Neurocomputing 198 (2016) 125–134

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Fuzzy clustering with the entropy of attribute weights

Jin Zhou [a,b], Long Chen [b], C.L. Philip Chen [b,*], Yuan Zhang [a], Han-Xiong Li [c]

[a] School of Information Science and Engineering, University of Jinan, China
[b] Faculty of Science and Technology, University of Macau, Macau, China
[c] Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Received 29 March 2015
Received in revised form
14 September 2015
Accepted 22 September 2015
Available online 8 March 2016

Keywords:
Attribute-weighted clustering
Maximum-entropy regularization
Kernel method
Feature selection

ABSTRACT

For many datasets, it is a difficult work to seek a proper cluster structure which covers the entire feature set. To extract the important features and improve the clustering, the maximum-entropy-regularized weighted fuzzy c-means (EWFCM) algorithm is proposed in this paper. A new objective function is developed in the proposed algorithm to achieve the optimal clustering result by minimizing the dispersion within clusters and maximizing the entropy of attribute weights simultaneously. Then the kernelization of proposed algorithm is realized for clustering the data with 'non-spherical' shaped clusters. Experiments on synthetic and real-world datasets have demonstrated the efficiency and superiority of the presented algorithms.

1. Introduction

Prototype-based partitioning clustering analysis is an efficient tool for data mining, pattern recognition and statistical analysis [1,2]. It partitions the data into clusters according to the similarity of objects, which helps in extraction of new information or discovering new patterns. In the past few decades, various partitioning clustering algorithms have been proposed. K-means algorithm [3] and fuzzy c-means (FCM) algorithm [4] are two well known clustering algorithms. Variants of these two algorithms are further discussed and popularized in [5–8]. However, one drawback of these clustering algorithms is that they treat all features equally in deciding the cluster memberships of objects. This is not desirable in some applications, such as high-dimensions sparse data clustering, where the cluster structure in the dataset is often limited to a subset of features rather than the entire feature set. A better solution is to introduce the proper attribute weight into the clustering process. Weight assignment and feature selection have been a hot research topic in partitioning clustering analysis [9,10].

Desarbo et al. [11] propose the first attribute-weighted method based on K-means algorithm. In [12], Friedmn et al. present the COSA method to calculate a weight for each dimension in each cluster. Jing et al. [13] extend Friedmn's method to optimal variable weighting for high-dimensional sparse data clustering. Recently, Tsai et al. [14] suggest a novel feature weight self-adjustment mechanism embedded into k-means, where feature weight searching is modeled as an optimization problem to simultaneously minimize the separations within clusters and maximize the separations between clusters.

Compared with the hard clustering methods discussed above, the fuzzy clustering technique is peculiarly effective when the boundaries between clusters of data are ambiguous. Moreover, the degree of membership will help us discover complicated relations between a given data object and all clusters [15]. So many attribute-weighted fuzzy clustering methods have been proposed in the past few decades. In [16], Wang et al. use the weighted Euclidean distance to replace the general Euclidean distance in FCM. Borgelt [17] carries out clustering on the selected subspace instead of the full data space by directly assigning zero weights to features that have little information. But one limitation is that they generate the attribute weights from all clusters, rather than each cluster. Keller et al. [18] introduce a basic attribute-weighted FCM algorithm by assigning one influence parameter to each single data dimension for each cluster, while Frigui et al. [19] put forward an approach searching for the optimal prototype parameters and the optimal feature weights simultaneously. Recently, Deng et al. present an enhanced soft subspace clustering (ESSC) algorithm by employing both within-cluster and between-cluster information [20]. In [21], a novel subspace clustering technique has been proposed by introducing the feature interaction using the concepts of fuzzy measures and the Choquet integral. Finally, in [22], Tang et al. give a survey of weighted clustering technologies.

