Clustering incomplete data using kernel-based fuzzy c-means algorithm

Dao-Qiang Zhang*, Song-Can Chen

Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, People's Republic of China

---

Abstract: There is a trend in recent machine learning community to construct a nonlinear version of a linear algorithm using the 'kernel method ', e.g. Support Vector Machines (SVMs), kernel principal component analysis, kernel fisher discriminant analysis and the recent kernel clustering algorithms. In unsupervised clustering algorithms using kernel method, typically, a nonlinear mapping is used first to map the data into a potentially much higher feature space, where clustering is then performed. A drawback of these kernel clustering algorithms is that the clustering prototypes lie in high dimensional feature space and hence lack clear and intuitive descriptions unless using additional projection approximation from the feature to the data space as done in the existing literatures. In this paper, a novel clustering algorithm using the 'kernel method' based on the classical fuzzy clustering algorithm (FCM) is proposed and called as kernel fuzzy c-means algorithm (KFCM). KFCM adopts a new kernel-induced metric in the data space to replace the original Euclidean norm metric in FCM and the clustered prototypes still lie in the data space so that the clustering results can be reformulated and interpreted in the original space. Our analysis shows that KFCM is robust to noise and outliers and also tolerates unequal sized clusters. And finally this property is utilized to cluster incomplete data. Experiments on two artificial and

* Corresponding author. Tel.:+86-25-489-2805.
E-mail address: daoqz@mail.com (D. Zhang), s.chen@nuaa.edu.cn (S. Chen).

one real datasets show that KFCM has better clustering performance and more robust than several modifications of FCM for incomplete data clustering.

Key words: Kernel methods, fuzzy c-means, clustering, incomplete data

1. Introduction

The fuzzy c-means (FCM) algorithm [1], as a typical clustering algorithm, has been utilized in a wide variety of engineering and scientific disciplines such as medicine imaging, bioinformatics, pattern recognition, and data mining. FCM partitions a given dataset, $X = \{x_1,...,x_n\} \subset R^p$, into c fuzzy subsets by minimizing the following objective function

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^m \| x_k - v_i \|^2 \qquad (1)$$

where c is the number of clusters and selected as a specified value in this paper, n the number of data points, u_{ik} the membership of x_k in class i, m the quantity controlling clustering fuzziness, and V the set of cluster centers ($v_i \in R^p$). The matrix U with the *ik*-th entry u_{ik} is constrained to contain elements in the range [0,1] such that $\sum_{i=1}^{c} u_{ik} = 1, \forall k = 1,2,...,n$. The function J_m is minimized by a famous alternate iterative algorithm.

Since the original FCM uses the squared-norm to measure similarity between prototypes and data points, it can only be effective in clustering 'spherical' clusters. To cluster more general dataset, a lot of algorithms have been proposed by replacing the squared-norm in Eq. (1) with other similarity measures [1, 2]. A recent development is to use kernel method to construct the kernel versions of the FCM algorithm [3-5]. A common ground of these algorithms is that clustering is performed in the transformed feature space after a (implicitly) nonlinear data transformation Φ. However, a drawback of these algorithms is that clustering result, especially those prototypes, is difficult to be exactly represented and reformulated due to having not

corresponding pre-images in the data space for some prototypes in the feature space. To avoid that problem, some additional approximate projection techniques must be used, as shown in [6,7].

On the other hand, all aforementioned algorithms are based on the assumption that the data in a dataset are complete, that is, all of the features (i.e., components) of every vector in X are known or exist. However, many real data sets such as medical data suffer from incompleteness, i.e., one or more of the components in X are missing, as a result of measurement errors, missing observations, etc [8]. Many algorithms have been proposed to deal with incomplete data [8-12]. An elementary but good summary was given in [9], where several principles for handling incomplete data were included. More recently, the triangle inequality was used to estimate the missing dissimilarity data [8]. And in [12], several ways were developed to continue the FCM clustering of incomplete data. One simple method is to use the partial distance strategy (PDS) in FCM, the other is to estimate the missing feature as the weighted sum of the prototypes (WSP), and another strategy is the nearest prototype strategy (NPS).

