Kernel Parameter Optimization in Stretched Kernel-Based Fuzzy Clustering

Chunhong Lu(✉), Zhaomin Zhu, and Xiaofeng Gu

Key Laboratory of Advanced Process Control for Light Industry
(Ministry of Education), Department of Electronics Engineering,
Jiangnan University, Wuxi 214122, China
sharon0510@126.com, zhuzhaomin@gmail.com,
xgu@jiangnan.edu.cn

Abstract. Although the kernel-based fuzzy c-means (KFCM) algorithm utilizing a kernel-based distance measure between patterns and cluster prototypes outperforms the standard fuzzy c-means clustering for some complex distributed data, it is quite sensitive to selected kernel parameters. In this paper, we propose the stretched kernel-based fuzzy clustering method with optimized kernel parameter. The kernel parameters are updated in accordance with the gradient method to further optimize the objective function during each iteration process. To solve the local minima problem of the objective function, a function stretching technique is applied to detect the global minimum. Experiments on both synthetic and real-world datasets show that the stretched KFCM algorithm with optimized kernel parameters has better performance than other algorithms.

Keywords: Kernel fuzzy c-means · Kernel parameter · Optimization · Stretching technique

1 Introduction

Clustering algorithms are universally employed to partition patterns into a couple of smaller homogeneous groups. The fuzzy c-means (FCM) [1] algorithm, a typical one has been widely used in pattern recognition and image segmentation. The FCM algorithm, which applies Euclidean distance measure between objects and prototypes can obtain good clustering results for spherically-structured data, but cannot obtain effective clustering analysis for some complex distributed data such as the mixture structure with heterogeneous cluster prototypes and non-spherical geometry of data. The kernel-based fuzzy c-means (KFCM) [2] algorithm was then presented to overcome this drawback. A kernel function is defined to transform nonlinear distributed data to the higher dimensional feature space so that the naturally distributed data can be partitioned linearly. Obviously, the KFCM algorithm can improve the results of FCM algorithm by selecting appropriate kernel function and reasonable parameters. Due to its superiority to other kernel functions, the RBF kernel function has been employed in all four kernel clustering algorithms [3], which is thus also chosen in this work.

Z.-H. Zhou and F. Schwenker (Eds.): PSL 2013, LNAI 8183, pp. 49–57, 2013.
DOI: 10.1007/978-3-642-40705-5_5,

However, the values of kernel parameters affect the performance of KFCM algorithm significantly. With respect to optimization of the kernel parameter, there are several methods in kernel-based techniques. Firstly, empirical, grid search and cross-validation methods are often applied to search the optimal kernel parameters [4]. On the one hand, users repeatedly execute their concerned algorithms for a couple of candidate kernel parameters according on their research experience until one corresponding to the best results is chosen as the final kernel parameter value. Unfortunately, because the number of these concerned candidate values is usually limited, the results of the kernel-based approaches are not distinctly preferable. On the other hand, the well-known cross-validation approach widely applied in the model selection can effectively obtain more favorable performance than selecting the parameter values empirically since the optimal value is chosen within a fairly broader range. But the kernel parameters need to be adjusted in real-time, which the cross-validation approach can not achieve on account of its time-consuming implementation. Secondly, the optimal kernel parameters can also be obtained by minimizing an objective function [5]. However, non-optimized parameters or sub-optimized values can be found by the above-mentioned methods. Although these kernel parameter learning methods can gain relative satisfactory evaluation results, searching the global optimized parameter is not promising. AL-Sultan and Fedjki [6] proposed a tabu search algorithm for globally solving fuzzy c-means clustering and obtained better performance in most of the test cases than the standard FCM algorithm.

In this paper, we propose an optimization method to select the kernel parameters for the KFCM algorithm, and employ a function stretching technique [7] to reformulate the objective function once it is trapped into a local minimum and to select the global optimized kernel parameter as the solution to the objective function. Incorporating the KFCM algorithm added by the stretching technique with optimization the kernel parameter is presented. Results of experiments on synthetic data sets and real data sets demonstrate that it is most important to select the desirable kernel parameters in kernel-based fuzzy clustering so that the KFCM algorithm with the best optimized parameters can outperform the other algorithms.

