ICPR.2004.1333848

[11] U. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395-416, Dec. 2007. http://dx.doi.org/10.1007/s11222-007-9033-z

[12] X. Peng, Y. Chai, L. Xu, and X. Man, "Research on fault diagnosis of marine diesel engine based on grey relational analysis and kernel fuzzy C-means clustering," in *The Fifth International Conference on Intelligent Computation Technology and Automation*, Zhangjiajie, China, January 12-14, 2012, pp. 283-286. http://dx.doi.org/10.1109/ICICTA.2012.78

[13] H. Liu, C. Wu, H. Wan, and H. Wang, "Parameter identification based on stabilization diagram with kernel fuzzy clustering method," in *International Conference on Transportation, Mechanical, and Electrical Engineering*, Changchun, China, December 16-18, 2011, pp. 1185-1188. http://dx.doi.org/10.1109/TMEE.2011.6199417

[14] J. Fan and Y. Wu, "Watermarking algorithm based on kernel fuzzy clustering and singular value decomposition in the complex wavelet transform domain," in *International Conference on Information Technology, Computer Engineering and Management Sciences*, Nanjing, China, September 24-25, 2011, pp. 42-46. http://dx.doi.org/10.1109/ICM.2011.121

[15] L. Y. Qi and K. G. Wang, "Kernel fuzzy clustering based classification of Ancient-Ceramic fragments," in *The 2nd IEEE International Conference on Information Management and Engineering*, Chengdu, China, April 16-18, 2010, pp. 348-350. http://dx.doi.org/10.1109/ICIME.2010.5477818

[16] D. M. Tsai and C. C. Lin, "Fuzzy C-means based clustering for linearly and nonlinearly separable data," *Pattern Recognition*, vol. 44, no. 8, pp. 1750-1760, Aug. 2011. http://dx.doi.org/10.1016/j.patcog.2011.02.009

[17] C. Mei, H. Xu, and J. Liu, "A novel NN-based soft sensor based on modified fuzzy kernel clustering for fermentation process," in *International Conference on Intelligent Human-Machine Systems and Cybernetics*, Hangzhou, China, August 26-27, 2009, pp. 113-117. http://dx.doi.org/10.1109/IHMSC.2009.153

[18] D. H. Liu, J. P. Bian, and X. Y. Sun, "The study of fault diagnosis model of DGA for oil-immersed transformer based on fuzzy means Kernel clustering and SVM multi-class object simplified structure," in *Proceedings of the 7th International Conference on Machine Learning and Cybernetics*, Kunming, China, July 12-15, 2008, pp. 1505-1509. http://dx.doi.org/10.1109/ICMLC.2008.4620644

[19] B. Qu and H. Wang, "Dynamic fuzzy kernel clustering analysis of enterprises independent: innovation capability based on artificial immunity," in *International Workshop on Modelling, Simulation and Optimization*, Hong Kong, December 27-28, 2009, pp. 216-220. http://dx.doi.org/10.1109/WMSO.2008.102

[20] X. Li and S. Bian, "A kernel fuzzy clustering algorithm with spatial constraint based on improved expectation maximization for image segmentation," in *International Conference on Measuring Technology and Mechatronics Automation*, Zhangjiajie, China, April 11-12, 2009, pp. 529-533. http://dx.doi.org/10.1109/ICMTMA.2009.59

[21] L. Liao and T. S. Lin, "A fast spatial constrained fuzzy kernel clustering algorithm for MRI brain image segmentation," in *International Conference on Wavelet Analysis and Pattern Recognition*, Beijing, China, November 2-4, 2008, pp. 82-87. http://dx.doi.org/10.1109/ICWAPR.2007.4420641

[22] R. J. G. B. Campello, "A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment," *Pattern Recognition Letters*, vol. 28, no. 7, pp. 833-841, May 2007. http://dx.doi.org/10.1016/j.patrec.2006.11.010

[23] B. Schölkopf, A. Smola, and K. R. Mller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299-1319, Jul. 1998. http://dx.doi.org/10.1162/089976698300017467

[24] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, no. 3, pp. 337-404, May 1950.