In this paper, an alternative kernel-based fuzzy c-means (KFCM) algorithm is proposed to cluster incomplete data. Unlike the usual way utilizing kernel method in FCM, the proposed KFCM clustering algorithm is performed still in original data space, i.e., prototypes lie in data space. Furthermore, KFCM adopts a more robust kernel-induced metric different from the Euclidean norm in original FCM. By a similar way as in [12], we applied the proposed KFCM to cluster incomplete data, and it is shown that WSP and NPS are two special cases of KFCM when clustering incomplete data. Furthermore, because KFCM has better outlier and noise immunity than FCM, it is especially suitable to dealing with incomplete data. In this paper, three artificial and real datasets are used for testing.. Experimental results show that KFCM has better

performance than WSP and NPS.

In Section 2, we first discuss the alternative kernel based fuzzy c-means clustering algorithm, and then we apply this algorithm for incomplete data clustering in Section 3. To demonstrate the effectiveness of the proposed algorithm, three experiments on incomplete datasets are conducted and results are given in Section 4. At last, conclusions and discussions are given in Section 5.

2. Kernel fuzzy c-means clustering (KFCM)

Define a nonlinear map as $\Phi : x \to \Phi(x) \in F$, where $x \in X$. X denotes the data space, and F the transformed feature space with higher or even infinite dimension. KFCM minimizes the following objective function

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^m \| \Phi(x_k) - \Phi(v_i) \|^2 \qquad (2)$$

where

$$\| \Phi(x_k) - \Phi(v_i) \|^2 = K(x_k, x_k) + K(v_i, v_i) - 2K(x_k, v_i) \qquad (3)$$

where $K(x,y) = \Phi(x)^T \Phi(y)$ and is an inner product kernel function. If we adopt the Gaussian function as a kernel function, i.e., $K(x,y) = \exp(-\| x - y \|^2 / \sigma^2)$, then $K(x,x) = 1$, according to Eq. (3), Eq. (2) can be rewritten as

$$J_m(U,V) = 2\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^m (1 - K(x_k, v_i)) \qquad (4)$$

Minimizing Eq. (4) under the constraint of U, we have

$$u_{ik} = \frac{(1/(1 - K(x_k, v_i)))^{1/(m-1)}}{\sum_{j=1}^{c} (1/(1 - K(x_k, v_j)))^{1/(m-1)}}, \forall i = 1,2,...,c, k = 1,2,...,n \qquad (5)$$

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m K(x_k, v_i) x_k}{\sum_{k=1}^{n} u_{ik}^m K(x_k, v_i)}, \forall i = 1,2,...,c \qquad (6)$$

It is worth to note that although Eqs. (5)~(6) are derived using the Gaussian kernel function, we can use other functions satisfying $K(x,x) = 1$ in Eqs. (5)~ (6) in real applications such as the

following RBF functions and hyper tangent functions:

(1) RBF functions:

$$K(x,y) = \exp(-\sum_i | x_i^a - y_i^a |^b / \sigma^2) \quad (0 < b \le 2) \tag{7}$$

(2) Hyper tangent functions:

$$K(x,y) = 1 - \tanh(- \| x - y \|^2 / \sigma^2) \tag{8}$$

Nothe that RBF function with $a = 1, b = 2$ reduces into the common-used Gaussian function. In fact, Eq.(3) can be viewed as kernel-induced new metric are in the data space, which is defined as the following

$$d(x,y) \overset{\Delta}{=} \| \Phi(x) - \Phi(y) \| = \sqrt{2(1 - K(x,y))} \tag{9}$$

In Appendix I, we prove that $d(x,y)$ defined in Eq. (9) is a metric in the original space in case that *K(x,y)* takes as the Gaussian, RBF functions and hyper tangent function as the kernel functions respectively. According to Eq. (6), the data point x_k is endowed with an additional weight $K(x_k, v_i)$, which measures the similarity between x_k and v_i. When x_k is an outlier, i.e., x_k is far from the other data points, $K(x_k, v_i)$ will be very small, so the weighted sum of data points shall be more robust. Since in incomplete dataset, a data point with missing components is likely to turn into an outlier, the algorithm based on KFCM to cluster incomplete data is of great potential.

