Feature Weighted Kernel Clustering with Application to Medical Data Analysis

Hong Jia[1] and Yiu-ming Cheung[1,2]

[1] Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[2] United International College, Beijing Normal University – Hong Kong Baptist University, Zhuhai, China
{hjia,ymc}@comp.hkbu.edu.hk

Abstract. Clustering technique is an effective tool for medical data analysis as it can work for disease prediction, diagnosis record mining, medical image segmentation, and so on. This paper studies the kernel-based clustering method which can conduct nonlinear partition on input patterns and addresses two challenging issues in unsupervised learning environment: feature relevance estimate and cluster number selection. Specifically, a kernel-based competitive learning paradigm is presented for nonlinear clustering analysis. To distinguish the relevance of different features, a weight variable is associated with each feature to quantify the feature's contribution to the whole cluster structure. Subsequently, the feature weights and cluster assignment are updated alternately during the learning process so that the relevance of features and cluster membership can be jointly optimized. Moreover, to solve the problem of cluster number selection, the cooperation mechanism is further introduced into the presented learning framework and a new kernel clustering algorithm which can automatically select the most appropriate cluster number is educed. The performance of proposed method is demonstrated by the experiments on different medical data sets.

Keywords: Kernel-based Clustering, Competitive Learning, Feature Weight, Cooperation Mechanism, Number of Clusters.

1 Introduction

As an important technique in the research areas of machine learning and pattern recognition, clustering analysis has extensive applications in data mining [10], computer vision [2], bioinformatics [8] and so forth. Traditional clustering algorithms include the k-means algorithm [12] and EM algorithm [14], which have been rated as top ten algorithms in data mining area. Generally, these methods are only suitable for linearly separable clusters. Nevertheless, nonlinearly separable cluster structure is common in the data sets from real-world applications. Under the circumstances, kernel-based clustering methods have been widely studied in the literature [6]. This kind of approach utilizes kernel functions to map the original data into a high dimensional feature space, in which a linear partition will result in a nonlinear partition in the input space.

K. Imamura et al. (Eds.): BHI 2013, LNAI 8211, pp. 496–505, 2013.

Existing kernel-based clustering algorithms, such as the kernel k-means [16] and kernel SOM [9], have played an important role in the analysis of nonlinearly separable data. Nevertheless, two key problems have not been considered by them. The first one is how to determine the number of clusters in unsupervised learning environment. The aforementioned kernel clustering algorithms need the users to specify the exact number of clusters as an input. However, choosing the cluster number is an ad hoc decision based on prior knowledge of given data and it becomes nontrivial when the data has many dimensions [7]. This problem also exists in the traditional methods, such as the k-means and EM algorithms. In the literature, competitive learning paradigm with special mechanism has show its effectiveness of automatic cluster number detection in linear cluster analysis. For example, with penalization mechanism, the Rival Penalized Competitive Learning (RPCL) [18] algorithm can automatically select the cluster number by gradually driving extra seed points far away from the input data set. In this learning approach, for each input, not only the winner among all seed points is updated to adapt to the input, but also the second winner is penalized by a much smaller fixed rate (i.e. delearning rate). Some improved variants of RPCL method include the Rival Penalization Controlled Competitive Learning (RPCCL) [5], Stochastic RPCL (S-RPCL) [4], and distance-sensitive RPCL (DSRPCL) [11]. Besides the penalization mechanism, cooperation strategy can also be utilized for detecting cluster number in competitive learning paradigm. One example is the Competitive and Cooperative Learning (CCL) [3] algorithm, in which the winner of each learning iteration will dynamically cooperate with several nearest rivals to update towards the input data together. Consequently, the CCL can make all the seed points converge to the corresponding cluster centers and the number of those seed points stably locating at different positions is exactly the cluster number. By contrast, to the best of our knowledge, conducting kernel-based clustering without knowing cluster number has not been well studied yet.