* Corresponding author. Tel.: +86 853 83974950.
E-mail addresses: ise_zhouj@ujn.edu.cn (J. Zhou), longchen@umac.mo (L. Chen), Philip.Chen@ieee.org (C.L.P. Chen), yzhang@ujn.edu.cn (Y. Zhang), mehxli@cityu.edu.hk (H.-X. Li).

http://dx.doi.org/10.1016/j.neucom.2015.09.127
0925-2312/

These weighted fuzzy clustering methods are effective in data clustering and feature selection, but there are still some questions. (1) There is still no good criterion, which is not only valid for the attribute weight assignment but also has a clear physical meaning; (2) for some particular applications, such as 'non-spherical' shaped data clustering, traditional weighted clustering algorithms can not show good performance. Different from other weighted fuzzy clustering methods, we propose a maximum-entropy-regularized weighted fuzzy c-means (EWFCM) clustering algorithm, in which the attribute-weight-entropy regularization is defined in the new objective function to achieve the optimal distribution of attribute weights. So that we can simultaneously minimize the dispersion within clusters and maximize the entropy of attribute weights to stimulate important attributes to contribute to the identification of clusters. Then the good clustering result can be yielded and the important attributes can be extracted for cluster identification. Moreover, for the 'non-spherical' shaped data clustering, developments on kernel technique and their applications have emphasized the need to consider kernel method into the data clustering, which is referred to as kernel-based clustering [25–29]. Therefore, the kernelization of the proposed algorithm is realized for clustering the data with 'non-spherical' shaped clusters. Experiments on synthetic and real-world datasets have demonstrated the efficiency of the proposed algorithms compared with traditional clustering approaches.

The rest of the paper is organized as follows. In Section 2, we present the EWFCM clustering algorithm based on attribute-weight-entropy regularization. The kernelization of EWFCM algorithm is given in Section 3. Section 4 illustrates experiment results of different clustering algorithms on synthetic and real world datasets. Finally, conclusions are drawn in Section 5.

2. Maximum-entropy-regularized weighted fuzzy c-means clustering algorithm

2.1. Algorithm description

In our research, the attribute weighted fuzzy clustering problem is considered as a maximum entropy inference problem [23] (the proof is shown in the Appendix), which aims to search for global regularity and obtain the smoothest reconstructions from the available data. We focus on this problem and propose the maximum-entropy-regularized weighted fuzzy c-means (EWFCM) clustering algorithm, in which a new objective function is developed as (1) by combining the weighted dissimilarity measure and an extra term for attribute-weight-entropy regularization.

$$F(\boldsymbol{U},\boldsymbol{C},\boldsymbol{W})=\sum_{k=1}^{K}\sum_{n=1}^{N}u_{kn}^{\alpha}\sum_{m=1}^{M}w_{km}(x_{nm}-c_{km})^2+\gamma^{-1}\sum_{k=1}^{K}\sum_{m=1}^{M}w_{km}\log w_{km} \tag{1}$$

Subject to $\sum_{k=1}^{K}u_{kn}=1, 0\le u_{kn}\le 1$

$\sum_{m=1}^{M}w_{km}=1, 0\le w_{km}\le 1$

where N is the number of objects, M is the number of object dimensions/attributes, K is the number of clusters, $\boldsymbol{U}=[u_{kn}]$ is a K-by-N matrix, u_{kn} denotes the degree of membership of the *n-th* object belonging to the *k-th* cluster, $\boldsymbol{C}=[c_{km}]$ is a K-by-M matrix, c_{km} denotes cluster center of the *k-th* cluster defined by u_{kn}, $\boldsymbol{W}=[w_{km}]$ is a K-by-M matrix, w_{km} indicates the attribute weight of the *m-th* object dimension in the *k-th* cluster. $\alpha(\alpha>1)$ is the fuzzification coefficient. γ is a positive scalar.

In this new objective function, the first distance-based term controls the shape and size of the clusters and encourages the agglomeration of clusters, while the second term is the negative entropy of attribute weights that regularize the optimal distribution of all attribute weights according with the available data. So that we can simultaneously minimize the dispersion within clusters and maximize the entropy of attribute weights to stimulate important attributes to contribute to the identification of clusters. $\gamma(\gamma>0)$ is a positive regularizing and adjustable parameter. With a proper choice of γ, we can balance the two terms and achieve the optimal and stable solution.

