Applied Mathematics & Information Sciences
An International Journal

A Robust Fuzzy Kernel Clustering Algorithm

Zhang Chen[1], *Xia Shixiong*[1], *Liu Bing*[1]

[1]School of Computer Science and Technology, China University of Mining and Technology, Xu Zhou, 221116, China

Received: 9 Sep. 2012, Revised: 2 Dec. 2012, Accepted: 16 Dec. 2012
Published online: 1 Feb. 2013

Abstract: Traditional fuzzy kernel clustering methods does Iterative clustering in the original data space or in the feature space by mapping the samples into high-dimensional feature space through a kernel function These methods with normalized fuzzy degree of membership has weak robustness against noises and outliers, and lack of effective kernel parameter selection method. To overcome these problems, a robust kernel clustering algorithm is proposed to enhance the robustness by using typical parameter. Meawhile, a kernel function parameter optimization method under the unsupervised condition is also proposed in this paper. The experimental results show that the new algorithm is not only effective to the linear inseparable datasets with noisy data, but also more robust compared with other similar clustering algorithms and can obtain better clustering accuracy under noise jamming.

Keywords: Kernel function; fuzzy kernel clustering; kernel optimization; robustness.

1. Introduction

Fuzzy C- means clustering (FCM) algorithm is one of the most widely used clustering algorithms [1]. FCM and its improved algorithms have been widely used in pattern recognition, data mining, image processing and other fields, due to their simple design and low complexity [2-5].

Although the FCM algorithm improved the clustering effect of the partially overlapping datasets, the algorithm requires that the sum of various degree of membership of each sample point is 1. Therefore, FCM is sensitive to the outliers or the noise jamming. Literature [6] introduced the uncertainty degree of membership, but only limited to 0 and 1, and the interference from noise is still large. Krishnapuram and Keller proposed the possibilitic c-means clustering (PCM) algorithm, which loosened the normalized constraint and enhanced the robustness using the typical parameters, but would easily produce a consistent clustering. Pal and et al. proposed the possibilistic fuzzy c-means clustering (PFCM) algorithm, which combines FCM and PCM [7], which although solved consistency clustering and noise sensitive problem, but has a slow convergence rate. Zhang and Leung proposed the improved possibilistic c-means (IPCM) algorithm [8], which has strong robustness and fast convergence rate. Literature [9] further accelerated the convergence rate of the IPCM algorithm.

In a noisy environment, the above algorithms, using the Euclidean metric clustering methods, are sometimes not stable, and also, they are sensitive to the initial clustering center, cluster shapes and sizes, and cannot deal with nonlinear data as well. Therefore, some algorithms based on the kernel function have been put forward in succession, among which, a kind of algorithms enhance the robustness using the kernel function distance, for example, Literature [10] proposed the alternative fuzzy c-means (AFCM) algorithm based on the FCM algorithm, and similar algorithms were proposed in Literature [11-13]. Literature [14] proposed kernel possibilistic c-means (KPCM) algorithm based on the PCM algorithm. Literature [15] presented a possibilistic fuzzy clustering algorithm based on the kernel function, which used the Gauss kernel function to design the distance based on the PFCM, overcoming the shortcomings of the FCM and the PCM algorithm, and obtaining good results for the linear dataset in the original data space. In order to process the nonlinear data, the other kind of kernel clustering algorithms map the samples into a high-dimensional feature space using the kernel function, then do clustering in high-dimensional feature space. For example, Literatures [16,17] proposed hard partition kernel clustering algorithm, based on which Literature [18] presented fuzzy kernel clustering with outliers (FKCO) algorithm, a soft partition clustering algorithm, which can obtain better clustering accuracy for

* Corresponding author e-mail: xiasx@cumt.edu.cn

linearly inseparable data, however, it is sensitive to the initial clustering center and noise data, and would easily obtain the local optimum solutions. The fuzzy kernel c-means (FKCM) algorithms proposed in [19] and the algorithms proposed in [20-23] are mostly based on the FCM algorithm. They have bad robustness, and have not discussed the parameter optimization problem of the kernel function.

When a smooth and continuous nonlinear kernel function is used to map the sample dataset into a high-dimensional feature space, the topological structure of the samples in the original data space holds the line, so the noise or outliers in high-dimensional feature space will have a big effect on clustering. To further enhance the robustness of the kernel clustering algorithm, a possibilistic fuzzy clustering algorithm is proposed, the convergence of the algorithm is proved, and a optimization method for the parameter of the kernel function under the unsupervised conditions is presented. The experimental results show that, compared with other similar kernel clustering algorithm, this algorithm can not only deal with the linear inseparable and partially overlapping dataset, but also get better clustering accuracy under the noise interference.