Another key problem to be solved in existing kernel clustering methods is the relevance of different features to the clustering analysis. Most clustering algorithms treat the features of data vector equally during clustering process. However, from the practical viewpoint, different features actually have different levels of contribution to the clustering structure. The existing of irrelevant features may even deteriorate the ability of utilized learning model. Therefore, it is expected to pay more attention to the relevant features during clustering process and reduce the negative effect from irrelevant features as much as possible. In supervised learning environment, the most relevant features can be extracted conveniently based on the class label information [15]. Nevertheless, for unsupervised learning, due to the absence of guiding information, evaluating the relevance of different features becomes a more challenging problem. Some methods have been proposed in the literature to address this issue. For example, Mitra et al. [13] proposed a feature similarity measure namely maximum information compression index, based on which the most dissimilar features are selected. Additionally, the Q-α algorithm presented in [17] defines the feature relevance based on the spectral properties of the graph Laplacian of data on

the candidate features and ranks all the features with a least-squares optimization technique in the feature selection process. In these methods, the features are selected prior to the clustering analysis and this operation goes against the fact that the selected feature subset and the clustering result are inter-related. Therefore, it is suggested to take into account the selection of relevant feature jointly with the clustering analysis [19].

To conduct nonlinear clustering analysis in unsupervised learning environment, this paper introduces the competition strategy into the mapped feature space and presents a kernel-based competitive learning method. Moreover, to take into account the relevance of different features, a feature weight variable has been integrated into the clustering framework. This weight estimates the contribution of each feature to the clustering structure by comparing the intra-cluster variance of observations with the whole variance of all patterns in feature space. Subsequently, the partition of clusters and the calculation of feature weights are implemented alternately so that the feature weights and cluster membership can be jointly optimized. Additionally, to learn the number of clusters automatically, we further introduce the cooperation mechanism into the feature weighted competitive learning framework and propose a new kernel-based clustering algorithm, which can conduct nonlinear partition on input data with the cluster number being initialized larger than or equal to the true one. Finally, to investigate the efficacy of presented method, we apply it to variant medical data sets. In practice, clustering technique is a kind of effective tool for medical data analysis as it can do disease prediction, diagnosis record mining, gene clustering, medical image segmentation, and so on. The results of our experiments have shown the good performance of proposed algorithm.

2 Unsupervised Feature Weighted Kernel Clustering

2.1 Kernel-Based Competitive Learning

Given the data set $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ with $\mathbf{x}_i \in \mathbb{R}^d$, the Mercer kernel $K : X \times X \to \mathbb{R}$ can be expressed as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j), \forall i, j \in \{1, 2, \ldots, N\}, \tag{1}$$

where $\Phi : X \to \mathcal{F}$ maps the original space X to a high dimensional feature space $\mathcal{F}$. The clustering in feature space is to find k centers (i.e., $\mathbf{m}_j^{\Phi} \in \mathcal{F}$ with $j = 1, 2, \ldots, k$), which partition the mapped patterns into different groups so that the summation of distances between each center and its cluster members in feature space is minimized. Generally, each center $\mathbf{m}_j^{\Phi}$ can be written as a combination of the mapped patterns [16]. Accordingly, we have

$$\mathbf{m}_j^{\Phi} = \sum_{i=1}^{N} \alpha_{ji} \Phi(\mathbf{x}_i), \tag{2}$$

where α_{ji} is a non-negative coefficient. Subsequently, based on the kernel trick [16], the squared distance between a mapped pattern $\Phi(\mathbf{x}_i)$ and a center $\mathbf{m}_j^{\Phi}$ can be calculated by

$$\begin{aligned}\left\|\Phi(\mathbf{x}_i)-\mathbf{m}_j^{\Phi}\right\|^2 &= \left\|\Phi(\mathbf{x}_i)-\sum_{t=1}^{N}\alpha_{jt}\Phi(\mathbf{x}_t)\right\|^2 \\ &= K(\mathbf{x}_i,\mathbf{x}_i)-2\sum_{t=1}^{N}\alpha_{jt}K(\mathbf{x}_i,\mathbf{x}_t)+\sum_{r,s=1}^{N}\alpha_{jr}\alpha_{js}K(\mathbf{x}_r,\mathbf{x}_s).\end{aligned} \tag{3}$$