Minimizing $F(\boldsymbol{U}, \boldsymbol{C}, \boldsymbol{W})$ with respect to normalization constraints is a constrained nonlinear optimization problem. Like the traditional FCM algorithm, Picard iteration is also applied to solve this problem. We first fix $\boldsymbol{C}$ and $\boldsymbol{W}$ and find necessary conditions on $\boldsymbol{U}$ to minimize $F(\boldsymbol{U})$. Then we fix $\boldsymbol{W}$ and $\boldsymbol{U}$ and minimize $F(\boldsymbol{C})$ with respect to $\boldsymbol{C}$. Finally, we fix $\boldsymbol{U}$ and $\boldsymbol{C}$ and minimize $F(\boldsymbol{W})$ with respect to $\boldsymbol{W}$. The matrices $\boldsymbol{U}$, $\boldsymbol{C}$ and $\boldsymbol{W}$ are updated according to the Eqs. (2), (3) and (4) respectively.

$$u_{kn}=\frac{1}{\sum_{h=1}^{K}\left(\frac{\sum_{m=1}^{M}w_{km}(x_{nm}-c_{km})^2}{\sum_{m=1}^{M}w_{hm}(x_{nm}-c_{hm})^2}\right)^{\frac{1}{\alpha-1}}} \tag{2}$$

for $1\le k\le K, 1\le n\le N$

$$c_{km}=\frac{\sum_{n=1}^{N}u_{kn}^{\alpha}x_{nm}}{\sum_{n=1}^{N}u_{kn}^{\alpha}} \tag{3}$$

for $1\le k\le K, 1\le m\le M$

$$w_{km}=\frac{\exp\left(-\gamma\sum_{n=1}^{N}u_{kn}^{\alpha}(x_{nm}-c_{km})^2\right)}{\sum_{s=1}^{M}\exp\left(-\gamma\sum_{n=1}^{N}u_{kn}^{\alpha}(x_{ns}-c_{ks})^2\right)} \tag{4}$$

for $1\le k\le K, 1\le m\le M$

The process is repeated until the stop criteria are satisfied like the improvement in the objective function values is small enough. The pseudo-code of EWFCM algorithm is summarized as follows.

Algorithm. EWFCM Clustering Algorithm

Input: The number of objects N, object attributes M, clusters K; parameter α and γ, and a small enough error ε.

Randomly choose K objects as initial cluster centers and set all initial weights of each attribute to $1/M$, t=0;

Repeat

Update the partition matrix $\boldsymbol{U}$ by (2);
Update the cluster center matrix $\boldsymbol{C}$ by (3);
Update the attribute weight matrix $\boldsymbol{W}$ by (4);
Calculate the objective function $F^{(t)}$ by (1);
$t++$;

Until $\left|F^{(t)}-F^{(t-1)}\right|<\varepsilon$

According to the definitions above, the computational complexity of EWFCM is $O(tKNM)$, where t is the total number of iterations required, and N, M, K indicate the number of data objects, data dimensions and clusters respectively. For the storage, we need $O(NM + 2KM+KN)$ space to keep the data objects (NM), the cluster centers (KM), the membership matrix (KN) and the attribute weight matrix (KM). In a word, EWFCM is an efficient clustering algorithm.

2.2. Iterative optimization analysis

Theorem 1. *Let $\boldsymbol{C}$ and $\boldsymbol{W}$ be fixed, $\boldsymbol{U}$ is a strict local minimum of $F(\boldsymbol{U})$ if and only if $\boldsymbol{U}$ is calculated via (2).*

Theorem 2. *Let $\boldsymbol{U}$ and $\boldsymbol{W}$ be fixed, $\boldsymbol{C}$ is a strict local minimum of $F(\boldsymbol{C})$ if and only if $\boldsymbol{C}$ is calculated via (3).*

Theorem 3. *Let $\boldsymbol{U}$ and $\boldsymbol{C}$ be fixed, $\boldsymbol{W}$ is a strict local minimum of $F(\boldsymbol{W})$ if and only if $\boldsymbol{W}$ is calculated via (4).*

The Theorems 1 and 2 can be proven by using the method as [24]. Because the main difference between FCM and EWFCM is the attribute weight introduced. Therefore, we will only provide a detailed analysis for Theorem 3.

