Outlier Detection by Weighted Mercer-Kernel Based Fuzzy Clustering Algorithm

Hongbin Shen, Jie Yang, Yifei Dong, and Shitong Wang

Abstract—**Outliers are data values that lie away from the general cluster of other data values. Detecting the outliers of a dataset is an important research topic for data cleaning and finding new useful knowledge in many research areas, i.e. data mining, pattern recognition, etc. In the past decades, many useful algorithms were proposed in the literature. In this paper, a new fuzzy kernel-clustering algorithm with outliers (FKCO) is presented to locate critical areas that are often represented by only a few outliers. Theoretic analysis also shows that FKCO can converge to a local minimum of the objective function. Finally, based on the information theory, a new criterion for finding outliers is also proposed. Simulations of different types of datasets demonstrate the feasibility of this new method.**

Index Terms—**Fuzzy clustering analysis, kernel function, feature space, outliers.**

I. Introduction

Outliers are data values that lie away from the general cluster of other data values. Each outlier needs to be examined carefully to determine if it represents a possible value from the population being studied, in which case it should be retained, or if it is non-representative (or an error) in which case it can be excluded. There is a true story that the ozone hole above the South Pole had been detected by a satellite years before it was detected by ground-based observations, but the values were tossed out by a computer program because they were smaller than thought possible. The damage to our atmosphere caused by chloroflourocarbons went undetected and untreated for up to nine years because outliers were discarded without being examined [1]. So in recent years, how to locate outliers have attracted more and more attention. In past decades, several different methods have been proposed to attain this goal. Generally speaking, outliers can be found through subjective and objective measures. In a subjective case, a user directly applies their own knowledge or belief to determine the parameters like "very far away" and "low frequency". From that perspective, subjective methods may be unreliable, low scalability and may vary with users [2]. Furthermore, manual detection of outliers is also a very time-assuming task, not suitable for large datasets. For objective measures, no prior knowledge is needed to find the outliers. This type of measure is likely to be more reliable since no user's biased preference is given while locating outliers. In [3]-[4], the author proposed objective methods to find outliers based on graphical measure of constructing a box plot, which is a type of graph used to show the shape of the distribution, its central value, and variability. The picture produced consists of the most extreme values in the data set (maximum and minimum values), the lower and upper quartiles and the median.

In fact, other pattern recognition technology based objective ways to identify outliers have been proposed recently, such as the method based on fuzzy clustering techniques [5]. The author modified the objective function of Fuzzy C-means (FCM) clustering algorithm [6]-[14] of adding an additional weighting factor for each datum. The modified objective function to be optimized is shown as follows

$$J(X,U,v)=\sum_{i=1}^{C}\sum_{k=1}^{K}\mu_{ik}{}^{m}\cdot\frac{1}{w_k{}^{q}}\cdot d^2(\vec{v}_i,\vec{x}_k)\,, \qquad (1)$$

where $m>1$ is the fuzzy coefficient; μ_{ik} represents the membership degree of sample $\vec{x}_k$ belonging to the i-th cluster, C is the number of clusters and K denotes the number of data points in the dataset. Often, we take $d^2(\vec{x}_k,\vec{v}_i)=\|\vec{x}_k-\vec{v}_i\|^2$, which is the Euclidean distance; w_k denotes the weight factor of k-th sample and satisfy

$$\sum_{k=1}^{K}w_k=w,\qquad w_k>0\,, \qquad (2)$$

where $w>0$ is a user-defined real number of the total weight factor to be distributed. The obtained weights determine a kind of representativeness of each datum for the data distribution, which can be interpreted as the importance of the corresponding datum and more important datum will have larger weight. When the algorithm converges, outliers will hold higher weight value than other data values which can be used to construct a criteria for finding outliers described in the following section.

The methods of [5] can have a good result for linear separable dataset, but if the separation boundary between clusters is nonlinear, then the conventional methods discussed above will fail. An alternative method is to perform clustering in the kernel feature space. Mapping the observed data to a higher dimensional space in a nonlinear manner forms the basis for nonlinear classification techniques such as radial basis function networks, support vector machines (SVM), and certain forms of nonlinear discriminant analysis [15]-[18]. In this paper, we present a new algorithm denoted as FKCO (fuzzy kernel clustering

Manuscript received September 3, 2004, revised July 8, 2005. This work was supported in part by the China-France advanced program under Grant No. PRA SI 03-032

H. Shen is with the College of Computer & Information Engineering, Hohai University, Changzhou, Jiangsu, P. R. China (e-mail: zjshenhongbin@yahoo.com).

J. Yang is with the Department of Image Processing & Pattern Recognition, Jiaotong University, Shanghai, P. R. China (e-mail: jieyang@sjtu.edu.cn).

Y. Dong is with the School of Computer Science and Engineering, University of New South Wales, Australia.

Shitong Wang is with School of Information, Southern Yantze University, Wuxi, Jiangsu, P. R. China.

Publisher Item Identifier S 1682-0053(05)0305

with outliers), tending to locate the outliers of a dataset in the corresponding feature space. Firstly, the observed data is mapped to a higher-dimensional feature space with some proper kernel functions, and then assign a weight to each datum in the feature space while clustering in the feature space. By iterative function derived below, weight for each datum will be obtained, which finally can be used to construct a new criteria for finding the outlier information in the dataset. Different kinds of simulations including linear inseparable dataset and an image dataset demonstrate the feasibility of the method proposed in this paper.