For the competitive learning method, given a data point $\mathbf{x}_t$ each time, the winner $\mathbf{m}_c^{\Phi}$ among k centers is determined by

$$c=\arg\min_{1\le j\le k}\{\gamma_j\left\|\Phi(\mathbf{x}_t)-\mathbf{m}_j^{\Phi}\right\|^2\} \tag{4}$$

with the relative winning frequency γ_j of $\mathbf{m}_j^{\Phi}$ defined as

$$\gamma_j=\frac{n_j}{\sum_{i=1}^{k}n_i}, \tag{5}$$

where n_j is the winning times of $\mathbf{m}_j^{\Phi}$ in the past [1]. Synthesizing Eq. (3) and Eq. (4), we can get

$$c=\arg\min_{1\le j\le k}\{\gamma_j[\sum_{r,s=1}^{N}\alpha_{jr}\alpha_{js}K(\mathbf{x}_r,\mathbf{x}_s)-2\sum_{i=1}^{N}\alpha_{ji}K(\mathbf{x}_t,\mathbf{x}_i)]\}. \tag{6}$$

Subsequently, $\mathbf{x}_t$ is assigned to the winning cluster and the corresponding cluster center is updated with

$$\mathbf{m}_c^{\Phi(t)}=\mathbf{m}_c^{\Phi(t-1)}+\eta(\Phi(\mathbf{x}_t)-\mathbf{m}_c^{\Phi(t-1)}), \tag{7}$$

where η is a small learning rate. Substituting Eq. (2) into Eq. (7) yields

$$\begin{aligned}\sum_{i=1}^{N}\alpha_{ci}^{(t)}\Phi(\mathbf{x}_i) &= \sum_{i=1}^{N}\alpha_{ci}^{(t-1)}\Phi(\mathbf{x}_i)+\eta\Phi(\mathbf{x}_t)-\eta\sum_{i=1}^{N}\alpha_{ci}^{(t-1)}\Phi(\mathbf{x}_i) \\ &= (1-\eta)\sum_{i=1}^{N}\alpha_{ci}^{(t-1)}\Phi(\mathbf{x}_i)+\eta\Phi(\mathbf{x}_t).\end{aligned} \tag{8}$$

Therefore, the updating of winning center $\mathbf{m}_c^{\Phi}$ can be handled indirectly by updating the coefficient α_{ci} according to

$$\alpha_{ci}^{(t)}=\begin{cases}(1-\eta)\alpha_{ci}^{(t-1)}, & \text{if } i\neq t,\\ (1-\eta)\alpha_{ci}^{(t-1)}+\eta, & \text{otherwise.}\end{cases} \tag{9}$$

2.2 Estimate of Feature Weights

Suppose the input patterns are represented by d features $\{f_1, f_2, \ldots, f_d\}$. To evaluate the relevance of different features to the clustering analysis, we associate a weight w_l ($w_l \in [0,1]$)with each feature f_l and let $\mathbf{w} = (w_1, w_2, \ldots, w_d)$ be the weight vector. In this paper, Gaussian kernel function is utilized. That is,

$$K(\mathbf{x}_r, \mathbf{x}_s) = \exp\left(-\frac{\|\mathbf{x}_r - \mathbf{x}_s\|^2}{2\sigma^2}\right), \tag{10}$$

where σ is a suitable constant. Integrating the feature weights, we can further get

$$K(\mathbf{x}_r, \mathbf{x}_s) = \exp\left(-\frac{\sum_{l=1}^{d} w_l (x_{rl} - x_{sl})^2}{2\sigma^2}\right). \tag{11}$$