[25] D. Zhang and S. Chen, "Fuzzy clustering using kernel method," in *International Conference on Control and Automation*, Xiamen, China, June 19, 2002, pp. 162-163. http://dx.doi.org/10.1109/ICCA.2002.1229535

[26] D. Q. Zhang and S. C. Chen, "Kernel-based fuzzy and possibilistic c-means clustering," in *Proceedings of the International Conference on Artificial Neural Network*, Istanbul, Turkey, June 26-29, 2003, pp. 122-125.

[27] R. J. Hathaway, J. M. Huband, and J. C. Bezdek, "A kernelized non-euclidean relational fuzzy c-means algorithm," in *IEEE International Conference on Fuzzy Systems*, Reno, NV, May 22-25, 2005, pp. 414-419. http://dx.doi.org/10.1109/FUZZY.2005.1452429

[28] R. J. Hathaway and J. C. Bezdek, "Nerf c-means: non-Euclidean relational fuzzy clustering," *Pattern Recognition*, vol. 27, no. 3, pp. 429-437, Mar. 1994. http://dx.doi.org/10.1016/0031-3203(94)90119-8

[29] O. Bchir and H. Frigui, "Fuzzy relational kernel clustering with local scaling parameter learning," in *Proceedings of the 20th IEEE International Workshop on Machine Learning for Signal Processing*, Kittila, Finland, August 29-September 1, 2010, pp. 289-294. http://dx.doi.org/10.1109/MLSP.2010.5589234

[30] O. Bchir and H. Frigui, "Fuzzy clustering with learnable cluster dependent kernels," in *IEEE International Conference on Fuzzy Systems*, Taipei, Taiwan, June 27-30, 2011, pp. 2521-2527. http://dx.doi.org/10.1109/FUZZY.2011.6007411

[31] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27-72, Jan. 2004.

[32] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, July 4-8, 2004, pp. 41-48.

[33] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "More efficiency in multiple kernel learning," in *Proceedings of the 24th International Conference on Machine Learning*, Corvalis, OR, June 20-24, 2007, pp. 775-782. http://dx.doi.org/10.1145/1273496.1273594

[34] J. Ye, S. Ji, and J. Chen, "Learning the kernel matrix in discriminant analysis via quadratically constrained quadratic programming," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, CA, August 12-15, 2007, pp. 854-863. http://dx.doi.org/10.1145/1281192.1281283

[35] M. Gönen and E. Alpaydin, "Localized multiple kernel learning," in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, July 5-9, 2008, pp. 352-359.

[36] B. Zhao, J. T. Kwok, and C. Zhang, "Multiple kernel clustering," in *Proceedings of the 9th SIAM International Conference on Data Mining, Sparks*, NV, April 30-May 2, 2009, pp. 638-649.

[37] H. C. Huang, Y. Y. Chuang, and C. S. Chen, "Multiple kernel fuzzy clustering," *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 1, pp. 120-134, Feb. 2012. http://dx.doi.org/10.1109/TFUZZ.2011.2170175

[38] N. Baili and H. Frigui, "Fuzzy clustering with multiple kernels in feature space," in *IEEE International Conference on Fuzzy Systems*, Brisbane, Australia, June 10-15, 2012. http://dx.doi.org/10.1109/FUZZ-IEEE.2012.6251146

[39] N. Baili and H. Frigui, "Relational fuzzy clustering with multiple kernels," in *Proceedings of the 11th IEEE International Conference on Data Mining*, Vancouver, Canada, December 11, 2011, pp. 488-495. http://dx.doi.org/10.1109/ICDMW.2011.145

[40] R. J. Hathaway, J. W. Davenport, and J. C. Bezdek, "Relational duals of the c-means clustering algorithms," *Pattern Recognition*, vol. 22, no. 2, pp. 205-212, 1989.

[41] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information," in *The 17th Annual Conference on Neural Information Processing Systems*, Vancouver & Whistler, Canada, December 8-13, 2003.