II. Fuzzy Kernel Clustering Algorithm with Outliers

A. Mercer Kernel Functions

Let $\vec{x}_k \in R^N (k = 1,2,...,K)$ denotes the samples of the observed space. We can map the dataset to a higher dimensional space H through a nonlinear mapping function ϕ, denoted as $\phi(\vec{x}_1), \phi(\vec{x}_2), \ldots, \phi(\vec{x}_K)$. Then the dot product of vectors in feature space will be represented as Mercer kernel in the original space

$$K(\vec{x}_i, \vec{x}_j) = (\phi(\vec{x}_i) \cdot \phi(\vec{x}_j)). \quad (3)$$

All these samples will form a kernel function matrix $K_{ij} = K(\vec{x}_i, \vec{x}_j)$ [16]. This is the basis for non-linear classification techniques such as radial basis function networks, support vector machines, and certain forms of nonlinear discriminant analysis [15]-[18].

In the literature, the following three kernel functions are the most frequently used [16]

- Polynomial kernel function

$$K(\vec{x}, \vec{y}) = (\vec{x} \cdot \vec{y} + 1)^d \quad (4)$$

where d is a user-defined parameter.

- Gaussian kernel function

$$K(\vec{x}, \vec{y}) = \exp(\frac{-\| \vec{x} - \vec{y} \|^2}{2\sigma^2}) \quad (5)$$

where σ is the width of the Gaussian function.

- Two-level sigmoid kernel function

$$K(\vec{x}, \vec{y}) = \tanh(-b(\vec{x} \cdot \vec{y}) - c) \quad (6)$$

where b and c are user-defined parameters.

Up to now, there is no criterion for choosing the kernel functions. In most cases, gaussian kernel function is a better choice than the other two kernel functions, because an infinite-dimensional feature space will be obtained when it serves as the mapping function. In other words, any linear inseparable dataset in the observation space can be clustered linearly in the feature space.

B. Kernel Clustering Algorithm

Girolami introduced Mercer kernel-based clustering in feature space and the objective function was defined as [17]

$$Tr(S_w^{\varphi}) = \frac{1}{N} \sum_{k=1}^{C} \sum_{n=1}^{K} \mu_{kn} Q_{kn} \quad (7)$$

where

$$Q_{kn} = K_{nn} - \frac{2}{N_k} \sum_{j=1}^{K} \mu_{kj} K_{nj} + \frac{1}{N_k^2} \sum_{i=1}^{K} \sum_{l=1}^{K} \mu_{ki} \mu_{kl} K_{il} \quad (8)$$

where $K_{ij} = k(\vec{x}_i, \vec{x}_j)$ and $N_k = \sum_{n=1}^{K} \mu_{kn}$ denotes the kernel function.

Numerous work have shown that performing clustering technique in the feature space will solve the linear inseparable problem which is a obstacle of the conventional clustering analysis.

C. Fuzzy Kernel Clustering Algorithm with Outliers

As stated in [6], the topographic order will be preserved in the feature space, if the dataset is mapped to the feature space H by a Mercer kernel function ϕ, and ϕ will provide linear separation of classes. Considering then a smooth, continuous nonlinear mapping ϕ from data space to feature space H such that

$$\phi: \quad R^N \to H \qquad \vec{x} \to \vec{X}$$

Then we can rewrite the objective function of (1) in H as follows

$$J_H = \sum_{i=1}^{C} \sum_{k=1}^{K} \mu_{ik}^{\ m} \frac{1}{w_k^{\ q}} (\phi(\vec{x}_k) - \vec{m}_i^{\ \phi})(\phi(\vec{x}_k) - \vec{m}_i^{\ \phi})^T \quad (9)$$

where $\vec{m}_i^{\ \phi}$ denotes the center of the i-th cluster in the feature space.

We should notice that J_H takes the form of a series of inner products in feature space. As stated in previous section, the inner product can be easily computed through Mercer kernel. Through a specific kernel function, the inner product which it returns implicitly defines the nonlinear mapping ϕ to feature space [5]. So the objective function (17) can be rewritten solely with respect to the elements of the symmetric $K \times K$ kernel matrix as follows

$$J_H = \sum_{i=1}^{C} \sum_{k=1}^{K} \mu_{ik}^{\ m} \frac{1}{w_k^{\ q}} Q_{ik} \quad (10)$$

where

$$Q_{ik} = K_{kk} - \frac{2}{N_i} \sum_{j=1}^{K} \mu_{ij} K_{kj} + \frac{1}{N_i^2} \sum_{j=1}^{K} \sum_{l=1}^{K} \mu_{ij} \mu_{il} K_{jl} \quad (11)$$

where N_i, K_{ij}, μ_{ij} have the same meaning as before.

Thus, function (10) is considered as the objective to be optimized by FKCO, and Q_{ik} is the distance of the k-th sample to the i-th cluster center. The aim of FKCO is to add a small weighting value w_k (large value for $1/w_k^{\ q}$) to the datum that belongs to at least one of the classes. Generally speaking, outliers are far away from all the clusters, in this case, we will assign a large value w_k to each outlier (small value for $1/w_k^{\ q}$). The parameter q plays an important role in the clustering process. When q is large enough, then the weight value of each datum is almost equal to w/k, in other words, the weight plays the same influence on all the data samples; and if $q \to 0$, then the influence of the weight will reach the maximum.