The contribution of each feature to the clustering analysis will depend on its weight value. Next, to estimate the feature weights, we take into account the relevance of different features to the cluster structure. As pointed out in [19], a feature can be regarded less relevant if the variance of observations in a cluster is closer to the global variance of observations in all clusters along this feature. Following this guidance, the feature weight can be estimated by

$$w_l = \frac{1}{k}\sum_{j=1}^{k} \max(0, 1 - \frac{\delta_{lj}^2}{\delta_l^2}),\ l = 1, 2, \ldots, d, \tag{12}$$

where δ_{lj}^2 calculates the variance of the observations in jth cluster along the lth dimension and δ_l^2 is the global variance of all observations on the lth feature. In the mapped feature space of kernel clustering, δ_{lj}^2 and δ_l^2 can be calculated respectively as follows:

$$\delta_{lj}^2 = \frac{1}{N_j - 1}\sum_{i=1}^{N_j} \left\| \Phi(x_{il}) - \frac{1}{N_j}\sum_{t=1}^{N_j} \Phi(x_{tl}) \right\|^2,\ \mathbf{x}_i, \mathbf{x}_t \in j\text{th cluster}, \tag{13}$$

$$\delta_l^2 = \frac{1}{N - 1}\sum_{i=1}^{N} \left\| \Phi(x_{il}) - \frac{1}{N}\sum_{t=1}^{N} \Phi(x_{tl}) \right\|^2, \tag{14}$$

where N_j stands for the number of patterns in the jth cluster. The squared distances in these two formulas are given by

$$\left\| \Phi(x_{il}) - \frac{1}{N_j}\sum_{t=1}^{N_j} \Phi(x_{tl}) \right\|^2 = 1 - \frac{2}{N_j}\sum_{t=1}^{N_j} K(x_{il}, x_{tl}) + \frac{1}{N_j^2}\sum_{r,s=1}^{N_j} K(x_{rl}, x_{sl}) \tag{15}$$

$$\left\| \Phi(x_{il}) - \frac{1}{N}\sum_{t=1}^{N} \Phi(x_{tl}) \right\|^2 = 1 - \frac{2}{N}\sum_{t=1}^{N} K(x_{il}, x_{tl}) + \frac{1}{N^2}\sum_{r,s=1}^{N} K(x_{rl}, x_{sl}), \tag{16}$$

where $K(x_{rl}, x_{sl}) = \exp\left(-\frac{(x_{rl}-x_{sl})^2}{2\sigma^2}\right)$. Subsequently, when an intermediate cluster membership is obtained during the learning process, the feature weights can be adjusted accordingly based on Eq. (12) to Eq. (16).

2.3 Implementation of Cooperation Mechanism

To learn the true number of clusters automatically, we introduce the cooperation mechanism into the competitive learning framework and propose a new algorithm which can conduct kernel-based clustering without knowing exact cluster number. Specifically, we set the number of initial cluster centers (also called *seed points* hereinafter) not less than the true one, i.e. $k \geq k^*$. Subsequently, once the winner $\mathbf{m}_c^{\Phi}$ is selected, the other cluster centers which have fallen into its territory will cooperate with it. That is, any center $\mathbf{m}_j^{\Phi}$ $(j \neq c)$ satisfies

$$\left\|\mathbf{m}_c^{\Phi} - \mathbf{m}_j^{\Phi}\right\|^2 \leq \left\|\mathbf{m}_c^{\Phi} - \Phi(\mathbf{x}_t)\right\|^2 \tag{17}$$

will be selected as a cooperator of the winner. Based on Eq. (2) and Eq. (3), Eq. (17) can be rewritten as

$$\sum_{r,s=1}^{N} (\alpha_{jr}\alpha_{js} - 2\alpha_{cr}\alpha_{js})K(\mathbf{x}_r, \mathbf{x}_s) \leq K(\mathbf{x}_t, \mathbf{x}_t) - 2\sum_{i=1}^{N} \alpha_{ci}K(\mathbf{x}_t, \mathbf{x}_i). \tag{18}$$

When the cooperating team is formed, each member $\mathbf{m}_u^{\Phi}$ among it will be adjusted towards the given data point with a dynamic learning rate according to

$$\mathbf{m}_u^{\Phi(t)} = \mathbf{m}_u^{\Phi(t-1)} + \eta\rho_u(\Phi(\mathbf{x}_t) - \mathbf{m}_u^{\Phi(t-1)}), \tag{19}$$

where

$$\rho_u = \frac{\left\|\mathbf{m}_c^{\Phi(t-1)} - \Phi(\mathbf{x}_t)\right\|^2}{\max\left(\left\|\mathbf{m}_c^{\Phi(t-1)} - \Phi(\mathbf{x}_t)\right\|^2, \left\|\mathbf{m}_u^{\Phi(t-1)} - \Phi(\mathbf{x}_t)\right\|^2\right)}. \tag{20}$$