Firstly, we need to prove that Eq. (4) is the necessary condition for the minimum of $F(\boldsymbol{W})$.

Proof. We use the Lagrangian multiplier technique to solve the following unconstrained minimization problem.

$$G(\boldsymbol{W},\boldsymbol{\Lambda})=\sum_{k=1}^{K}\sum_{n=1}^{N}u_{kn}^{\alpha}\sum_{m=1}^{M}w_{km}(x_{nm}-c_{km})^2 +\gamma^{-1}\sum_{k=1}^{K}\sum_{m=1}^{M}w_{km}\log w_{km}-\sum_{k=1}^{K}\lambda_k\Big(\sum_{m=1}^{M}w_{km}-1\Big) \quad (5)$$

where $\boldsymbol{\Lambda}=[\lambda_1,\lambda_2,...,\lambda_K]^T$ is a vector containing the Lagrange multipliers corresponding to the constraints.

The optimization problem in (5) can be decomposed into k independent sub-minimization problems.

$$G_k(\boldsymbol{W},\lambda_k)=\sum_{n=1}^{N}u_{kn}^{\alpha}\sum_{m=1}^{M}w_{km}(x_{nm}-c_{km})^2 +\gamma^{-1}\sum_{m=1}^{M}w_{km}\log w_{km}-\lambda_k\left(\sum_{m=1}^{M}w_{km}-1\right) \quad (6)$$

for $1\le k\le K$

By setting the gradient of $G_k(\boldsymbol{W},\lambda_k)$ to zero with respect to λ_k and w_{km}, we obtain

$$\frac{\partial G_k(\boldsymbol{W},\lambda_k)}{\partial\lambda_k}=-\left(\sum_{m=1}^{M}w_{km}-1\right)=0 \quad (7)$$

$$\frac{\partial G_k(\boldsymbol{W},\lambda_k)}{\partial w_{km}}=\sum_{n=1}^{N}u_{kn}^{\alpha}(x_{nm}-c_{km})^2+\gamma^{-1}(\log w_{km}+1)-\lambda_k=0 \quad (8)$$

From (8), we obtain

$$\begin{aligned} w_{km}&=\exp\Big(-\gamma\sum_{n=1}^{N}u_{kn}^{\alpha}(x_{nm}-c_{km})^2-\gamma^2+\gamma\lambda_k\Big)\\ &=\exp\Big(-\gamma\sum_{n=1}^{N}u_{kn}^{\alpha}(x_{nm}-c_{km})^2\Big)\exp(-\gamma^2+\gamma\lambda_k) \end{aligned} \quad (9)$$

Substituting (9) into (7), we have

$$\begin{aligned} \sum_{s=1}^{M}w_{ks}&=\sum_{s=1}^{M}\exp\left(-\gamma\sum_{n=1}^{N}u_{kn}^{\alpha}(x_{ns}-c_{ks})^2\right)\exp(-\gamma^2+\gamma\lambda_k)\\ &=\exp(-\gamma^2+\gamma\lambda_k)\sum_{s=1}^{M}\exp\left(-\gamma\sum_{n=1}^{N}u_{kn}^{\alpha}(x_{ns}-c_{ks})^2\right)\\ &=1 \end{aligned} \quad (10)$$

It follows that

$$\exp(-\gamma^2+\gamma\lambda_k)=\frac{1}{\sum_{s=1}^{M}\exp\left(-\gamma\sum_{n=1}^{N}u_{kn}^{\alpha}(x_{ns}-c_{ks})^2\right)} \quad (11)$$

Substituting (11) into (9), we obtain (4). This completes the proof. □

Secondly, we can prove that the Eq. (4) is the sufficient condition for the minimum of $F(\boldsymbol{W})$.