Considering the constraint of (2), we have the following Lagrange function

$$J_H = \sum_{i=1}^{C} \sum_{k=1}^{K} \mu_{ik}^{\ m} \frac{1}{w_k^{\ q}} Q_{ik} + \lambda(\sum_{k=1}^{K} w_k - w) \quad (12)$$

differentiating (12) with respect to w_k, we obtain the following partial differentiation equation

$$\frac{\partial J_H}{\partial w_k} = -q \cdot \frac{1}{w_k^{q+1}} \cdot \sum_{i=1}^{C} \mu_{ik}{}^m \cdot Q_{ik} + \lambda \tag{13}$$

with $\partial J_H / \partial w_k = 0$, we will obtain

$$\lambda = q \cdot \frac{1}{w_k^{q+1}} \cdot \sum_{i=1}^{C} \mu_{ik}{}^m \cdot Q_{ik} \tag{14}$$

resolving (14) for w_k

$$w_k = \left(\frac{q \cdot \sum_{i=1}^{C} \mu_{ik}{}^m \cdot Q_{ik}}{\lambda}\right)^{\frac{1}{q+1}} \tag{15}$$

with constraint (3)

$$w = \sum_{k=1}^{K} \left(\frac{q \cdot \sum_{i=1}^{C} \mu_{ik}{}^m \cdot Q_{ik}}{\lambda}\right)^{\frac{1}{q+1}} \tag{16}$$

according to (15), we obtain the following equation

$$\lambda^{\frac{1}{q+1}} = \sum_{k=1}^{K} \left(q \cdot \sum_{i=1}^{C} \mu_{ik}{}^m \cdot Q_{ik}\right)^{\frac{1}{q+1}} \cdot \frac{1}{w} \tag{17}$$

so

$$\lambda = \left(\sum_{k=1}^{K} \left(q \cdot \sum_{i=1}^{C} \mu_{ik}{}^m \cdot Q_{ik}\right)^{\frac{1}{q+1}} \cdot \frac{1}{w}\right)^{q+1} \tag{18}$$

from (14) and (18), we can finally obtain the iterative function of w_k as follows

$$\begin{aligned} w_k &= \frac{\left(q \cdot \sum_{i=1}^{C} \mu_{ik}{}^m \cdot Q_{ik}\right)^{\frac{1}{q+1}}}{\sum_{k=1}^{K} \left(q \cdot \sum_{i=1}^{C} \mu_{ik}{}^m \cdot Q_{ik}\right)^{\frac{1}{q+1}}} \cdot w \\ &= \frac{\left(\sum_{i=1}^{C} \mu_{ik}{}^m \cdot Q_{ik}\right)^{\frac{1}{q+1}}}{\sum_{k=1}^{K} \left(\sum_{i=1}^{C} \mu_{ik}{}^m \cdot Q_{ik}\right)^{\frac{1}{q+1}}} \cdot w \end{aligned} \tag{19}$$

Similar to the famous fuzzy clustering algorithm FCM, the iterative function of the membership degree can be easily derived according to the distance tolerance Q

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{C} \left(\frac{Q_{ik}}{Q_{jk}}\right)^{\frac{1}{m-1}}}. \tag{20}$$

Based on the above derivations, the fuzzy kernel clustering algorithm with outliers (FKCO) can be described as follows:

Fuzzy Kernel Clustering Algorithm with Outliers (FKCO)

Step1. Initialize the parameters such as C, q, m for the algorithm and set the loop counter $t = 1$. Initialize the membership of each datum randomly.

Step2. Compute the distance $Q_{ik}{}^{(t)}$ of each datum to the cluster center according to (11).

Step3. Compute the membership $\mu_{ik}{}^{(t)}$ of each data to the i-th cluster.

Step4. Obtain the weight $w_k{}^{(t)}$ for each datum according to (19).

Step5. IF $abs(J_H(t+1) - J_H(t)) > \varepsilon$), increment t and go to Step 3, ELSE stop.

Theorem I: FKCO algorithm will finally converge to a locally minimum of the objective function (10) see appendix for the detailed proof.

III. Criteria for Finding Outliers

Here, we present a new fuzzy measurement of each datum in a dataset based on the fuzzy memberships obtained by FKCO. Suppose there is only one model in a given dataset, then obviously the membership $\mu_k = 1$ for points which overlap with the center of the model, while $0 \le \mu_k < 1$ for other member datum, apparently, the farer from the center, the smaller the membership μ_k. So similar to the famous Hartley tolerance method of information theory [19], we can define a new monotonous descending function as the representation of the datum

$$F(k) = \left(\frac{1 - \mu_k}{\mu_k}\right)^{\alpha} \tag{21}$$

where $\alpha \ge 1$ is the fuzzy coefficient, e.g., let α equals to 1 for simplicity in this paper.

This new fuzzy representation can be extended to multi-models easily as follows

$$F(k) = \min\left(\left(\frac{1-\mu_{1k}}{\mu_{1k}}\right)^{\alpha}, \left(\frac{1-\mu_{2k}}{\mu_{2k}}\right)^{\alpha}, \ldots\ldots, \left(\frac{1-\mu_{ck}}{\mu_{ck}}\right)^{\alpha}\right). \tag{22}$$

It is clear that the smaller μ_k, the larger $F(k)$. Usually, outliers which lie away from the general cluster of other data values will have much smaller membership, so they hold much higher information index $F(k)$ for outliers may be the most important feature of a dataset as discussed above and this character conforms to the basic information theory.