2 KFCM Algorithm

The classical FCM algorithm is remarkably effective only in partitioning spherical data based on the sum-of-squares error criterion. The KFCM algorithm has been adopted to overcome the problem in partitioning complex data with nonlinear boundaries by mapping nonlinear structure in the input space into a higher dimensional feature space. Given a data set $X = \{x_1,\ldots, x_N\}$, x_i $(i = 1\ldots N)$ from the input space R^p, a transformation function ϕ nonlinearly maps the data points to a higher dimensional feature space R^q. That is, $\phi : R^p \rightarrow R^q$, $p < q$. The inner product of two patterns obtained by the mapping function can be simply defined as a kernel function by using 'kernel trick' method:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j). \tag{1}$$

There are several examples of a kernel function. Using the RBF kernel as a kernel function in this paper, σ^2 as the variance parameter, we have

$$K(x_i, x_j) = e^{-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}}. \tag{2}$$

The kernel-induced Euclidean distance between patterns x_i and x_j in the input space can be calculated in the higher dimensional feature space through the kernel function $K(x_i, x_j)$as

$$d_{ij}^2\big(\phi(x_i), \phi(x_j)\big) = \left\|\phi(x_i) - \phi(x_j)\right\|^2 \tag{3}$$

$$\begin{aligned} &= \phi(x_i)\cdot\phi(x_i) - 2\phi(x_i)\cdot\phi(x_j) + \phi(x_j)\cdot\phi(x_j) \\ &= K(x_i, x_i) - 2K(x_i, x_j) + K(x_j, x_j). \end{aligned} \tag{4}$$

For a RBF kernel function, the distance measure in (3) using (2) can be easily described by

$$d^2(x_i, v_j) = 2\big(1 - 2K(x_i, v_j)\big). \tag{5}$$

The main goal of the KFCM algorithm is to minimize the following objective function as in the FCM.

$$J = \sum_{j=1}^{C}\sum_{i=1}^{N} u_{ij}^m d^2(x_i, v_j), \quad 2 \le C \le N, \tag{6}$$

where m represents the fuzzifier constant; C is the cluster number; N is the number of patterns; u_{ij} ∈[0,1], as elements of an order matrix $U_{C\times N}$, represents the membership degree of x_i in cluster j; V = (v_1,…,v_j,….v_c), v_j as centroids or prototypes. In addition, the elements of partition matrix U satisfy the following condition:

$$\sum_{j=1}^{C} u_{ij} = 1 \quad \text{and} \quad 0 < \sum_{i=1}^{N} u_{ij} < N, \forall j. \tag{7}$$

The memberships to minimize the objective function in (5) can be calculated by

$$u_{ij} = \frac{\left(1/d^2(x_i, v_j)\right)^{1/(m-1)}}{\sum\limits_{j=1}^{C}\left(1/d^2(x_i, v_j)\right)^{1/(m-1)}}, \tag{8}$$

and the formula of the prototypes v_j (j=1,2,…,C) proceeds as follows:

$$v_j = \frac{\sum\limits_{k=1}^{N} u_{jk}^m(x_k, v_j)x_k}{\sum\limits_{k=1}^{N} u_{jk}^m(x_k, v_j)}. \tag{9}$$

3 Learning the Kernel Parameters

The minimization problem of the objective function J given in (5) subjected to the constraint specified by Eq. (6) is solved by minimizing a constraint objective function defined as

$$\tilde{J}(\sigma) = \sum_{j=1}^{C}\sum_{i=1}^{N} u_{ij}^{m} d^{2}(x_i, v_j) + \sum_{j=1}^{C} \lambda_j (\sum_{i=1}^{N} u_{ij} - 1), \tag{10}$$

where λ_j ($j = 1,\ldots,C$) are Lagrangian multipliers. The optimal values of kernel parameter is obtained by minimizing (9),

$$\sigma^{*} = \arg\min_{\sigma} \tilde{J}(\sigma). \tag{11}$$

To calculate the kernel parameter σ, the general gradient method is applied to generate the optimal values σ^{*} by continually updating the following Eq. (11)

$$\frac{\partial \tilde{J}}{\partial \sigma} = -2\sum_{j=1}^{C}\sum_{i=1}^{N} u_{ij}^{m} K(x_i, v_j) \frac{\left\| x_i - v_j \right\|^2}{\sigma^3}, \tag{12}$$

$$\sigma^{(n+1)} = \sigma^{(n)} + \theta\left(\frac{\partial \tilde{J}}{\partial \sigma}\right). \tag{13}$$

Herein, θ is called the learning rate, commonly a positive constant and n is the iteration step. When $\left|\sigma^{(n+1)} - \sigma^{(n)}\right| < \varepsilon$, and ε is a very small positive number, the searching procedure reaches the convergence.