Proof. To show sufficiency of Theorem 3, we can examine Hessian matrix of Lagrangian of $F(\boldsymbol{W})$ denoted as $\boldsymbol{H}(\boldsymbol{W})$. According to (8), we have

$$h_{jl,km}(\boldsymbol{W},\lambda_k)=\frac{\partial}{\partial w_{jl}}\left(\frac{\partial G_k(\boldsymbol{W},\lambda_k)}{\partial w_{km}}\right)=\begin{cases}\frac{1}{\gamma w_{km}}, & \text{if } j=k,\ l=m\\ 0, & \text{otherwise}\end{cases} \quad (12)$$

Since we know $\gamma>0$ and $w_{km}>0$ (according to (4)), the $\boldsymbol{H}(\boldsymbol{W})$ is a diagonal matrix, and the diagonal entries are positive. So the $\boldsymbol{H}(\boldsymbol{W})$ is positive definite. $F(\boldsymbol{W})$ must have the minimum point, and Eq. (4) is sufficient for $\boldsymbol{W}$ to be a local minimum of $F(\boldsymbol{W})$. Then Theorem 3 can be validated. This completes the proof. □

3. Kernelization of the EWFCM clustering algorithm

Since FCM uses the squared-norm to evaluate similarity between objects and prototypes, it can only be helpful in clustering the data with 'spherical' clusters. For the data with 'non-spherical' clusters, the idea of performing clustering in high-dimensional feature space with mercer kernel based mapping need to be considered [25]. Recently, the kernel method has been widely applied to fuzzy clustering, which is referred to as kernel-based fuzzy clustering [25–27]. The essence of kernel method is to perform a non-linear mapping Φ from the original d-dimensional space $\mathbf{R}^d$ to a high-dimensional kernel space $\mathbf{H}$ [28,29]. Then the linear classifier in the kernel space can be used to solve the clustering problem which could be highly non-linear in the original feature space. Kernel method takes advantage of the fact that dot products in the kernel space can be expressed as a Mercer kernel K given by $K(\mathbf{x},\mathbf{y})=\Phi(\mathbf{x})^{\mathrm{T}}\Phi(\mathbf{y})$ where $\mathbf{x},\mathbf{y}\in\mathbf{R}^d$. Proverbially used Mercer kernels include Gaussian kernel $\left(K(\mathbf{x},\mathbf{y})=\exp\left(\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}\right)\right)$, Polynomial kernel $(K(\mathbf{x},\mathbf{y})=(\mathbf{x}^{\mathrm{T}}\mathbf{y}+\theta)^p)$, and so on [28,29]. In our research, only the Gaussian kernel is used. The detail of the kernel-based EWFCM (KEWFCM) clustering algorithm is shown in the following.

KEWFCM minimizes the following objective function subject to the same constraints as EWFCM.

$$F(\boldsymbol{U},\overline{\boldsymbol{C}},\boldsymbol{W})=\sum_{k=1}^{K}\sum_{n=1}^{N}u_{kn}^{\alpha}\sum_{m=1}^{M}w_{km}\|\Phi(x_{nm})-\overline{\mathbf{c}}_{km}\|^2 +\gamma^{-1}\sum_{k=1}^{K}\sum_{m=1}^{M}w_{km}\log w_{km} \quad (13)$$

Subject to $\sum_{k=1}^{K}u_{kn}=1,\quad 0\le u_{kn}\le 1$

$\sum_{m=1}^{M}w_{km}=1,\quad 0\le w_{km}\le 1$

where $\boldsymbol{U}=[u_{kn}]_{K\times N}$ is the membership degree matrix in the original space. $\boldsymbol{W}=[w_{km}]_{K\times M}$ is the attribute weight matrix in the original space. $\overline{\boldsymbol{C}}=[\overline{\mathbf{c}}_{km}]_{K\times M}$ is the cluster center matrix in the kernel space. Φ is the non-linear mapping from the original feature space to the kernel space. α is the fuzzification coefficient and γ is a positive scalar.