When FKCO finally converges, two attributes of each datum will be obtained in the dataset, one is the fuzzy representation derived from the membership and the other is the weight allocated to each data value which can be interpreted as the importance of the datum. Because outliers usually contain some important information and may be the most important feature of the dataset, i.e., outliers will have much higher weights than other data values and this also conforms to the results of Keller [5]. So each datum can be represented by a vector $\mathbf{v} = [w_k, F(k)]^T$, then we will take the inner product of the corresponding attribute vector as the new criteria for finding outlier information in the dataset:

$$S_k = \mathbf{v}^T \mathbf{v} = \left(w_k{}^2 + (F(k))^2\right)^{1/2} \tag{23}$$

Just as discussed above, the farer away the datum, the larger S_k and we will demonstrate below that the criteria for finding outliers of (23) should be more reliable and interpretable than the single weight factor in [5] which outlines the outliers only according to the weight factor and neglect the membership attribute of each datum with S_k,

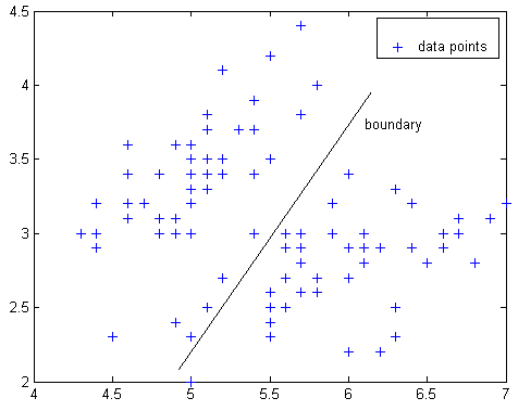

Fig. 1 The constructed two-dimensional iris data.

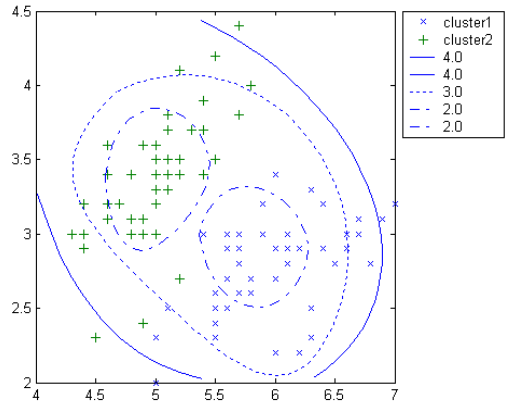

Fig. 3 Results of algorithm in [5] identifying outliers with the weights

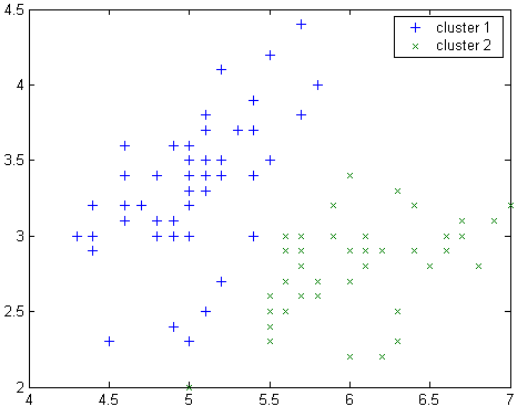

Fig. 2 The clustering result of FCM on the two-dimensional iris dataset.

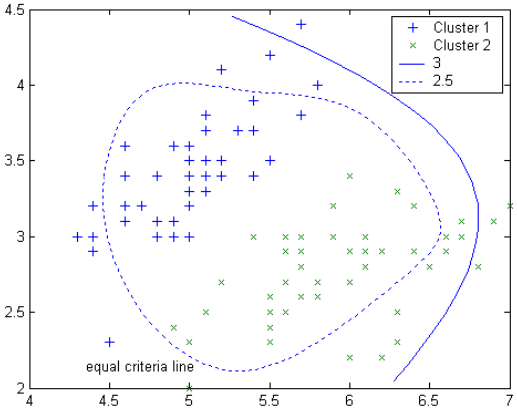

Fig. 4 Results of FKCO algorithm identifying outliers with S_k.

we can easily find the outlier information when $S_k > \Re$, where $\Re$ is the threshold given by experts.

IV. Simulations

In this section, four different datasets are used to illustrate the performance of FKCO. Experiment results will show the better behavior than the conventional methods, especially for the linear inseparable dataset, on which the conventional algorithms fail clustering. In our experiments, we suppose the parameters $q = 1, m = 2$ and $W = 200$ during clustering process.

A. Test on IRIS Dataset

Firstly, we use the famous iris dataset to investigate how FKCO behaves and performs on the clustering and finding out the outlier information of a linear separable and high dimensional datasets.

Case 1: For simplicity, we firstly select the first 2 dimensions of the IRIS data [10] in the first 2 groups as the testing dataset, and the dataset displays the attribute of linear separable as shown in Fig. 1. Three algorithms are tested on this simple iris dataset, Fig. 2 illustrates the clustering result of FCM; Figs. 3 and 4 show the performances of clustering and identifying outliers of algorithm in [5] and FKCO algorithm, respectively. As can be seen from Fig. 4, FKCO can obtain rational clustering results on this dataset. Based on S_k defined above, we can easily identify the outlier information according to the given threshold. Comparing the results of Fig. 3 with Fig. 4, it is not difficult for us to realize that FKCO with the new defined criteria deals with the outlier information more cautiously, e.g., we will get more outlier points in [5] more cautiously, e.g., we will get more outlier points in [5] (Fig. 3) than that of FKCO (Fig. 4) under the same threshold such as 3. This is in accordance with the viewpoints of dealing with outlier information by human and it is careful for us to accept the outlier information and surely will pay more attention.

TABLE I
CLUSTERING PERFORMANCE OF FKCO AND OTHER CLUSTERING ALGORITHMS ON 4-DIMENSIONAL IRIS DATASET

Algorithms	Number of samples clustered *wrongly*
FKCO	14
FCM	15
Algorithm in [5]	16

Case 2: We employ the complete iris dataset to test the performance of FKCO. Table I illustrates the clustering performance of FKCO and other algorithms.