The objective function J is calculated according to Eq. (5) by serially updating the value of σ during executing kernel-based fuzzy clustering process until termination criteria satisfied. That is, the two consecutive Js almost remain unchanged. The iterative process can learn a minimizer of J. However, this acceptable minimizer may be local optimum in many cases. The stretching technique is applied for alleviating the local minima problem so that more optimized values of σ can be found. The idea of stretching technique is to perform a two-stage transformation of the objective function. In the first stage, local minima with higher functional values than the stretched local minimizer are eliminated, while other lower local minima remain unchanged; in second transformation stage, the neighbors of the local minimizer are stretched upwards and the local minimizer is turned into a local maximum. Thus the location of the global minimum is left unaltered in the final. This means a minimum of the objective function J can be obtained at the unchanged minimizer. Supposing an optimization objective function corresponds to a local minimize $\bar{\sigma}$ and the transformation function is defined as

$$G(\sigma) = J(\sigma) + \gamma_1 \|\sigma - \bar{\sigma}\| (sign(J(\sigma) - J(\bar{\sigma})) + 1), \tag{14}$$

$$J_{new}(\sigma) = G(\sigma) + \gamma_2 \frac{sign(J(\sigma) - J(\bar{\sigma})) + 1}{\tanh(\tau(G(\sigma) - G(\bar{\sigma})))}, \tag{15}$$

where γ_1, γ_2 and τ are arbitrary chosen positive constants, and $sign(\cdot)$ is the well-known triple valued sign function:

$$sign(x) = \begin{cases} 1, & if \quad x \quad > \quad 0, \\ 0, & if \quad x \quad = \quad 0, \\ -1, & if \quad x \quad < \quad 0. \end{cases}$$

Herein, $J(\sigma)$ is replaced by $J_{new}(\sigma)$ and a newly formulated $J(\sigma)$ is regarded as the objective function. The updating process repeating and stretching function transforming many times, the optimized solution to the original objective function can be obtained.

4 Stretched KFCM Algorithm with Optimal Kernel Parameter

Kernel-based fuzzy clustering algorithm is derived from the above section by the stretching technique and optimization of kernel parameter as follows:

Stretched kernel-based fuzzy c-means algorithm with optimal kernel parameter (SKFCM-opt *σ)*

Step 1. Set learning rate θ, the maximum iteration number n, stopping criterion ε, initial iteration k=0;γ_1>0, γ_2>0 and τ to a very small positive number.

Step 2. Initialize fuzzifier $m > 1$, usually set to 2, fuzzy partition U, prototypes v_j, kernel parameter σ_0 using Eq. (15).

Step 3. Update the memberships, cluster center based on formula (7) and (8); update kernel parameter σ according to Eq. (12) and $\sigma > 0$ must be satisfied during the iteration process. It is assumed that the kernel width exceeds zero at each iterating. If it is close to zero or a negative number, just giving 2ε to it in order to avoid the risk of degeneracy.

Step 4. Calculate the value difference of the objective function between consecutive iterations. Once a local minimizer is found, update the objective function J using newly-obtained J according to Eqs. (13) and (14).

Step 5. Repeat the total iteration process until termination criteria satisfied or maximum iterations reached.

Step 6. Select one minimizer that yields the best optimal solution of formula (5) and the minimizer is regarded as the best optimal value.

In this section, the computational complexity of the proposed algorithm is O (C N q l), where l is the iteration time in the algorithm implementing, C is clustering number, and N is the number of samples and q is feature attribute. The kernel matrix is calculated between the patterns and prototypes during each iteration. From the viewpoint of storage complexity, the space O (NC+Nq+$2Cq$) is used to storage samples, cluster prototypes and partition matrix. Additionally, the algorithm is proper for clustering some large datasets.

5 Experiments

The experimental results of SKFCM-opt are evaluated in this section to demonstrate effectiveness of the proposed method, compared with the other methods, KFCM with non-optimized σ (KFCM-σ_0), FCM with the tabu search technique (FCM-tabu) detailed in [6] and standard FCM without mapping respectively. An artificial data set Circles and two real datasets (Iris and Pendigits from UCI Machine Learning Repository [8]) are applied in these tests. We choose m=2 which is a common choice for fuzzy clustering. All obtained experimental results use the following parameters: average on 10 runs, iteration error ε=0.00001, max_iter=100, γ_1=3000, γ_2=0.5,$\tau = 10^{-5}$, θ=0.05.The initial value for the kernel parameter σ is set to

$$\sigma_0 = \frac{1}{C}\sqrt{\frac{1}{N}(\sum_{i=1}^{N}\|x_i - \bar{x}\|^2)}, \tag{16}$$

$\bar{x}$ is the centroid of the total patterns.