2. Possibilistic fuzzy clustering algorithm

2.1. *Description of the algorithm*

In order to solve the sensitive issues of the fuzzy kernel clustering algorithm to the initial clustering center and noise data, and to enhance the robustness, the kernel method is extended to the possibilistic fuzzy clustering algorithm, namely the fuzzy membership and the typical value matrix are used in the mapped high-dimensional feature space, and in the iterative updates of the algorithm. In the high-dimensional feature space H, the objective function is defined as:

$$\begin{aligned} J_{KIPCM}(T,U,V) = & \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} t_{ij}^{p} \left\|\Phi(x_j) - v_i^{\Phi}\right\|^2 \\ & + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{n} u_{ij}^{m} (1 - t_{ij})^{p} \end{aligned} \tag{1}$$

where, $x_j \in R^N (j = 1, 2, \ldots n)$ is a sample vector in the sample space, $\Phi_j)$ is the mapping of x_j in the feature space, v_i^{Φ} is the i^{th} center in the feature space, and U is the fuzzy membership matrix. u_{ij} is a fuzzy membership. which is to express the relative degree of the j^{th} sample belonging to the i^{th} class. T is the typical value matrix or the possibilistic membership matrix, wherein the element t_{ij} is a typical value or a possibilistic membership used to denote the absolute degree of the j^{th} sample belonging to the i^{th} class. V is a center point set, m is weight of the fuzzy membership, p is its weight index, and η_i is a right positive value, which is defined as:

$$\eta_i = K \frac{\sum\limits_{j=1}^{n} u_{ij}^{m} t_{ij}^{p} \left\|\Phi(x_j) - v_i^{\Phi}\right\|^2}{\sum\limits_{j=1}^{n} u_{ij}^{m} t_{ij}^{p}} \tag{2}$$

K usually is set as 1.

The necessary conditions for Equation (1) to obtain the minimum value are as follows:

$$t_{ij} = \frac{1}{1 + \left(\frac{\left\|\Phi(x_j) - v_i^{\Phi}\right\|^2}{\eta_i}\right)^{\frac{1}{p-1}}} \tag{3}$$

$$u_{ij} = \frac{1}{\sum\limits_{k=1}^{c} \left(\frac{t_{ij}^{p-1}\left\|\Phi(x_j) - v_i^{\Phi}\right\|^2}{t_{kj}^{p-1}\left\|\Phi(x_j) - v_k^{\Phi}\right\|^2}\right)^{\frac{1}{m-1}}} \tag{4}$$

$$v_i^{\Phi} = \frac{\sum\limits_{j=1}^{n} u_{ij}^{m} t_{ij}^{p} \Phi(x_j)}{\sum\limits_{j=1}^{n} u_{ij}^{m} t_{ij}^{p}} \tag{5}$$

Where, v_i^{Φ} cannot be calculated directly. When the Gauss kernel function is used, the following method can be used to indirectly solve u_{ij} and t_{ij}.

The sample and the distance of the i^{th} class center in the feature space are calculated as follows:

$$\begin{aligned} D_{ij}^2 = \left\|\Phi(x_j) - v_i^{\Phi}\right\|^2 = K(x_j, x_j) - 2K(x_j, \hat{v}_i) \\ + K(\hat{v}_i, \hat{v}_i) \end{aligned} \tag{6}$$

Where,

$$K(x_j, x_j) = \Phi(x_j) \cdot \Phi(x_j) = \exp\left(\frac{\left\|x_j - x_j\right\|^2}{2\sigma^2}\right) = 1 \tag{7}$$

$$\begin{aligned} K(x_j, \hat{v}_i) &= \Phi(x_j) \cdot v_i^{\Phi} = \Phi(x_j) \cdot \frac{\sum\limits_{k=1}^{n} u_{ik}^{m} t_{ik}^{p} \Phi(x_k)}{\sum\limits_{k=1}^{n} u_{ik}^{m} t_{ik}^{p}} \\ &= \frac{\sum\limits_{k=1}^{n} u_{ik}^{m} t_{ik}^{p} K(x_j, x_k)}{\sum\limits_{k=1}^{n} u_{ik}^{m} t_{ik}^{p}} \end{aligned} \tag{8}$$

$$\begin{aligned} K(\hat{v}_i, \hat{v}_i) &= v_i^{\Phi} \cdot v_i^{\Phi} \\ &= \frac{\sum\limits_{j=1}^{n} u_{ij}^{m} t_{ij}^{p} \Phi(x_j)}{\sum\limits_{j=1}^{n} u_{ij}^{m} t_{ij}^{p}} \cdot \frac{\sum\limits_{j=1}^{n} u_{ij}^{m} t_{ij}^{p} \Phi(x_j)}{\sum\limits_{j=1}^{n} u_{ij}^{m} t_{ij}^{p}} \end{aligned}$$

$$= \frac{\sum\limits_{j=1}^{n}\sum\limits_{k=1}^{n} u_{ij}^m t_{ij}^p u_{ik}^m t_{ik}^p K(x_j, x_k)}{(\sum\limits_{j=1}^{n}(u_{ij}^m t_{ij}^p))^2} \tag{9}$$

Therefore, through solving Equation (8) and (9), without the explicit solution of v_i^{Φ}, the fuzzy membership matrix and the typical value matrix can be updated. The possibilistic fuzzy kernel clustering algorithm is described as follows.