[42] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," in *The 19th Annual Conference on Neural Information Processing Systems*, Vancouver & Whistler, Canada, December 5-10, 2005.

[43] A. Globerson and S. Roweis, "Metric learning by collapsing classes," in *The 19th Annual Conference on Neural Information Processing Systems*, Vancouver & Whistler, Canada, December 5-10, 2005.

[44] S. Shalev-Shwartz, Y. Singer, and A. Y. Ng, "Online and batch learning of pseudo-metrics," in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, July 4-8, 2004, p. 94.

[45] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, June 20-24, 2007, pp. 209-216.

[46] R. Chatpatanasiri, T. Korsrilabutr, P. Tangchanachaianan, and B. Kijsirikul, "On kernelization of supervised Mahalanobis distance learners," arXiv:0804.1441. http://arxivorg/abs/08041441

[47] H. Zhang and J. Lu, "Semi-supervised fuzzy clustering: a kernel-based approach," *Knowledge-Based Systems*, vol. 22, no. 6, pp. 477-481, Aug. 2009. http://dx.doi.org/10.1016/j.knosys.2009.06.009

[48] O. Bchir, H. Frigui, and M. M. Ben Ismail, "Semi-supervised clustering and local scale learning algorithm," in *World Congress on Computer and Information*

Technology, Sousse, Tunisia, June 22-24, 2013, article number 6618774. http://dx.doi.org/10.1109/WCCIT.2013.6618774

[49] O. Bchir, H. Frigui, and M. M. B. Ismail, "Semi-supervised fuzzy clustering with learnable cluster dependent kernels," *International Journal on Artificial Intelligence Tools*, vol. 22, no. 3, article number 1350013, Jun. 2013. http://dx.doi.org/10.1142/S0218213013500139

[50] N. Baili and H. Frigui, "Semi-supervised clustering with cluster-dependent multiple kernels," in *The 4th International Conference on Information*, Intelligence, Systems and Applications Piraeus, Greece, July 10-12, 2013.

[51] J. H. Chiang and P. Y. Hao, "A new kernel-based fuzzy clustering approach: support vector clustering with cell growing," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 4, pp. 518-527, Aug. 2003. http://dx.doi.org/10.1109/TFUZZ.2003.814839

[52] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443-1471, Jul. 2001. http://dx.doi.org/10.1162/089976601750264965

[53] F. C. H. Rhee, K. S. Choi, and B. I. Choi, "Kernel approach to possibilistic C-means clustering," *International Journal of Intelligent Systems*, vol. 24, no. 3, pp. 272-292, Mar. 2009. http://dx.doi.org/10.1002/int.20336

[54] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, pp. 98-110, May 1993. http://dx.doi.org/10.1109/91.227387

[55] H. Shen, J. Yang, S. Wang, and X. Liu, "Attribute weighted mercer kernel based fuzzy clustering algorithm for general non-spherical datasets," *Soft Computing*, vol. 10, no. 11, pp. 1061-1073, Sep. 2006. http://dx.doi.org/10.1007/s00500-005-0043-5

[56] H. Frigui and O. Nasraoui, "Unsupervised learning of prototypes and attribute weights," *Pattern Recognition*, vol. 37, no. 3, pp. 567-581, Mar. 2004. http://dx.doi.org/10.1016/j.patcog.2003.08.002

[57] M. Sato-Ilic, S. Ito, and S. Takahashi, "Generalized kernel fuzzy clustering model," in *IEEE International Conference on Fuzzy Systems*, Jeju, Korea, August 20-24, 2009, pp. 421-426. http://dx.doi.org/10.1109/FUZZY.2009.5276876

[58] M. Sato and Y. Sato, "On a general fuzzy additive clustering model," *Intelligent Automation & Soft Computing*, vol. 1, no. 4, pp. 439-448, Jan. 1995. http://dx.doi.org/10.1080/10798587.1995.10750648