3. Clustering incomplete data using KFCM

To implement clustering on incomplete data, we derive the following algorithm based on KFCM:

(1) Fix c, $m > 1$ and $\varepsilon > 0$ for some positive constant;

(2) Set $x_{kj} = 0$, if x_{kj} is a missing feature;

(3) Initilize the prototypes v_i using FCM;

(4) For t=1,2,…, $t_{\max}$, do:

(a) Update all memberships u_{ik}^t with Eq. (5);

(b) Update all prototypes v_i^t with Eq. (6);

(c) Calculate the missing value using Eq. (10);

(d) Compute $E^t = \max_{i,k} | u_{ik}^t - u_{ik}^{t-1} |$, if $E^t \le \varepsilon$, stop;

Endo.

$$x_{kj} = \frac{\sum_{i=1}^{c} u_{ik}^m K(x_k, v_i) v_{ij}}{\sum_{i=1}^{c} u_{ik}^m K(x_k, v_i)} \qquad (10)$$

Here the kernel function is the same as in Section 2. From Eqs. (5), (6) and (10), as $\sigma \to \infty$, $K(x_k, v_i) \sim 1 - \| x_k - v_i \|^2 / \sigma^2$, then KFCM reduces to the classical FCM, and Eq. (10) changes into the expression used in WSP algorithm. Furthermore, if $\sigma \to 0$, Eq. (10) reduces to $x_{kj} = v_{pj}$, where $p = \min_i(\| x_k - v_i \|^2)$, which is just the strategy used in NPS algorithm.

4. Experiments Results and Discussions

In this section, we compare the performance of FCM with that of KFCM on some incomplete datasets. In all cases, the incomplete data are artificially generated the by randomly selecting a specified percentage of its components to be designated as missing values. The random selection of missing values is constrained so that: 1) each original feature vector retains at least one component; 2) each feature has at least one value present in the incomplete data set [12]. The initial values for prototypes are obtained using FCM on the original (complete) data sets. And the kernel functions used are Gaussian, RBF, and hyper tangent functions.

The first dataset (Data Set A) is an artificially generated one [2]. It contains two clusters; one

has 25 points and the other 125 points. Fig. 1 shows the clustering result on the complete data using FCM and KFCM. We use both Gaussian and hyper tangent kernel function in this experiment, with $\sigma = 2$, and experiments are repeated 10 times, with each time having the same result. It can be seen from Fig. 2 that FCM misclassifies 8 data points, while KFCM correctly classifies the data. Table I gives the clustering result on incomplete Data Set A. Both Gaussian and RBF kernels are used in KFCM, with $\sigma = 0.2$. Results are got under a total of 1000 trials. Obviously, KFCM has much better performance than FCM on Data Set A and the best result is got using the RBF function at a=1.5,b=1.2, with the shading in Table I.

The second artificial dataset (Data Set B) consists of two clusters having 100 points each in R^5 [12]. They are randomly distributed from a Gaussian distributions with mean (-1 -1 -1 -1 -1) and (1 1 1 1 1) respectively and have identity covariance. Fig. 2 shows the 2-D plot of Data Set B using the standard PCA technique [13]. For Data Set B with Gaussian distributions, both FCM and KFCM can correctly classify the data. Next, we perform clustering on incomplete Data Set B. Table II gives the clustering result on incomplete Data Set B. The kernel function used in KFCM is the Gaussian and hyper tangent function, both taking $\sigma = 2$. Results are averaged under a total of 1000 trials. The Gaussian function has the smallest misclassifications in average in all cases, which are highlighted by shading in Table II.

The last dataset is the well-known Iris dataset. It consists of 150 four-dimensional feature vectors, with 50 vectors for each of three physically labeled classes. In order to acquire better clustering performance, each vector is normalized. The clustering result on the incomplete Iris data is shown in Table III. The kernel functions used are Gaussian and RBF functions, both with $\sigma = 1$. The result is averaged under a total of 1000 trials. The RBF function under a=0.5, b=2

has the averaged smallest misclassifications in all cases, which are highlighted by shading in Table III.

From our experiments, we found that different kernels with different parameters lead to different clustering results. Thus a key point is to choose an appropriate kernel parameter. However, there is not a general theory to guide the selection of the best parameter in most kernel based algorithms. This is an open problem. Here in our experiments, we adopted an approach similar to that in [6], *i.e.*, running a 5-fold cross-validation procedure only on a few realizations of the data set. Each time, this is done in two stages: first taking a large interval in the exponential scale to find a good initial guess of the parameter, and then shortening gradually the interval to refine the parameter at the second stage. We use the median of the five estimations throughout the remaining trials on the data set. Usually, it needs about several tens of trials on the data set to get an appropriate parameter. Remembering in our experiments, a total of 1000 trials are done, the computation cost on choosing an appropriate parameter is still less.