From Table I, we can see clearly that FKCO has better clustering performance on IRIS dataset than that of the algorithm of [5] and FCM because of the more explicit features of IRIS dataset in the feature space.

To compare the importance index w and S_k, Figs. 5 and 6 show the weights of the 150 4-dimensional datum obtained by the algorithm in [5] and the new index S_k obtained by FKCO, respectively.

It has been generally accepted that there are no obvious outliers in the iris dataset, hence, the smaller of the interval scope for w or S_k, the better. For the weight vector $\mathbf{w} = [w_1, w_2, ..., w_N]^T$ of [5] and the new criteria vector $\mathbf{S} = [S_1, S_2, ..., S_N]^T$ of FKCO respectively, we compute the interval $V_W = \max_{i,j}(W_i - W_j)$ and $V_S = \max_{i,j}(S_i - S_j)$ respectively, and the results are shown in Table II.

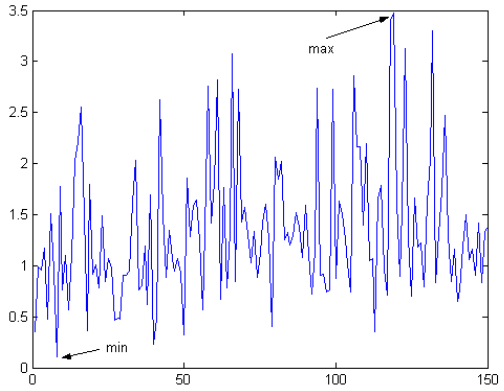

Fig. 5 weight factor of the 150 data of iris dataset of the algorithm in [5].

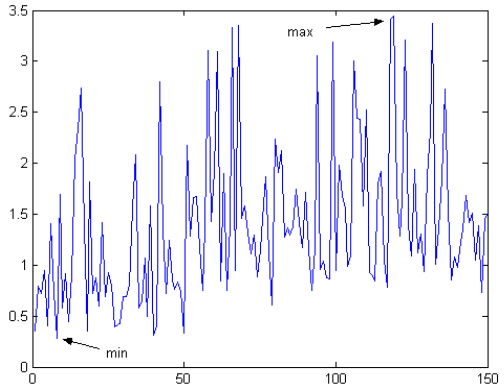

Fig. 6 S_k factor of the 150 data of iris dataset of FKCO

TABLE II
INTERVAL SCOPE COMPARISON OF DIFFERENT CRITERIA

Different Criteria	Interval Scope
S for FKCO	3.1661 (V_S)
Weight factor **w** in [5]	3.3654 (V_W)

Since **S** holds smaller change scope than that of **w**, so we can believe that FKCO algorithm and the newly defined criteria in (23) should be more reliable.

B. Test on Linear Inseparable Dataset

In the real word, the case of dataset with nonspherical boundaries occurs frequently. In such a case, the conventional sum-of-squares methods such as K-means will not work well. Recently, some new methods are proposed to solve this problem, such as maximum-certainty portioning. Another alternative method is to map the datum from original observation space into high-dimensional feature space with a nonlinear kernel function such as Guassian kernel function and then performing the clustering in the feature space. As stated above, we can perform linear clustering for such datasets in the mapped feature space.

Fig. 7 illustrates the linear inseparable dataset used in this example. As expected, FCM algorithm and the algorithm in [5] which work in the observation space, *fails* to perform the clustering properly on this dataset as illustrated in Figs. 8 and 9, whereas, FKCO algorithm *works well* as demonstrated in Fig. 10. It is worth to point out that FKCO can also identify the outliers more reliably with the additional constructed criteria of the datum.

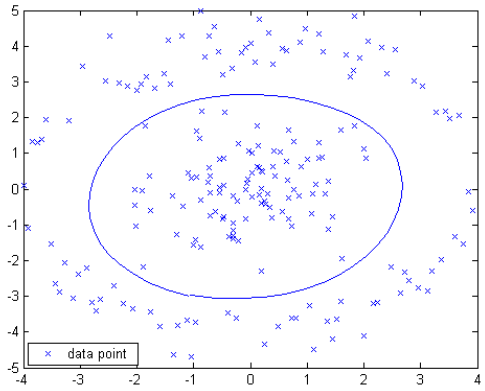

Fig. 7 Linear inseparable but nonlinear separable dataset.

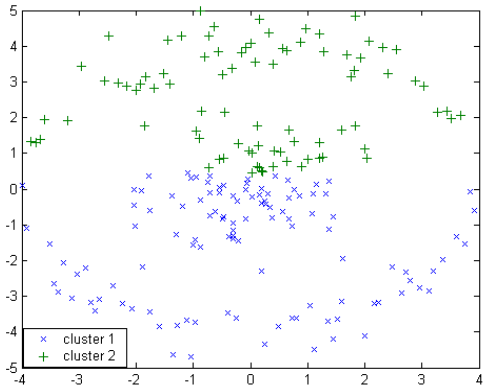

Fig. 8 Clustering performance of FCM on the third nonlinear separable sample.

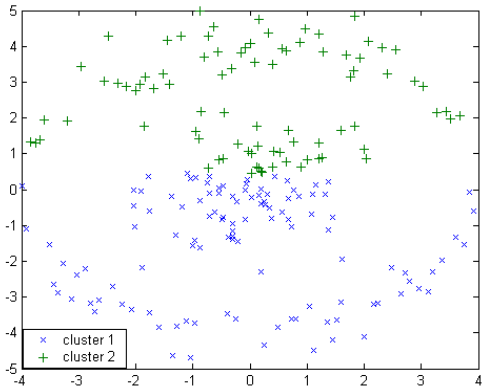

Fig. 9 Clustering performance of the algorithm of [5] on the third nonlinear separable sample.