Given the Euclidean distance and optimizing F with respect to $\bar{\mathbf{c}}_{km}$ located in the kernel space such that $\frac{\partial F}{\partial \bar{\mathbf{c}}_{km}}=0$, we have

$$\bar{\mathbf{c}}_{km}=\frac{\sum_{n=1}^{N} u_{kn}^{\alpha}\Phi(x_{nm})}{\sum_{n=1}^{N} u_{kn}^{\alpha}} \tag{14}$$

According to the definition of the kernel function [26], we have

$$\begin{aligned}
&\|\Phi(x_{nm})-\bar{\mathbf{c}}_{km}\|^2\\
&=(\Phi(x_{nm})-\bar{\mathbf{c}}_{km})^T(\Phi(x_{nm})-\bar{\mathbf{c}}_{km})\\
&=\Phi(x_{nm})^T\Phi(x_{nm})-2\Phi(x_{nm})^T\bar{\mathbf{c}}_{km}+\bar{\mathbf{c}}_{km}{}^T\bar{\mathbf{c}}_{km}\\
&=K(x_{nm},x_{nm})-2\frac{\sum_{s=1}^{N}u_{ks}^{\alpha}K(x_{nm},x_{sm})}{\sum_{s=1}^{N}u_{ks}^{\alpha}}+\frac{\sum_{s=1}^{N}\sum_{t=1}^{N}u_{ks}^{\alpha}u_{kt}^{\alpha}K(x_{sm},x_{tm})}{\left(\sum_{t=1}^{N}u_{kt}^{\alpha}\right)^2}
\end{aligned} \tag{15}$$

Picard iteration is also applied to solve this problem. The matrices $\boldsymbol{U}$ and $\boldsymbol{W}$ are updated corresponding to the Eqs. (16) and (17) respectively.

$$u_{kn}=\frac{1}{\sum_{h=1}^{K}\left(\frac{\sum_{m=1}^{M}w_{km}\|\Phi(x_{nm})-\bar{\mathbf{c}}_{km}\|^2}{\sum_{m=1}^{M}w_{hm}\|\Phi(x_{nm})-\bar{\mathbf{c}}_{hm}\|^2}\right)^{\frac{1}{\alpha-1}}} \tag{16}$$

for $1\le k\le K,\quad 1\le n\le N$

$$w_{km}=\frac{\exp(-\gamma\sum_{n=1}^{N}u_{kn}^{\alpha}\|\Phi(x_{nm})-\bar{\mathbf{c}}_{km}\|^2)}{\sum_{l=1}^{M}\exp(-\gamma\sum_{n=1}^{N}u_{kn}^{\alpha}\|\Phi(x_{nl})-\bar{\mathbf{c}}_{km}\|^2)} \tag{17}$$

for $1\le k\le K, 1\le m\le M$

4. Experiments

To evaluate the performance of the proposed algorithms (EWFCM and KEWFCM), several FCM-type clustering algorithms (the FCM algorithm [4], the attribute weighting FCM algorithm (WFCM) [17], the KFCM algorithm [26], the weighted fuzzy kernel-clustering algorithm (WFKCA) [30]) are chosen for comparative analysis. A series of experiments are performed with various datasets. All data are normalized to be between zero and one. Firstly, some real world datasets from UCI Machine Learning Repository are chosen to verify the efficiency of the proposed algorithms compared with the traditional fuzzy clustering methods and other attribute weighted clustering methods. To better understand the effectiveness of the kernel method for clustering, some synthetic datasets with 'non-spherical' shaped clusters are then created for further experiment. The main motivation for development of the EWFCM algorithm is to cluster high-dimensional sparse data, the Reuters Transcribed text dataset is finally used to demonstrate the efficiency of the attribute-weight-entropy regularization technique for data clustering and feature extraction. All the clustering algorithms are implemented with C++ language on a computer with 3.4 GHz CPU and 4 GB RAM. The fuzzification coefficient α is uniformly set to 2.0. The error threshold value ε is set to 2^{-6}. The Gaussian kernel, $K(\mathbf{x},\mathbf{y})=\exp\left(\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}\right)$, $\sigma^2>0$, is adopted in the kernel-based clustering algorithms and the parameter σ^2 is varied from 2^{-10} to 2^5. Each algorithm is executed on each dataset for 100 times, and the cluster centers are randomly initialized at each time.

4.1. Performance metrics

In order to comparative analysis, three kinds of performance metrics are applied in the experiments.