Possibilistic fuzzy kernel clustering algorithm (PFKCA):

Step 1: The value of$c$, m,$p$are fixed, $1 < c < n$, $1 < m < +\infty$, $1 < p < +\infty$, the initial value of the loop is set as $r = 1$ and the maximum cycle number is r_{max}, and the threshold of stopping the algorithm is ε;

Step 2: The KFCM algorithm is run, and the obtained clustering centers is used as the initial clustering center V$^{(0)}$, the obtained fuzzy membership matrix is used as the initial fuzzy membership matrix U$^{(0)}$, the obtained kernel function values $K(x_j, \hat{v_i})$ and $K(\hat{v_i}, \hat{v_i})$ are used as the initial kernel function value; at the same time, the parameters of the Gauss kernel function are optimized, and the optimization parameter σ is obtained (the specific optimization method of the kernel function is described in Section 1.3).

Step 3: The parameter $\eta_i$is calculated according to Equation (2);

Step 4: Using the initial clustering center V$^{(0)}$, the fuzzy membership matrix U$^{(0)}$, the parameter η_i and the optimized Gauss kernel parameter σ, the loop is run as follows:

The distance between the sample vector in the high-dimensional feature space and the clustering center is updated using Equation (6), (7), (8) and (9);

The typical value matrix T$^{(r)}$is updated using Equation (3);

The fuzzy membership matrix U$^{(r)}$ is updated using Equation (4);

The loop variable $r$is added 1;

Until the conditions $||$V$^{(r)}$-V$^{(r-1)}||<\varepsilon$ or $r>r_{max}$ are satisfied

The following is the convergence proof of possibilistic fuzzy kernel clustering algorithm.

Theorem 1. In the possibilistic fuzzy kernel clustering algorithm, the necessary condition of $U = [u_{ij}]_{c\times n}$ and $T = [t_{ij}]_{c\times n}, V = [\varphi(v_i)]_{1\times c}$ being J_K local optimum, is

$$u_{ij} = \left(\sum_{k=1}^{c}(\frac{t_{ij}^{p-1}D_{ij}^2}{t_{kj}^{p-1}D_{kj}^2})^{\frac{1}{m-1}}\right)^{-1}$$

$$t_{ij} = \left(1 + (\frac{D_{ij}^2}{\eta_i})^{\frac{1}{p-1}}\right)^{-1},$$

u_{ij} satisfies the constraint conditions $\sum\limits_{i=1}^{c} u_{ij} = 1 \forall j$. $(i = 1, 2,..., c, j = 1, 2, ...n)$.

Proof: under the constraint condition $\sum\limits_{i=1}^{c} u_{ij} = 1$, the minimum of the objective function $J_K(T, U, V)$ is calculated using the Lagrange multiplier method , the Lagrange function is obtained as follows:

$$L(T, U, V, \lambda) = J_K(T, U, V) - \sum_{j=1}^{n}\lambda_j(\sum_{i=1}^{c} u_{ij} - 1)$$

$$= \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m t_{ij}^p \left\|\Phi(x_j) - \Phi(v_i)\right\|^2$$

$$+ \sum_{i=1}^{c}\eta_i \sum_{j=1}^{n} u_{ij}^m (1 - t_{ij})^p - \sum_{j=1}^{n}\lambda_j(\sum_{i=1}^{c} u_{ij} - 1) \tag{10}$$

The partial differential equations are solved:

$$\frac{\partial L(T, U, V, \lambda)}{\partial u_{ij}} = m u_{ij}^{m-1} t_{ij}^p D_{ij}^2 + \eta_i m u_{ij}^{m-1}(1 - t_{ij})^p - \lambda_j = 0 \tag{11}$$

$$\frac{\partial L(T, U, V, \lambda)}{\partial t_{ij}} = p t_{ij}^{p-1} u_{ij}^m D_{ij}^2 - \eta_i p (1 - t_{ij})^{p-1} u_{ij}^m = 0 \tag{12}$$

$$\frac{\partial L(T, U, V, \lambda)}{\partial \lambda_j} = \sum_{i=1}^{c} u_{ij} - 1 = 0 \tag{13}$$