C. Test on Image Processing

Finally, a test image is used here to illustrate the power of FKCO algorithm. We selected a satellite image of size 864×1024 and tried to find the outlier information in the image while obtain a good segmentation result. Fig. 11 is the satellite image of a city of China. Based on the histogram, we obtain the pixel number $h(k)$ in accordance with a particular gray value k $(0 \le k \le 255)$. Histogram represents a specialist character of an image and is regarded as the clustering dataset of FKCO in this paper.

While assuming the number of cluster $C = 2$, we will consider the gray value corresponding to the smaller cluster center as the segmentation threshold when FKCO finally converges, and the segmentation result is shown in Fig. 12.

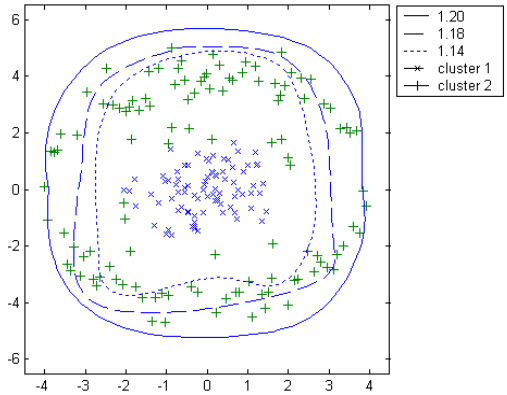

Fig. 10 Performance of the clustering and identifying the outliers with S of FKCO.

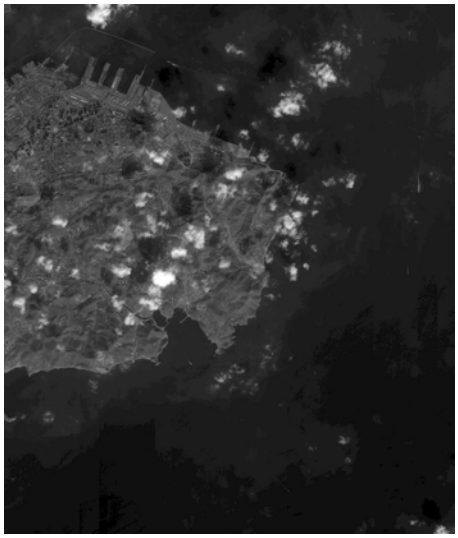

Fig. 11 Satellite image of a city in China.

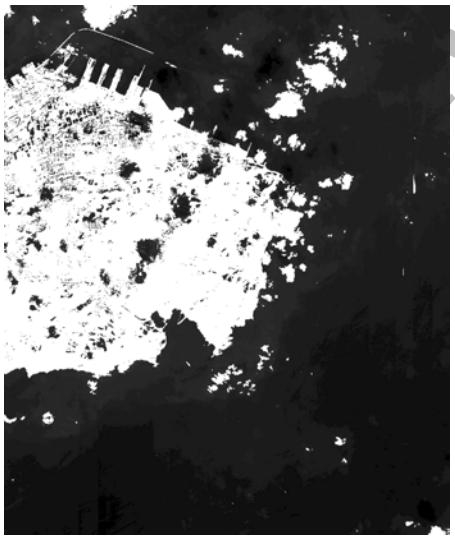

Fig. 12 Segmentation result of the satellite image.

Experiment results show that the criteria S_k of lower gray level is larger than that of higher gray level demonstrated in Fig. 13, in other words, FKCO algorithm has considered some lower gray level as the outlier information because there are much more pixels of such gray levels than that of lighter pixels which is shown in the original image. So we will perform a NOT operation while finding the outlier pixels in the image. That is to say when experts decide a threshold criteria T, we can find the

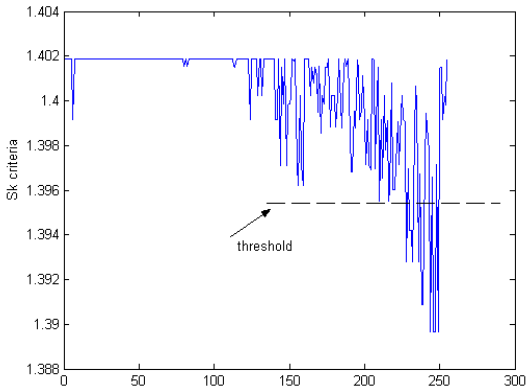

Fig. 13 criteria S_k for the histogram dataset of the image.

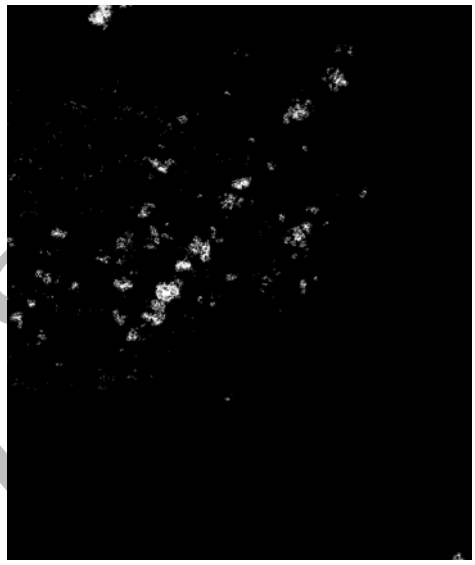

Fig. 14 Outlier information found by FKCO algorithm.

outlier pixels with $S_k < T$ and such an example is demonstrated by Fig. 14 and it also conforms to the reality well.