- **Iteration number:** the iteration number is a common measure used to indicate the speed of convergence of the clustering algorithm. For 100 times, the average iteration number (AIN) is used in the experiments.
- **Classification rate:** the classification rate is a measure used to determine how well clustering algorithms perform on the given dataset with a known cluster structure [26]. It can be measured by (18), which is expressed as a percentage in this paper. For 100 times, the average classification rate (ACR) is used in the experiments.

$$CR=\frac{\sum_{k=1}^{K}d_k}{N} \tag{18}$$

where d_k is the number of objects correctly identified in the *k-th* cluster, and N is the number of all objects in the dataset.

- **Normalized mutual information:** the normalized mutual information provides a symmetric measure to quantify the statistical information shared between two cluster distributions [31]. For 100 times, the average normalized mutual information (ANMI) is used in the experiments.

$$NMI(R,Q)=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}P(i,j)\log\frac{P(i,j)}{P(i)P(j)}}{\sqrt{H(R)H(Q)}} \tag{19}$$

where R, Q are two partitions of the dataset. Assume R and Q have I and J clusters respectively. $P(i)$ is the probability that a randomly selected object from the dataset falls into cluster R_i in partition R. $P(i,j)$ denotes the probability that an object belongs to cluster R_i in R and cluster Q_j in Q. $H(R)$ is the entropy associated with all probabilities $P(i)$ $(1\le i\le I)$ in partition R.

4.2. UCI machine learning datasets

Five real-world datasets selected from UCI repository [32] are experimented with including: Iris (N=150, K=3, M=4), Glass (N=214, K=6, M=9), Ionosphere (N=351, K=2, M=33), Haberman (N=306, K=2, M=3), Heart (N=267, K=2, M=44). Here N is the number of objects, M is the number of object dimensions, K is the number of clusters. The clustering results of different algorithms are shown in Table 1. The first observation is that the attribute-weighted clustering algorithms (the WFCM method, the EWFCM method, the WFKCA method, and the KEWFCM method) obtain better clustering performance in ACR (more than 6% promotion on the Iris dataset, and more than 10% promotion on the Haberman dataset and the Heart dataset) and ANMI with compared to the traditional approaches without considering attribute weight assignment (the FCM method and the KFCM method). And more important, further performance improvement in ACR (more than 10% promotion on the Glass dataset and the Heart dataset) and ANMI are bought by our proposed clustering algorithms with the attribute-weight-entropy regularization technique.

In order to have an intuitive understanding of the physical properties of the attribute weight assignment, we further investigate the distribution of four attributes of the Iris dataset shown in Fig. 1. We can clearly see that attribute 3 and attribute 4 are

Table 1
Statistics of different clustering algorithms on UCI machine learning datasets in terms of AIN, ACR, ANMI.

Dataset	FCM	WFCM	EWFCM	KFCM(σ^2)	KFKCA(σ^2)	KEWFCM(σ^2)
Iris	21.2	*15.8	**26.2**	31.4(2^{-2})	32.8(2^{-4})	**39.6(2^{-3})**
	0.8933	0.9533	**0.9666**	0.9266	0.9603	***0.9733**
	0.7433	0.8497	**0.8801**	0.7899	0.8616	***0.9011**
Glass	56.2	50.2	**63.8**	45.4(2^{-6})	*43.7(2^{-4})	**71.6(2^{-5})**
	0.4208	0.4318	**0.5439**	0.5271	0.5306	***0.5775**
	0.2974	0.3061	**0.4263**	0.3624	0.4017	***0.4651**
Ionosphere	*13.9	17.4	**44.8**	32.8(2^{2})	37.1(2^{-1})	**58.8(2°)**
	0.7094	0.7151	**0.7658**	0.7350	0.7483	***0.7920**
	0.1299	0.1343	**0.2026**	0.1861	0.1970	***0.2275**
Haberman	17.1	21.2	**18.2**	*16.8(2^{-8})	20.4(2^{-5})	**27.4(2^{-4})**
	0.5196	0.7516	**0.7712**	0.7078	0.7595	***0.7745**
	0.0024	0.0767	**0.0992**	0.0304	0.0812	***0.1061**
Heart	41.0	*32.6	**48.4**	43.8(2^{-3})	47.3(2^{0})	**54.4(2^{-1})**
	0.5131	0.6142	**0.7288**	0.6281	0.7043	***0.7378**
	0.0052	0.0146	**0.0445**	0.0216	0.0350	***0.0503**

*The best performance among the group.