Based on Eq. (2), Eq. (19) can be further rewritten as

$$\alpha_{ui}^{(t)} = \begin{cases} (1-\eta\rho_u)\alpha_{ui}^{(t-1)}, & \text{if } i \neq t, \\ (1-\eta\rho_u)\alpha_{ui}^{(t-1)} + \eta\rho_u, & \text{otherwise.} \end{cases} \tag{21}$$

The adjusting factor ρ_u here ensures that the learning rate of cooperators is not more than the winner's and also adaptively adjusts the cooperating rate based on the distance between the cooperator and the current input. This competitive learning model with cooperation mechanism can make all the seed points converge to the corresponding cluster centers. Finally, the number of those seed points stably locating at different positions is exactly the cluster number.

2.4 Feature Weighted Kernel Clustering Algorithm

Based on the description given in the former sub-sections, the feature weighted competitive learning algorithm with cooperation mechanism for kernel-based clustering analysis can be summarized as Algorithm 1. Specifically, to randomly initialize the k cluster centers in feature space, we make a random permutation on the order of input data and then initialize the centers as the first k mapped patterns. That is, we set $\alpha_{ji} = \delta_{ji}$, where $\delta_{ji} = 1$ if $i = j$ and 0 otherwise. In the stopping criterion, T stands for the number of learning epochs and scanning the whole data set once means an epoch. ε is a very small number, which has been set at 10^{-6} in our experiments. The convergency index e^{Φ} is calculated by

$$\begin{aligned} e^{\Phi} &= \sum_{j=1}^{k} \left\| \mathbf{m}_j^{\Phi(T)} - \mathbf{m}_j^{\Phi(T-1)} \right\|^2 \\ &= \sum_{j=1}^{k} \left\| \sum_{i=1}^{N} \alpha_{ji}^{(T)} \Phi(\mathbf{x}_i) - \sum_{i=1}^{N} \alpha_{ji}^{(T-1)} \Phi(\mathbf{x}_i) \right\|^2 \\ &= \sum_{j=1}^{k} \left[\sum_{r,s=1}^{N} \alpha_{jr}^{(T)} \alpha_{js}^{(T)} K(\mathbf{x}_r, \mathbf{x}_s) - 2 \sum_{r,s=1}^{N} \alpha_{jr}^{(T)} \alpha_{js}^{(T-1)} K(\mathbf{x}_r, \mathbf{x}_s) \right. \\ &\qquad \left. + \sum_{r,s=1}^{N} \alpha_{jr}^{(T-1)} \alpha_{js}^{(T-1)} K(\mathbf{x}_r, \mathbf{x}_s) \right], \end{aligned} \tag{22}$$

where $T-1$ and T are two sequential learning epochs.

Algorithm 1. Feature Weighted Kernel Clustering Algorithm (FWKC)

1: **Input:** data set X, learning rate η and an initial value of k ($k \geq k^*$)
2: **Output:** cluster label $Y = \{y_1, y_2, \ldots, y_N\}$ and cluster number k^*
3: Randomly initialize the k cluster centers, denoted as $\{\mathbf{m}_1^{\Phi(0)}, \mathbf{m}_2^{\Phi(0)}, \ldots, \mathbf{m}_k^{\Phi(0)}\}$. Set $n_j^{(0)} = 1$ with $j = 1, 2, \ldots, k$, $w_l = 1$ with $l = 1, 2, \ldots, d$, and $t = 1$.
4: **repeat**
5: **for** $i = 1$ **to** N **do**
6: Determine the winning unit $\mathbf{m}_c^{\Phi(t-1)}$ according to Eq. (6) and assign $\mathbf{x}_i$ to cluster c.
7: Let $S_u^{\Phi} = \emptyset$, and then add $\mathbf{m}_j^{\Phi(t-1)}$ ($j \in \{1, 2, \ldots, k\}$, $j \neq c$) into S_u^{Φ} if it satisfies Eq. (18).
8: Update all members in S_u^{Φ} by Eq. (21).
9: Update the winner $\mathbf{m}_c^{\Phi}$ by Eq. (9).
10: Update n_c by $n_c^{(t)} = n_c^{(t-1)} + 1$, and increase t by 1.
11: **end for**
12: Calculate the feature weights $\mathbf{w}$ according to Eq. (12).
13: **until** $e^{\Phi} \leq \varepsilon$ or $T \geq T_{max}$