V. CONCLUSION

In this paper, we present a new algorithm FKCO to find the outlier information. An outstanding property of FKCO is that it cannot only obtain a satisfying clustering performance, but also can identify the outliers easily with the newly defined criteria. Different from the conventional algorithms for identifying outliers, FKCO works in the feature space instead of the original observation space, so it can work well for the linear inseparable dataset when the conventional methods fail. Our theoretic analysis shows that FKCO can converge to a local minimum of the objective function. Several experiments are also performed to show the validity and effectiveness of our FKCO. It is believed that FKCO can have applications in many research areas, e.g., economic analysis, data mining, etc.

APPENDIX

Now, let us prove the convergence properties of the new FKCO algorithm.

Lemma 1: In FKCO, $\mu_{ij}(i=1,2,...,C, j=1,2,...,K)$ and $w_j(j=1,2,...,K)$ is a local minimum for J_H only if

$$\mu_{ij} = 1 \Big/ \sum_{k=1}^{C} \left(\frac{Q_{ij}}{Q_{kj}}\right)^{\frac{1}{m-1}} \qquad \text{(A-1)}$$

subject to $\sum_{i=1}^{C}\mu_{ij}=1$ and w_j is a solution of (19) with the constraint $\sum_{j=1}^{K}w_j=w$, where $\mu_{ij}>0$ for all i,j.

Proof: First, we assume that w_j is fixed. Then the problem is to minimize J_H with respect to μ_{ij} under the constraint $\sum_{i=1}^{C}\mu_{ij}=1$. Using Lagrange multiplier method, we find that the problem is equivalent to minimizing

$$L(W,\lambda)=J_H-\sum_{j=1}^{K}\lambda_j\left(\sum_{i=1}^{C}\mu_{ij}-1\right) \tag{A-2}$$

without constraints. The necessary condition of this problem is

$$\frac{\partial L(W,\lambda)}{\partial\mu_{ij}}=m{\mu_{ij}}^{m-1}\frac{1}{{w_j}^{q}}Q_{ij}-\lambda_j=0 \tag{A-3}$$

$$\frac{\partial L(W,\lambda)}{\partial\lambda_j}=\sum_{i=1}^{C}\mu_{ij}-1=0 \tag{A-4}$$

from (A-3), we have

$$\mu_{ij}=\left[\frac{\lambda_j{w_j}^{q}}{mQ_{ij}}\right]^{\frac{1}{m-1}} \tag{A-5}$$

substituting (A-5) into (A-4), we have

$$\left(\frac{\lambda_j{w_j}^{q}}{m}\right)^{\frac{1}{m-1}}=\frac{1}{\left(\sum_{k=1}^{C}\frac{1}{Q_{kj}}\right)^{\frac{1}{m-1}}} \tag{A-6}$$

substituting (A-6) into (A-5), we obtain (A-1).

To show w_j is a solution of (19) with the constraint $\sum_{j=1}^{K}w_j=w$, the proof is similar to the process of deriving the iterative function of w_j.

Lemma 2: Let $\phi(U)=J_H$, where $U=[\mu_{ij}]_{C\times K}$, $w_j(j=1,2,...,K)$ is fixed, and $Q_{ij}\neq 0$, for all $1\le i\le C, 1\le j\le K$, then U is a local minimum of $\phi(U)$ if and only if w_j $(j=1,2,...,K)$ is computed via (19).

Proof: The only-if part has been proved where is in Theorem 3. To show the sufficiency, we examine $H(\phi)$, the $CK\times CK$ Hessian of the Lagrangian of $\phi(U)$ evaluated at the U given by (20) or (A-1). From (A-2), we have

$$h_{st,ij}(U)=\frac{\partial}{\partial\mu_{st}}\left[\frac{\partial\phi(U)}{\partial\mu_{ij}}\right]=\begin{cases}m(m-1){\mu_{ij}}^{m-2}\dfrac{1}{{w_j}^{q}}Q_{ij} & \text{if } s=i,t=k\\ 0 & \text{otherwise}\end{cases} \tag{A-7}$$

where μ_{st} is computed from (20). Thus, $H(U)=[h_{st,ij}(U)]$ is a diagonal matrix. Since $m>1$, and $Q_{ij}>0, w_j>0$ for all $1\le i\le C, 1\le j\le K$, we know from the above formula that Hessian $H(U)$ is positive definite and consequently, (20) is also a sufficient condition for minimizing $\phi(U)$.

Lemma 3: Let $\phi(W)=J_H$, where $U=[\mu_{ij}]_{C\times K}$ is fixed, and $Q_{ij}\neq 0$, for all $1\le i\le C$, $1\le j\le K$, $m>1$. Then $w_j(j=1,2,...,K)$ is a local minimum of $\phi(W)$ if and only if $w_j(j=1,2,...,K)$ is computed via (19).

Proof: The necessity was proved in Lemma 1. To show the sufficiency, we have from (12) that

$$\frac{\partial}{\partial w_i}\left[\frac{\partial\phi(w)}{\partial w_j}\right]=\begin{cases}q(q+1)\sum_{i=1}^{C}{\mu_{ij}}^{m}\dfrac{1}{{w_j}^{q+2}}Q_{ij}>0 & \text{if } i=j\\ 0 & \text{otherwise}\end{cases} \tag{A-8}$$

i.e., the Hessian is positive definite and consequently (19) is a sufficient condition for minimizing $\phi(w)$.