The iteration formula of the typical value is obtained using Equation (12),

$$t_{ij} = \frac{1}{1 + (\frac{D_{ij}^2}{\eta_i})^{\frac{1}{p-1}}}, \quad \forall i, j \tag{14}$$

Then according to Equation (11),

$$u_{ij} = \left[\frac{\lambda_j}{m(t_{ij}^p D_{ij}^2 + \eta_i(1 - t_{ij})^p}\right]^{\frac{1}{m-1}} \tag{15}$$

According to Equation (12),

$$t_{ij}^{p-1} D_{ij}^2 (1 - t_{ij}) = \eta_i (1 - t_{ij})^p \tag{16}$$

Equation (7) is substituted into Equation (11), then

$$u_{ij} = \left[\frac{\lambda_j}{m t_{ij}^{p-1} D_{ij}^2}\right]^{\frac{1}{m-1}} \tag{17}$$

$$\sum_{k=1}^{c}(\frac{\lambda_j}{m t_{kj}^{p-1} D_{kj}^2})^{\frac{1}{m-1}} = 1 \tag{18}$$

$$(\frac{\lambda_j}{m})^{\frac{1}{m-1}} = \frac{1}{\sum\limits_{k=1}^{c}(\frac{1}{t_{kj}^{p-1}D_{kj}^2})^{\frac{1}{m-1}}} \tag{19}$$

Equation (19) is substituted into Equation (17), and then the iteration formula of the membership is:

$$u_{ij} = \frac{1}{\sum\limits_{k=1}^{c} (\frac{t_{ij}^{p-1} D_{ij}^2}{t_{kj}^{p-1} D_{kj}^2})^{\frac{1}{m-1}}}, \forall i, j, Q.E.D. \qquad (20)$$

Theorem 2. Let $\varphi(U) = J_K$, where $U = [u_{ij}]_{c\times n}$ satisfies the constraint conditions $\sum\limits_{i=1}^{c} u_{ij} = 1 \quad \forall j$, $T = [t_{ij}]_{c\times n}$ are fixed, and for all $1 \le i \le c, 1 \le j \le n$, $m > 1, p > 1, D_{ij}^2 > 0$ exists, then U is a local optimum of $\varphi(U)$, if and only if $u_{ij} (i = 1, 2, ..., c, j = 1, 2, ...n)$ are calculated by Equation (4).

Proof: The necessity has been proven by Theorem 1. To prove its sufficiency, the Hessian matrix $H(\varphi(U))$ of $\varphi(U)$ is obtained using the Lagrange function (10):

$$h_{mn,ij}(U) = \frac{\partial}{\partial u_{mn}} \left[\frac{\partial \varphi(U)}{\partial u_{ij}} \right] = \begin{cases} m(m-1) t_{ij}^p u_{ij}^{m-2} D_{ij}^2, & \text{if} m = i, n = j \\ 0, & \text{otherwise} \end{cases} \qquad (21)$$

According to Equation (21), $H(\varphi) = h_{mn,ij}(U)$ is a diagonal matrix. For all $1 \le i \le c, 1 \le j \le n, t_{ij}, u_{ij}$ are separately calculated by Equation (3) and (4), $u_{ij} > 0, t_{ij} > 0, m > 1, D_{ij}^2 > 0$, The above Hessian matrix is a positive definite matrix. So Equation (4) is the sufficient condition to minimize $\varphi(U)$.

Theorem 3. Let $\varphi(T) = J_K$, where $T = [t_{ij}]_{c\times n}, U = [u_{ij}]_{c\times n}$ are fixed and satisfies the constraint conditions $\sum\limits_{i=1}^{c} u_{ij} = 1 \, \forall j$, for all $1 \le i \le c, 1 \le j \le n, m > 1, p > 1$, $D_{ij}^2 > 0$ exists, then $T$is a local optimum of $\varphi(T)$, if and only if t_{ij} $(i = 1, 2, ..., c, j = 1, 2, ...n)$ are calculated by Equation (3).

Proof: The necessity has been proven by Theorem 1. The sufficiency proof is same as Theorem 2, the Hessian matrix $H(\varphi(T))$ of $\varphi(T)$ is obtained using the Lagrange function (10):

$$h_{mn,ij}(T) = \frac{\partial}{\partial t_{mn}} \left[\frac{\partial \varphi(T)}{\partial t_{ij}} \right] = \begin{cases} p(p-1) u_{ij}^m (t_{ij}^{p-2} D_{ij}^2 + \eta_i (1 - t_{ij})^{p-2}), & \text{if } m = i, \\ & n = j \\ 0, & \text{otherwise} \end{cases} \qquad (22)$$

According to Equation (22), $H(\varphi(T))$ is a diagonal matrix. For all $1 \le i \le c, 1 \le j \le n$, t_{ij}, u_{ij} are calculated by Equation (3) and (4), respectively, and $u_{ij} > 0, 0 < t_{ij} < 1, p \ge 2, \eta_i > 0, D_{ij}^2 > 0$, The above Hessian matrix is a positive definite matrix. So Equation (3) is the sufficient condition to minimize $\varphi(T)$.