The Circles data set involves 200 data points in a 2-dimensional space and two classes which respectively contains 100 samples of rectangle (4 × 4) distribution and circular(radius is 4) distribution. The Circles data set is shown in Fig. 1.

The Iris data set contains 50 4-dimensional samples each of three species (Versicolor, Virginica, Setosa). One of the classes is well separated from the other two, while the remaining two are partly overlapped.

The Pendigits data set comprises of 16-dimensional 7494 samples each of ten classes. We randomly select 100 data for each of the four classes {1,3,5,7}, total 400 data used in the experiment.

The Defense Advanced Research Projects Agency (DARPA) 1998 Basic Security Module (BSM) datasets are increased for experimental results. Among these datasets,

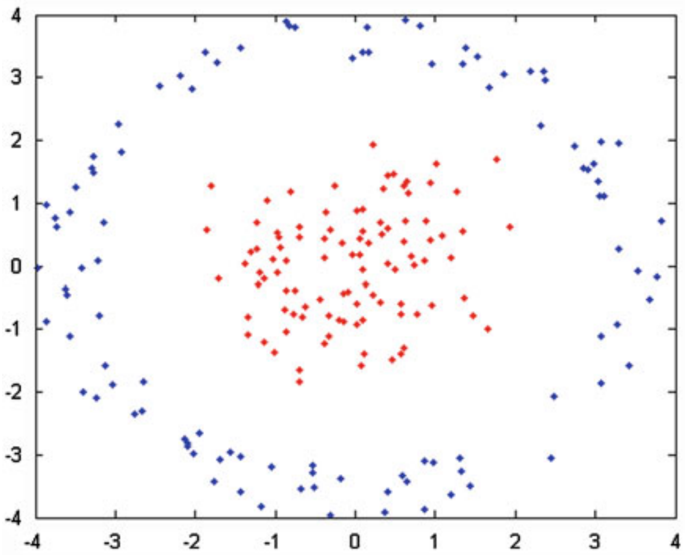

Fig. 1. The Circles data set

host-based BSM audit data from the seven-week training data and two weeks of testing data to evaluate the performance of the two algorithms, which involves two classes with 246 features, 7632 normal sessions and 456 attack sessions, respectively [9].

These data sets are used to analyze the quality of clustering. Table 1 shows the clustering accuracies of the presented method by comparing it with the other algorithms. The Circles data is turned into linear separation in the mapped feature space by the kernel function so that its clustering accuracy is evidently improved. In general, the performance of FCM-tabu method is better than the FCM algorithm, but is slightly worse than KFCM-σ_0 in the most cases. Additionally, SKFCM with the best kernel parameter (SKFCM-opt σ) obtains the best clustering accuracy among all these methods. It is apparently that stretching technique and optimization have an important influence on the clustering performance.

The clustering accuracies of SKFCM under a series of kernel parameter σ on Iris are shown in Fig. 2. It verifies the fact that the optimization of kernel parameters considerably affects the results of SKCM. Varying values of objective function under

Table 1. The clustering accuracies using FCM, FCM-tabu, KFCM and the proposed SKFCM with optimized σ for the aforementioned data sets.

Data sets	FCM (%)	FCM-tabu (%)	KFCM-σ_0 (%)	SKFCM-opt σ (%)
Circles data	51.16	78.24	93.24	100
Iris data	89.31	90.08	92.38	97.56
Pendigits data	42.82	48.34	59.74	67.23
DARPA data	40.52	54.05	53.61	65.29