With Lemma 2 and Lemma 3, similar to the proof of [9], we can prove that

$$J_H(U^{t+1},W^{t+1})\le J_H(U^{t},W^{t}) \tag{A-9}$$

In other words, J_H is a decreasing function with the increase of t. So, the FKCO algorithm will finally converge.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for excellent comments on this paper.

REFERENCES

[1] J. C Farman and B. G. Gardinar, "Large losses of total Ozone in Antarctica," *Nature*, vol. 315, no. 16, pp. 207-210, May 1985.
[2] M. Last, A. Kandel, "Automated perceptions in data mining", in *Proc. of the Eighth Int. Conf. on Fuzzy System*, Seoul, Korea. Part I, pp. 190-197, Seoul, South Korea, 1999.
[3] J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, Reading, Massachusetts, USA, 1977.
[4] W. Mendenhall, J. E. Reinmuth, and R. J. Beaver, *Statistics for Management and Economics*, Duxbury Press, Belmont, CA, 1993.
[5] Annette Keller, "Fuzzy Clustering with Outliers," *in Proc. NAFIPS00*, Atlanta, pp. 143-147, US, 2000.
[6] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms,* New York, 1981.
[7] S. J. Roberts, R. Everson, and I. Rezek, "Maximum certainty data partitioning," *Pattern Recognition*, vol. 33, no. 5, pp. 833-839, May 2000.
[8] Z. Li, Z. W. Da, and J. L. Cheng, "Kernel clustering algorithm," *Chinese Journal of Computers*, vol. 25, no. 6, pp. 587-590, Jun. 2002.
[9] R. Hathaway and J. Bezdek, "Switching regression models and fuzzy clustering," *IEEE Trans. Fuzzy Syst.*, vol. 1, no. 3, pp. 195-204, Jun. 1993.
[10] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data,* Prentice-Hall, Upper Saddle River, NJ, 1988.
[11] R. Krishnapuran and James M. Keller "The possibilistic C-means algorithm: insight and recommendations," *IEEE Trans Fuzzy Syst.*, vol. 4, no. 3, pp. 385-393, Jun. 1996.
[12] O. Nasraoui, H. Frigui, A. Joshi, and R. Krishnapuram, "Extracting web user profiles using relational competitive fuzzy clustering," *Int. J. on Artificial Intelligence Tools*, vol. 9, no. 4, pp. 509-526, 2000.
[13] H. Frigui and O. Nasraoui, "Unsupervised learning of prototypes and attribute weights," *Pattern Recognition*, vol. 37, no. 3, pp. 567-581, Mar. 2004.
[14] J. M. Buhmann, "Stochastic algorithms for exploratory data analysis: data clustering and data visualization," *in Learning in Graphical Models*, ed., M. I. Jordan, pp. 405-420, Boston, MA: Kluwer, 1998.
[15] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.
[16] V. Roth and V. Steinhage, "Nonlinear discriminant analysis using kernel functions," *Advances in Neural Information processing Systems*, vol. 12, pp. 568-574, eds., S. A. Solla, T. K. Leen, and K. -R. Muller, Cambridge, MA: MIT Press, 1999.
[17] M. Girolami, "Mercer kernel-based clustering in feature space," *IEEE Trans. on Neural Networks*, vol. 13, no. 3, pp. 780-784, May 2002.
[18] B. Sch□lkopf, *et al*., "Input space versus feature space in kernel-based methods," *IEEE Trans. on Neural Networks*, vol. 10, no. 5, pp. 1000-1017, Sep. 1999.
[19] R. V. L. Hartley, "Transmission of Information," T*he Bell Systems Technical Journal*, vol. 7, pp. 535-563, Jul. 1928.

Hong-Bin Shen received the B.Sc. and M.Sc. degrees in computer science from Jiangsu University of Science and Technology in 2001 and 2004, respectively and now he is a Ph.D. candidate of institute of image processing and pattern recognition, Shanghai Jiaotong University, P. R. China. His research interests include image processing, pattern recognition, data mining, computational biology. His research results has been published in Biochemical and Biophysical Research Communications(BBRC), Journal of Theoretical Biology, Applied Soft Computing, Soft Computing Journal as well as other international journals and conferences.

Dr. Shen's master dissertation was selected as one of the best dissertations of Jiangsu Province.

Jie Yang was born in Shanghai, China. He received bachelor's degree in Automatic Control in Shanghai Jiaotong University, where master's degree in Pattern Recognition & Intelligent System was achieved three years later. In March of 1989, as one of the nation's first-class graduate students, he was enrolled by the Department of Computer, University of Hamburg, Germany and majored in image process, pattern recognition and intelligent system, with the guidance of Prof. Bernd Neumann, who are world-famous in Artificial Intelligence. He is the author of more than 100 scientific papers in computer vision, pattern recognition and artificial intelligence. Prof. Yang is a high-ranking member of IEEE and Institute of National Automation.

Yi-Fei Dong was born in Jiangsu, China. He is now a Ph.D. candidate with School of Computer Science and Engineering, The University of New South Wales, Australia. His research interests include pattern recognition, medical image processing.

Shi-Tong Wang was born in Jiangsu, China. He is now a Professor and doctoral supervisor in computer science of Department of Information, Southern Yangtse University. He is a visiting Professor of Polytechnic University of Hongkong and Advanced Computing Research Center, Bristol University, UK. Professor Wang has published more than 100 papers in fuzzy system, neural networks, pattern recognition and artificial intelligence

.