According to Theorem 1 and 2, $J_K(U^{t+1}, T^{t+1}) \le J_K(U^t, T^t)$ can be proved, therefore, the possibilistic fuzzy kernel clustering algorithm will converge.

2.2. Optimization method of kernel function parameter

Based on the kernel parameter optimization idea proposed in Literature[24], an optimization method for Gauss kernel function parameter under the unsupervised case is presented. In this method, firstly the unsupervised kernel clustering results are used as prior knowledge to guide the kernel parameter optimization, then the optimized kernel parameters are used to do possibilistic fuzzy kernel clustering. The optimization process is as follows:

First of all, according to the initialization results of the KFCM algorithm, the subsets $X_k (k = 1, 2, \ldots c)$ where fuzzy membership is greater than a threshold value M are selected from each cluster. In this experiment, M is 0.9. Two pairwise constraint sets *ML* and *CL* are constructed from c subsets, where $ML = \{(x_i, x_j), x_i \in X_k, x_j \in X_k\}$ is a must-link constraint set denoting that two samples belong to the same category; $CL = \{(x_i, x_j), x_i \in X_l, x_j \in X_k, l \ne k\}$ is a cannot-link constraint set denoting that two samples are from different classes. *ML* and *CL* are used as a priori information to do kernel function learning. We expect that the learned kernel function should let two samples satisfying the must-link constraint condition be as close as possible after mapped into the high-dimensional feature space, and the two samples satisfying the cannot-link constraint conditions be separated as far as possible in the high-dimensional space. Therefore, the objective function is defined as follows:

$$F_{\ker nel} = \sum_{(x_i,x_j)\in CL} \|\Phi(x_i) - \Phi(x_j)\|^2 - \sum_{(x_i,x_j)\in ML} \|\Phi(x_i) - \Phi(x_j)\|^2 \qquad (23)$$

By maximizing the objective function, the effective kernel parameter can be found out, therefore the similarity between sample points can be more accurately expressed in the high-dimensional feature space, to obtain better clustering effect. The objective function is expanded as follows:

$$\sum_{(x_i,x_j)\in CL} \|\Phi(x_i) - \Phi(x_j)\|^2 = \sum_{(x_i,x_j)\in CL} \{K(x_i, x_i) + K(x_j, x_j) - 2K(x_i, x_j)\} = \sum_{(x_i,x_j)\in CL} \{2 - 2K(x_i, x_j)\}$$

$$= 2\mathrm{N}_{CL} - 2 \sum_{(x_i,x_j)\in CL} K(x_i,x_j) \tag{24}$$

Similarly available:

$$\begin{aligned} &\sum_{(x_i,x_j)\in ML} \|\Phi(x_i) - \Phi(x_j)\|^2 \\ &= 2\mathrm{N}_{ML} - 2 \sum_{(x_i,x_j)\in ML} K(x_i,x_j) \end{aligned} \tag{25}$$

Where N_{ML} and N_{CL} denote the number of constraint pair in a must-link constraint set and a cannot-link constraint set, respectively. After removing the constant terms, the objective function is obtained:

$$Fk = \sum_{(x_i,x_j)\in ML} K(x_i,x_j) - \sum_{(x_i,x_j)\in CL} K(x_i,x_j) \tag{26}$$

In this experiment, the Gauss kernel function is used to do mapping, and the gradient descent method is used for solving the Gauss kernel parameter σ. The process is as follows:

$$\frac{\partial Fk}{\partial \sigma} = \sum_{x_i,x_j\in ML} \frac{\partial K(x_i,x_j)}{\partial \sigma} - \sum_{x_i,x_j\in CL} \frac{\partial K(x_i,x_j)}{\partial \sigma} \tag{27}$$

Where,

$$\frac{\partial K(x_i,x_j)}{\partial \sigma} = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right)\frac{\|x_i - x_j\|^2}{\sigma^3}$$

$$\sigma^{(new)} = \sigma^{(old)} + \rho\frac{\partial Fk}{\partial \sigma} \tag{28}$$

Where ρ is a step parameter, which can be achieved through linear search method. The optimization algorithm is described as follows:

Step 1: Do initialization, the variance of the set $X = X_1 \cup X_1 \cup \ldots \cup$Xc is calculated to be as the initial value $\sigma 0$ of the Gauss kernel parameter.