3 Experimental Results

To investigate the performance of proposed FWKC algorithm, we applied it to four medical data sets from UCI Machine Learning Data Repository[1] and compared its results to that obtained by standard kernel k-means method [16]. The general information of utilized data sets has been summarized in Table 1. In the experiments, each algorithm has executed 20 times under different settings of k. Table 1 has given the chosen value of σ in the Gaussian kernel function for each data set. The learning rate η in FWKC algorithm was set at 0.0001.

Table 1. Main statistics of utilized data sets

Data set	N	d	k*	σ	Diagnosis task
Breast Cancer	569	30	2	500	Malignant or benign breast tumor
Indians Diabetes	768	8	2	150	Diabetes positive or negative
Mammographic Mass	961	4	2	20	Benign or malignant mammographic masses
Cardiotocography	2126	21	3	45	Fetal state: normal, suspect, or pathologic

According to [7], the performance of clustering algorithms with capability of cluster number selection can be evaluated by Partition Quality (PQ) index:

$$PQ = \begin{cases} \frac{\sum_{i=1}^{k^*} \sum_{j=1}^{k'} [p(i,j)^2 \cdot (p(i,j)/p(j))]}{\sum_{i=1}^{k^*} p(i)^2}, & \text{if } k' > 1, \\ 0, & \text{otherwise,} \end{cases} \tag{23}$$

where k^* is the true number of clusters and k' is the cluster number learned by the algorithm. The term $p(i,j)$ calculates the frequency-based probability that a data point is labeled i by the true label and labeled j by the obtained label. This PQ metric achieves the maximum value 1 when the obtained labels induce the same partition as the true ones. Additionally, we have also utilized the Rand Index (RI) to measure the clustering accuracy for reference, which is given by

$$RI = \frac{TP + TN}{TP + FP + FN + TN}, \tag{24}$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively.

Table 2 has given the experimental results obtained by kernel k-means and FWKC algorithms in terms of cluster number, Partition Quality, and Rand Index. From the records we can find that the kernel k-means algorithm cannot learn the cluster number as its results always fit the initial values of k. The observation that sometimes the cluster number presented by kernel k-means was less than the setting value is due to the generation of empty clusters. By contrast, the FWKC algorithm can give a good estimate for the cluster number

[1] `http://archive.ics.uci.edu/ml/`

during the clustering process. Therefore, when the initial cluster number was set much larger than the true one, the partition quality of kernel k-means degraded significantly while the FWKC did not. Moreover, we can find that the difference of clustering accuracy between kernel k-means and FWKC on Breast Cancer and Cardiotocography data sets is larger than that on the other two data sets. The reason is that the dimensionality of these two data sets is much higher, therefore, the benefit of feature weighting method is more prominent on them.

Table 2. Comparison of clustering results on different data sets

Data set	k	Methods	No. of Clusters	PQ	RI
Breast Cancer	3	Kernel k-means	3.0±0.0	0.3378	0.5048
		FWKC	2.15±0.32	0.5243	0.6850
	5	Kernel k-means	4.6±0.55	0.1746	0.4853
		FWKC	2.35±0.36	0.5012	0.6590
Indians Diabetes	3	Kernel k-means	3.0±0.0	0.3242	0.5152
		FWKC	2.3±0.24	0.3865	0.5946
	5	Kernel k-means	4.4±0.89	0.1947	0.4845
		FWKC	2.45±0.51	0.3673	0.5889
Mammographic	3	Kernel k-means	3.0±0.0	0.2573	0.4896
		FWKC	2.15±0.24	0.3056	0.5259
	5	Kernel k-means	4.6±0.88	0.1274	0.4582
		FWKC	2.3±0.47	0.2713	0.5208
Cardiotocography	4	Kernel k-means	4.0±0.0	0.2346	0.4648
		FWKC	3.2±0.16	0.3508	0.6158
	8	Kernel k-means	8.0±0.0	0.0980	0.4024
		FWKC	3.45±0.65	0.3258	0.5749