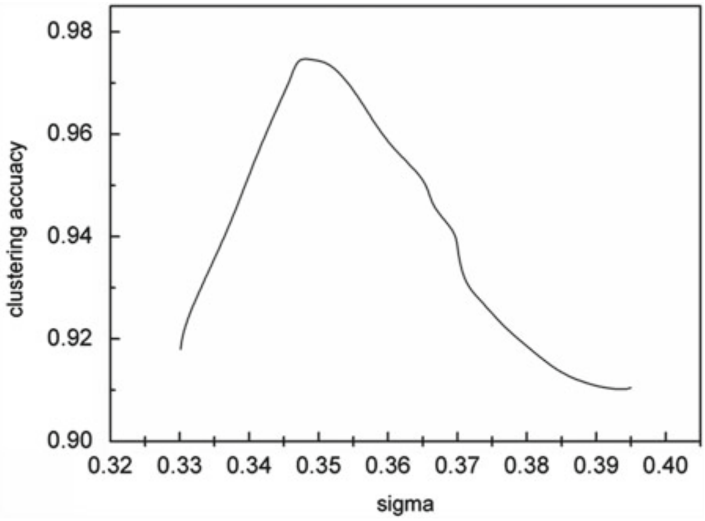

Fig. 2. Clustering accuracies under different kernel parameters σ on the Iris data

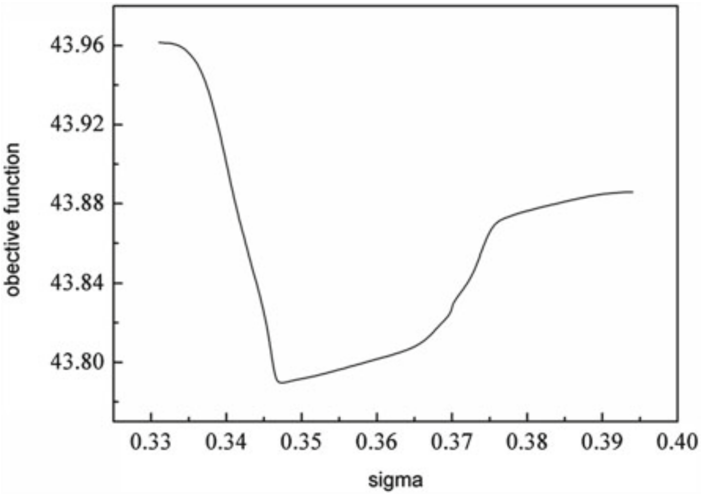

Fig. 3. Values of objective function under a series of σ values on the Iris data

a series of σ values on the Iris data is described in Fig. 3. From these two figures, we can observe when the proposed method achieves its best performance the objective function almost reaches its global minimum, given the same kernel parameter σ.

6 Conclusions

We have proposed a novel method for kernel-based fuzzy clustering. Based on the gradient method, the optimal parameter solution to the objective function of the KFCM algorithm is obtained, and a stretched technique reformulating the objective function can assure its optimal solution unaltered. Experimental results show that the proposed stretched kernel fuzzy clustering method with the optimal kernel parameter can be successfully applied compared with the KFCM algorithm with non-optimized σ, FCM with the tabu search algorithm and the standard FCM algorithm. However, there is no standard learning mechanism for evaluating the selection of the kernel parameters, a significant drawback of the kernel-based fuzzy clustering algorithms, which is worth of studying in the future work.

Acknowledgements. This work is supported by the PAPD of Jiangsu Higher Education Institutions, the Fundamental Research Funds for the Central Universities (JUDCF12027 JUSRP211A37) and the Fund of the State Key Laboratory of ASIC and System in Fudan University (11KF003).

References

1. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York (1981)
2. Wu, Z., Xie, W., Yu, J.: Kernel method-based fuzzy clustering algorithm. J. Syst. Eng. Electron. **16**, 160–166 (2005)

3. Girolami, M.: Mercer kernel-based clustering in feature space. IEEE Trans. Neural Netw. **13**, 780–784 (2002)
4. Wang, L., Chan, K.L.: Learning kernel parameters by using class separability measure. In: Proceedings of the Advances in Neural Information Processing Systems, NIPS (2002)
5. Zhang, D.Q., Chen, S.C., Zhou, Z.H.: Learning the kernel parameters in kernel minimum distance. Pattern Recognit. **39**, 133–135 (2006)
6. AL-Sultan, K.S., Fedjki, C.A.: A tabu search-based algorithm for the fuzzy clustering problem. Pattern Recognit. **30**, 2023–2030 (1997)
7. Parsopoulos, K.E., Vrahatis, M.N.: Recent approaches to global optimization problems through particle swarm optimization. Nat. Comput. **1**, 235–306 (2002)
8. Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J.: UCI machine learning repository. University of California, School of Information and Computer Sciences, Irvine. http://www.ics.uci.edu/~mlearn/MLRepository.html (1998)
9. Jeong, Y.S., Kang, I.H., Jeong, M.K.: A new feature selection method for one-class classification problems. IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. **42**, 1500–1509 (2012)