Step 2: Equation (27) is used to calculate the gradient $\frac{\partial Fk}{\partial \sigma}$.

Step 3: The linear search method is used to obtain the step parameter ρ, the Gauss kernel parameter is updated according to Equation (28).

Step 4: Return to the second step to do iterative computation, until a local optimal solution of Fk is obtained.

Step 5: Output finally calculated Gauss kernel parameter σ.

2.3. Time complexity of the algorithm

In this paper, Nexpresses the sample number, C is the cluster number, and L is the loops. The time complexity of the KFCM algorithm is $O(N^2CL_1)$, where L_1 is the loops of the KFCM algorithm. The time complexity counting η_i is $O(CN^2)$, and the time complexity to update the typical value and fuzzy membership is $O(NCL_2)$, with L_2 be the loops of this algorithm. The time complexity of updating $K(x_j, \hat{v}_i)$ and $K(\hat{v}_i, \hat{v}_i)$ is $O(\mathrm{N}^2\mathrm{CL}_2)$. Therefore, the total time complexity of this algorithm is $O(N^2CL_1) + O(CN^2) + O(NCL_2) + O(N^2CL_2) = O(N^2CL)$. Literature [13] points out that the time complexity of the FCM type algorithm iterating in the original data space is O(NCL), however, the time complexity of the KFCM type algorithm iterating in the high-dimensional feature space is $O(N^2CL)$. These are consistent with the analysis for time complexity of the algorithm in this paper.

3. Experimental results and analysis

In order to test the running time and clustering accuracy of the algorithm, we've run the traditional FCM and PCM algorithm, two kinds of typical possibilistic fuzzy clustering algorithm IPCM and PFCM, KPFCM algorithm based on the kernel function distance, FKCM and FKCO algorithm based on Mercer kernel mapping, and the possibilistic fuzzy clustering algorithm PFKCA proposed in this paper, according to the standard IRIS datasets high-dimensional WINE dataset and artificial datasets. At the same time, in order to verify the robustness of the algorithm, algorithm is tested under noise jamming, and the results are analyzed to verify the validity of this method. The experimental parameters are as follows: ε =0.00001, the maximum number of the loop r$_{max}$=200, m=2.0, p=2.0, a=1, and b=1. The threshold of the membership M=0.9. Gauss kernel function parameter σ is optimized by the method described in Subsection 1.3. The computer configuration is: Intel core 2 Duo CPU, frequency 2.93GHz, and memory 2.00GB. Microsoft Visual C++6.0 and MATLAB tools are used to carry out simulation experiments.

3.1. Evaluation index of clustering results

A clustering algorithm generates a class label for each data point, the close degree between the generated class labels and the true class label need to be considered for the clustering performance evaluation of the algorithm. Two commonly used evaluation indexes are: Rand Index (*RI*) and Normalized Mutual Information index (*NMI*)[25].

3.2. Test on complex dataset

To further verify the effectiveness and the robustness of the PFKCA algorithm, two datasets (named Test1 and Test2) shown in Figure 1 are tested. These datasets are simultaneously composed of the linear inseparable data, partially overlapping data and noise data.

The Test1 dataset are made of 4000 two-dimensional data points that are divided into three classes, with each one containing 1000 sample points: one is the annular dataset with the center at the origin, which is uniformly distributed with the radius of 5, and is added by the Gauss noise with mean 0 and variance 0.1; the other two classes follow the following normal distributions, respectively:

Class2: $N(\begin{bmatrix}1\\0\end{bmatrix},\begin{bmatrix}0.2&0\\0&0.4\end{bmatrix})$, 1000 points;

Class 3: $N(\begin{bmatrix}3\\0\end{bmatrix},\begin{bmatrix}0.3&0\\0&0.6\end{bmatrix})$, 1000 points;

Class 1 and Class 2 or Class 3 are linear inseparable, and Class 2 and Class 3 are approximately linear separable. In addition, there are 1000 noise data, uniformly distributed in the region of [-6,6]×[-6,6].

There are 4000 data points in the Test2 dataset, divided into three classes, with each one containing 1000 sample points. Class 1 is the annular dataset with the center at the origin, with is uniformly distributed, with the radius of 3, and is added by the Gauss noise with mean 0 and variance 0.1, while the other two classes follow the following normal distributions:

Class2: $N(\begin{bmatrix}1\\0\end{bmatrix},\begin{bmatrix}0.2&0\\0&0.4\end{bmatrix})$, 1000 points;

Class 3: $N(\begin{bmatrix}3\\0\end{bmatrix},\begin{bmatrix}0.3&0\\0&0.6\end{bmatrix})$, 1000 points;

Class 1 and Class 2 are linear inseparable, and Class 3 is approximately linear separable with Class 1 or Class 2. In addition, there are 1000 noise data, uniformly distributed in the region of [-4,10]×[-4,4].