4 Conclusion

This paper has presented a novel kernel-based competitive learning model for clustering analysis. In this method, each feature is associated with a weight factor, which is utilized to estimate the relevance of each feature and adjust its contribution to the clustering structure. Moreover, to select the number of clusters automatically in unsupervised learning environment, cooperation mechanism has been further introduced into the competitive learning model and a new kernel-based clustering algorithm which can conduct nonlinear partition on input data without knowing the true cluster number has been presented. Experiments on medical data sets have shown the efficacy of the proposed method.

Acknowledgments. The work described in this paper was supported by the NSFC grant under 61272366, the Faculty Research Grant of Hong Kong Baptist University (HKBU) with the project: FRG2/12-13/082, and the Strategic Development Fund of HKBU: 03-17-033.

References

1. Ahalt, S.C., Krishnamurty, A.K., Chen, P., Melton, D.E.: Competitive learning algorithms for vector quantization. Neural Networks 3(3), 277–291 (1990)
2. Cai, W., Chen, S., Zhang, D.: Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. Pattern Recognition 40(3), 825–838 (2007)
3. Cheung, Y.M.: A competitive and cooperative learning approach to robust data clustering. In: Proceedings of IASTED International Conference on Neural Networks and Computational Intelligence, pp. 131–136 (2004)
4. Cheung, Y.M.: Maximum weighted likelihood via rival penalized em for density mixture clustering with automatic model selection. IEEE Transactions on Knowledge and Data Engineering 17(6), 750–761 (2005)
5. Cheung, Y.M.: On rival penalization controlled competitive learning for clustering with automatic cluster number selection. IEEE Transactions on Knowledge and Data Engineering 17(11), 1583–1588 (2005)
6. Filippone, M., Camastra, F., Masulli, F., Rovetta, S.: A survey of kernel and spectral methods for clustering. Pattern Recognition 41, 176–190 (2008)
7. Hamerly, G., Elkan, C.: Learning the k in k-means. In: Proceedings of the 17th Annual Conference on Neural Information Processing Systems (NIPS), pp. 281–288 (2003)
8. Huang, D.S., Zhao, X.M., Huang, G.B., Cheung, Y.M.: Classifying protein sequences using hydropathy blocks. Pattern Recognition 39(12), 2293–2300 (2006)
9. Inokuchi, R., Miyamoto, S.: Lvq clustering and som using a kernel function. In: Proceedings of IEEE International Conference on Fuzzy Systems, vol. 3, pp. 1497–1500 (2004)
10. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recognition Letters 31(8), 651–666 (2010)
11. Ma, J., Wang, T.: A cost-function approach to rival penalized competitive learning (rpcl). IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics 36(4), 722–737 (2006)
12. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967)
13. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised feature selection using feature similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3), 301–312 (2002)
14. Render, R.A., Walker, H.F.: Mixture densities, maximum likelihood and the em algorithm. SIAM Review 26(2), 195–239 (1984)
15. Robnik-Šikonja, M., Kononenko, I.: Theoretical and empirical analysis of relieff and rrelieff. Machine Learning 53(1), 23–69 (2003)
16. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10(5), 1299–1319 (1998)
17. Wolf, L., Shashua, A.: Feature selection for unsupervised and supervised inference: the emergence of sparsity in a weight-based approach. Journal of Machine Learning Research 6, 1855–1887 (2005)
18. Xu, L., Krzyzak, A., Oja, E.: Rival penalized competitive learning for clustering analysis, rbf net, and curve detection. IEEE Transactions on Neural Networks 4(4), 636–648 (1993)
19. Zeng, H., Cheung, Y.M.: A new feature selection method for gaussian mixture clustering. Pattern Recognition 42, 243–250 (2009)