[59] M. R. P. Ferreira and F. D. A. T. de Carvalho, "Kernel fuzzy clustering methods based on local adaptive distances," in *IEEE International Conference on Fuzzy Systems*, Brisbane, Australia, June 10-15, 2012, pp. 1-8. http://dx.doi.org/10.1109/FUZZ-IEEE.2012.6251352

[60] D. D. Nguyen, L. T. Ngo, and L. T. Pham, "GMKIT2-FCM: a genetic-based improved multiple kernel interval type-2 fuzzy C-means clustering," in *IEEE International Conference on Cybernetics*, Lausanne, Switzerland, June 13-15, 2013, pp. 104-109. http://dx.doi.org/10.1109/CYBConf.2013.6617457

[61] J. A. Abhishek and F. C. H. Rhee, "Interval type-2 fuzzy C-means using multiple kernels," in *IEEE International Conference on Fuzzy Systems*, Hyderabad, India, July 7-10, 2013, pp. 1-8. http://dx.doi.org/10.1109/FUZZ-IEEE.2013.6622306

[62] M. A. Raza and F. C. H. Rhee, "Interval type-2 approach to kernel possibilistic C-means clustering," in *IEEE International Conference on Fuzzy Systems*, Brisbane, Australia, June 10-15, 2012, pp. 1-7. http://dx.doi.org/10.1109/FUZZ-IEEE.2012.6251233

[63] K. Lin, "A novel evolutionary kernel intuitionistic fuzzy C-means clustering algorithm," *IEEE Transactions on Fuzzy Systems*, vol. PP, no. 99, article number TFS-2013-0121.R2, Aug. 2013. http://dx.doi.org/10.1109/TFUZZ.2013.2280141

[64] K. T. Atanassov, "Intuitionistic fuzzy sets," *Fuzzy Sets and Systems*, vol. 20, no. 1, pp. 87-96, Aug. 1986. http://dx.doi.org/10.1016/S0165-0114(86)80034-3

[65] R. J. Hathaway and J. C. Bezdek, "Extending fuzzy and probabilistic clustering to very large data sets," *Computational Statistics & Data Analysis*, vol. 51, no. 1, pp. 215-234, Nov. 2006. http://dx.doi.org/10.1016/j.csda.2006.02.008

[66] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, "Fuzzy c-Means algorithms for very large data," *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 6, pp. 1130-1146, Dec. 2012. http://dx.doi.org/10.1109/TFUZZ.2012.2201485

[67] T. C. Havens, J. C. Bezdek, and M. Palaniswami, "Incremental kernel fuzzy c-means," in *Computational Intelligence, Studies in Computational Intelligence vol 399*, K. Madani, A. D. Correia, A. Rosa, and J. Filipe, Eds. Heidelberg: Springer Berlin, 2012, pp. 3-18. http://dx.doi.org/10.1007/978-3-642-27534-0_1

[68] N. Baili and H. Frigui, "Incremental fuzzy clustering with multiple kernels," *ATSP under review*, 2014.

Hichem Frigui received the Ph.D. degree in computer engineering and computer science from the University of Missouri, Columbia, in 1997. From 1998 to 2004, he was an Assistant Professor with the University of Memphis, Memphis, TN. He is currently a Professor and the Director of the Multimedia Research Lab, University of Louisville, Louisville, KY. He has been active in the research fields of fuzzy pattern recognition, data mining, and image processing with applications to content-based multimedia retrieval and land mine detection. He has participated in the development, testing, and real-time implementation of several land mine detection systems. He has published over 160 journal and refereed conference articles. Dr. Frigui has received the National Science Foundation Career Award for outstanding young scientists.

Ouiem Bchir is an Assistant professor in the Computer Science Department, College of Computer and Information Systems (CCIS), King Saud University, Riyadh, Saudi Arabia. She got her PHD degree from the University of Louisville, KY, USA. Her research interests include kernel clustering and learning, pattern recognition, and hyperspectral image analysis.

Naouel Baili received the Ph.D. degree from the Department of Computer Engineering and Computer Science, University of Louisville, USA, in 2013. She is currently a data scientist with Quintiles, USA. Her research interests include pattern recognition, computer vision, multimedia analysis and machine learning.