Noisytwins: twins+1000 uniform noise [-6,6]*[-6,6]

Ring: R=5, u=0, σ^2=0.1, 1000 points

Normal distribution 1: u=(2,0), σ^2=[0.5 0;0 0.5], 1000 points

Normal distribution 2: u=(-2,0), σ^2=[0.5 0;0 0.5], 1000 points

NoisyMirror: Mirror+1000 uniform noise $[-4, 10] * [-4, 4]$

Ring: $R = 3$, $u = 0$, σ^2=0.1, 1000 points

Normal distribution 1: $u = (0, 0)$, $\sigma^2 = [0.50; 00.5]$, 1000 points

Normal distribution 2: $u = (6, 0)$, $\sigma^2 = [0.50; 00.5]$, 1000 points

Let's summarize the above experimental results: for linear separable datasets or partially overlapping dataset, except PCM would easily produce consistent clustering, the clustering effects of the other algorithms have little differences, where the KPFCM, FKCM, FKCO and PFKCA algorithms based on the kernel function have more running time than the FCM, PCM, PFCM and IPCM algorithms based on the Euclidean distance; the FKCM, FKCO and PFKCA algorithms iterating in the high-dimensional feature space have more running time than KPFCM algorithm, while the IPCM, PFKCA algorithms have stronger robustness than the other algorithms; for linear inseparable dataset, the FCM, PCM, PFCM, IPCM and KPFCM algorithms iterating in the original data space are invalid, furthermore, under the condition of a random initial clustering center and noise interferences, FKCM and FKCO algorithms are sensitive and with poor clustering effect, but the PFKCA algorithm presented in this paper still has good clustering effect.

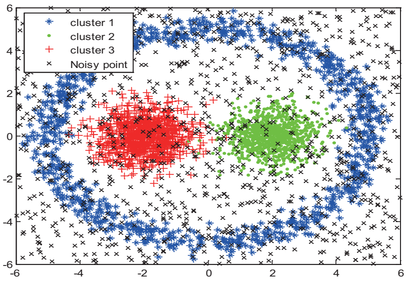

Figure 1 Test1 dataset with 1000 noises

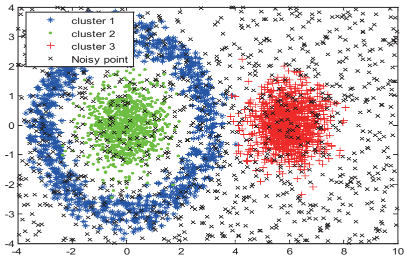

Figure 2 Test2 dataset with 1000 noises

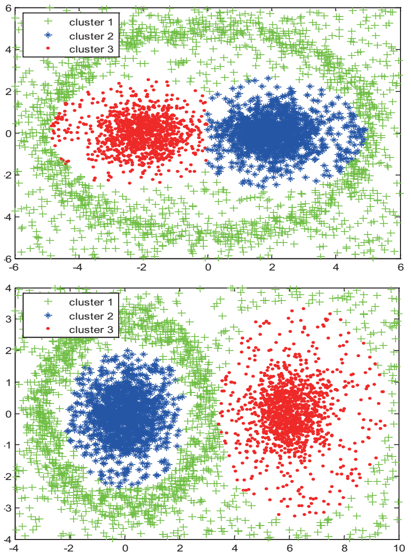

Figure 3 The clustering result of Test1

Table 1 Comparison of clustering performances on the linear inseparable dataset with 300 noisy points

Algorithms	Iterations	Time(s)	Errors	RI	NMI
FCM	262	3.36	847	0.72	0.56
PCM	18	4.05	946	0.68	0.69
PFCM	370	6.88	960 (963)	0.68 (0.68)	0.54 (0.56)
IPCM	272	12.50	1070 (994)	0.64 (0.67)	0.56 (0.56)
KPFCM ($\sigma 2$=10)	204	12.80	970 (757)	0.68 (0.75)	0.52 (0.69)
FKCM ($\sigma 2$=10)	33	1483.78	362	0.89	0.81
FKCO ($\sigma 2 = 10$, q=1, w=4000)	36	2394.55	298	0.92	0.85
PFKCA ($\sigma 2$=10)	48	8281.34	75(75)	0.98 (0.98)	0.90 (0.90)

4. Conclusion

In this paper, a new possibilistic fuzzy kernel clustering algorithm is proposed. At first, the samples in the sample space are mapped into the high-dimensional feature space using Mercer kernel function, then the possibilistic fuzzy clustering algorithm is used for clustering in the high-dimensional space. At the same time, the kernel function parameter optimization method under the unsupervised condition is presented, and compared with the other similar kernel clustering algorithm, the new algorithm can not only deal with the linearly inseparable dataset, and can get better clustering accuracy under noise jamming. Simulation results have proved the effectiveness of the possibilistic fuzzy kernel clustering algorithm. Since this algorithm has a weak point of high time complexity, fast clustering problems for a large linear inseparable dataset need to be studied, in addition, semi-supervised kernel clustering algorithm with noise robustness also needs to be future researched.