5. Conclusions

In this paper, we proposed a kernel-induced new metric to replace the Euclidean norm in fuzzy c-means algorithm in the original space and then derived the alternative kernel-based fuzzy c-means algorithm. Unlike the common way using the 'kernel method' to represent a variable in dual form as in SVM [7], kernel PCA [6], kernel Fisher discriminant analysis [6] and kernel clustering algorithms [3-5], we adopted a new kernel-induced metric as in Eq.(2). It was shown the proposed kernel clustering algorithm is robust to noise and outliers and also tolerates unequal sized clusters. That property is further utilized to cluster incomplete data and results in better

performance than those classical counterparts.

Appendix I: Proof that $d(x,y)$ defined in Eq. (9) is a metric

Proof. To prove $d(x,y)$ is a metric, the necessary and sufficient condition is that $d(x,y)$ satisfies the following three conditions [14]

(i) $d(x,y) > 0, \forall x \neq y, d(x,x) = 0$,

(ii) $d(x,y) = d(y,x)$,

(iii) $d(x,y) \leq d(x,z) + d(z,y), \forall z$.

It's easy to verify that for Gaussian, RBF and hyper tangent kernel functions, $d(x,y)$ defined in Eq. (9) satisfies $\forall x \neq y$, $d(x,y) = d(y,x) > 0$, and $d(x,x) = 0$, so condition (i) and (ii) are satisfied. From Eq. (9), we have

$$d(x,y) = \| \Phi(x) - \Phi(y) \| \leq \| \Phi(x) - \Phi(z) \| + \| \Phi(z) - \Phi(y) \| = d(x,z) + d(z,y).$$

Thus condition (iii) is also satisfied due to the properties of the norm. So $d(x,y)$ is a metric.

Acknowledgements

The authors are grateful to the anonymous reviewers for their comments and suggestions to improve the presentation of this paper. This work was supported in part by the National Science Foundations of China and of Jiangsu under Grant Nos. 60271017 and BK2002092, "QingLan project" foundation of Jiangsu and Returnees foundation of Educational Ministry.

References

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

2. K.L. Wu, M.S. Yang, Alternative c-means clustering algorithms, Pattern Recognition vol.

35, pp. 2267-2278, 2002

3. M. Girolami . Mercer kernel-based clustering in feature space. IEEE Trans. Neual Networks, ,vol. 13(3), pp. 780-784, 2002
4. L. Zhang, W.D. Zhou, L.C. Jiao. Kernel clustering algorithm. Chinese J. Computers, vol. 25(6), pp. 587-590, 2002 (in Chinese)
5. D.Q Zhang, S.C. Chen. Fuzzy clustering using kernel methods. in Procedings of Inter. Conf. Control and Automatation (ICCA'02), pp. 123-128, Xiamen, China, June 16-19, 2002
6. K.R. Muller, S. Mika, etal. An Introduction to Kernel-based Learning algorithms. IEEE Trans. Neural Networks, vol. 12(2), pp. 181-202, 2001
7. V.N. Vapnik. Statistical learning theory. Wiley, New York, 1998
8. R.J. Hathaway, J.C. Bezdek. Clustering incomplete relational data using the non-Euclidean relational fuzzy c-means algorithm. Pattern Recognition Letters, vol. 23, pp.151-160, 2002
9. A.K. Jain, R.C. Dubes. Algorithms for Clustering Data. Englewood Cliffs, NJ, 1988
10. W. Gaul, M. Schader. Pyramidal classification based on incomplete data. J. Classification, vol. 11, pp.171-193, 1994,
11. J.L. Schafer. Analysis of Incomplete Multivariate Data. Chapman &Hall, London, 1997
12. R.J. Hathaway, J.C. Bezdek. Fuzzy c-means clustering of incomplete data. IEEE Trans. Syst. Man. Cybernetics. vol. 31(5), pp.735-744, 2001
13. L.T. Jolliffe. Principal Component Analysis. Springer-Verlag, 1986
14. W. Rudin, Principles of Mathematical Analysis, McGraw-Hill Book Company, New York, 1976

Figure Captions:

Fig.1 Clustering result of FCM and KFCM algorithms on Data Set A.

+ cluster 1

• cluster 2

○ data points misclassified by FCM

Fig.2 Two dimension plot of the distribution of Data Set B.

+ cluster 1

○ cluster 2

Table Captions:

Table I. Averaged number of misclassifications on incomplete Data Set A

Table II. Averaged number of misclassifications on incomplete Data Set B

Table III. Averaged number of misclassifications on incomplete Iris dataset

Fig.1

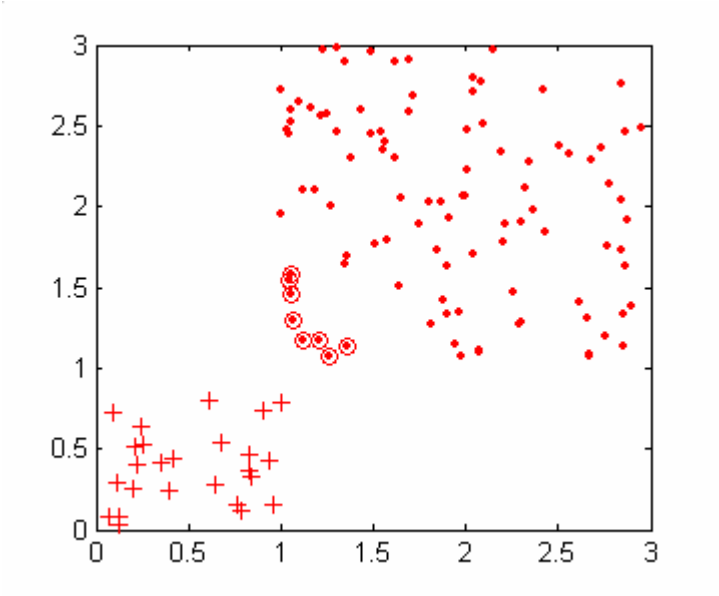

Fig.2

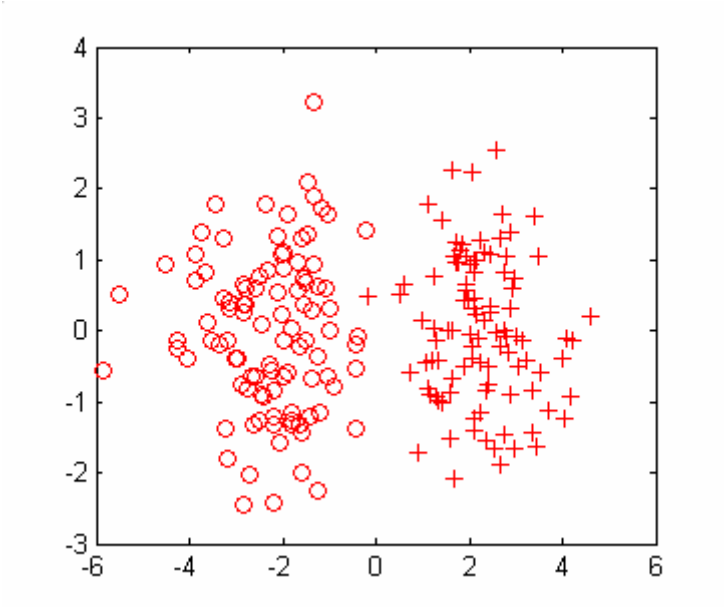

Table I

% Missing	PDS	WSP	NPS	Gaussian	RBF(a=1.5,b=1.2)
10	21.84	11.92	17.95	7.62	7.27
20	33.56	16.12	31.45	16.08	14.93

Table II

% Missing	PDS	WSP	NPS	Gaussian	Hyper tangent
20	2.57	2.54	2.61	2.43	2.51
40	6.39	6.33	6.71	6.07	6.10
60	15.70	14.66	30.50	14.32	14.39

Table III

% Missing	PDS	WSP	NPS	Gaussian	RBF(a=0.5,b=2)
25	63.96	16.33	29.14	13.57	12.73
50	77.79	37.21	50.75	37.66	31.26