References

[1] Sun JG,Liu J,Zhao LY, Journal of Software **19**, 48-61 (2008).

[2] S.Kirindis and V. Chatzis, IEEE Trans. Image Process **19**, 1328-1337 (2010).

[3] W. Cai,S. Chen,D. Zhang,Pattern Recognition **40**, 825-838 (2007).

[4] Zhang M,Yu J, Journal of Software **15**, 858-869 (2004).

[5] Tian Jun-wei,Huang yong-xuan,Yu ya-lin,Pattern Recognition & Artificial Intelligence **21**, 221-226 (2008).

[6] Chen Jian-mei, Lu Hu, Journal of Computer Research and Development **45**, 1486-1492 (2008).

[7] N R Pal,K Pal, J C Bezdek,IEEE Trans Fuzzy Systems **13**, 517-530 (2005).

[8] J S Zhang,Y W Leung, IEEE Trans Fuzzy systems,2004 **2**, 209-217 (2004).

[9] Liu Bing, Xia Shi-xiong, Acta Electronica Sinica **40**, 371-375 (2012).

[10] Wu KL,Yang MS, Pattern Recognition **35**, 2267-2278 (2002).

[11] H. Shen,J. Yang,S. Wang , X. Liu, Soft Computing **10**, 1061-1073 (2006).

[12] D.Q. Zhang,S.C. Chen, Neural Processing Letters **18**, 155-162 (2003).

[13] D. Graves,W. Pedrycz, Fuzzy Sets and Systems **164**, 522-543 (2010).

[14] Wu Xiao-hong,Zhou Jian-jiang, Joint IAPR International Workshop on Structural and Syntactic Pattern Recognition and Statistical Techniques in Pattern Recognition,LNCS **4109**, 783-791 (2006).

[15] Meena Tushir,Smriti Srivastava, Applied Soft Computing **10**, 381-389 (2010).

[16] Girolami M, IEEE Trans on Neural Networks **13**, 780-784 (2002).

[17] Zhang L,Zhou WD,Jiao LC, Chinese Journal of Computers **25**, 587-590 (2002).

[18] Shen HB,Wang ST,Wu XJ, Journal of Software **15**, 1021-1029 (2004).

[19] WU Zhong-dong,GAO Xin-bo,XIE Wei-xin,Journal of Xidian University **31**, 533-537 (2004) .

[20] J.H. Chiang and P.Y. Hao, IEEE Transactions on Fuzzy Systems **11**, 518-527 (2003).

[21] L. Zeyu,T. Shiwei,X. Jing,J. Jun, Proceedings of the Internat. Society for Optical Engineering **11**, 241-245 (2001).

[22] Daoqiang Zhang, Songcan Chen, Proceedings of the International Conference on Control and Automation, 123-127 (2002).

[23] Shangming Zhou, John Q. Gan, Proceedings of the Intelligent Data Engineering and Automated Learning **3177**, 613-618 (2004).

[24] Xuesong Yin, Songcan Chen, Enliang Hu, Daoqiang Zhang, Pattern Recogniton **43**, 1320-1333 (2010).

[25] Weifu Chen,Guocan Feng, Neurocomputing **77**, 229-242 (2012).

Zhang Chen is current a Ph.D candidate at China University of Mining and Technology(CUMT), China. She received her MS degree in Computer Application Technology from CUMT in 2004, and her BS degree in Computer Science from CUMT in 2001. She is currently a lecture at school of Computer Science and Technology, CUMT. Her research interest is computation intelligence and machine learning.

Xia Shi Xiong is born in 1962, Ph.D. He is a professor at school of Computer Science and Technology in CUMT. He has published more than 60 research papers in journals and international conferences. His research interest is Wireless sensor networks and intelligent information processing et al.

Liu Bing is current a Ph.D candidate at China University of Mining and Technology(CUMT), China. He received his MS degree in Computer Application Technology from CUMT in 2005, and his BS degree in Computer Science from CUMT in 2002. He is currently a lecture at school of Computer Science and Technology, CUMT. His research interest is computation intelligence and machine learning et al.