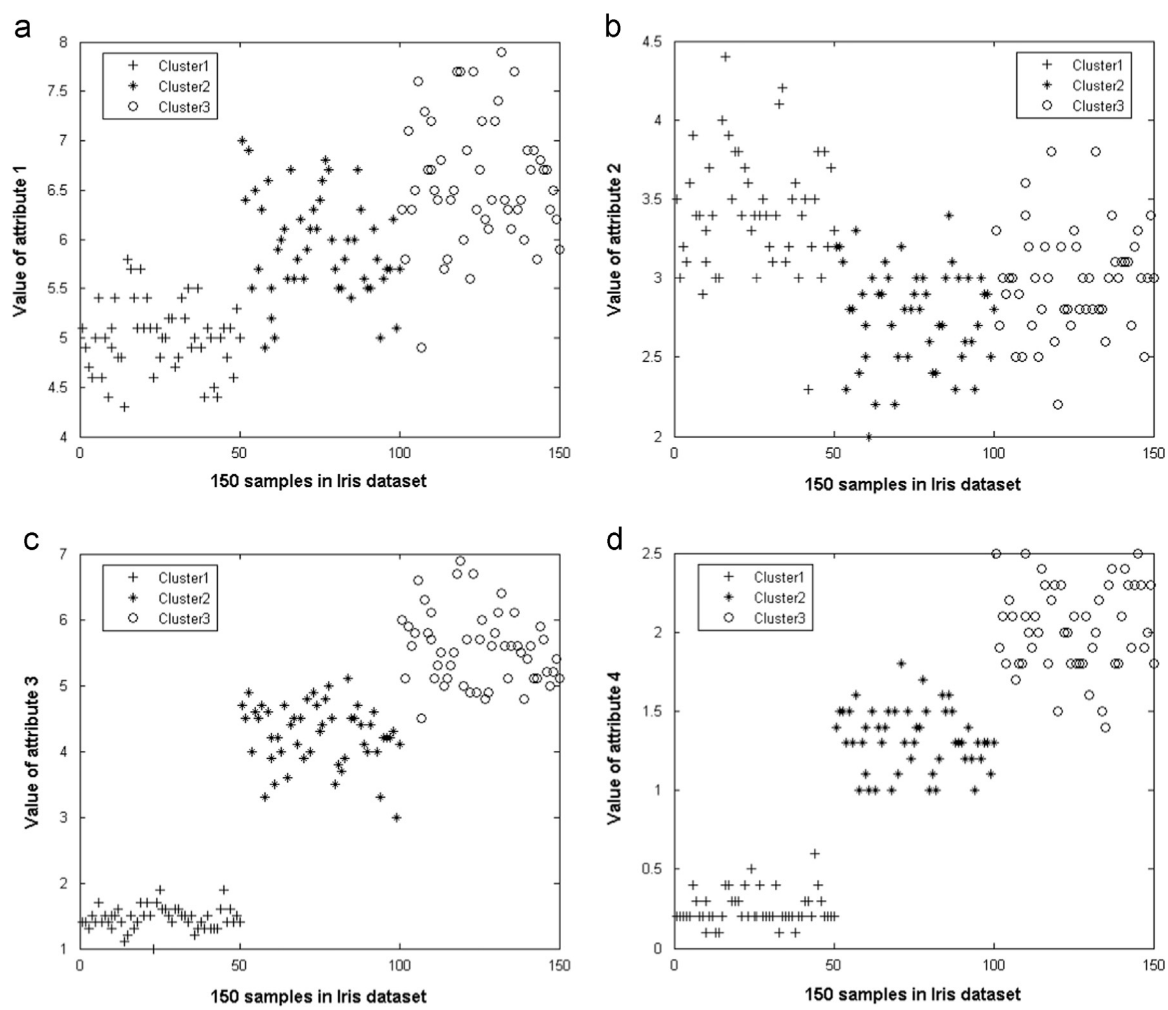

Fig. 1. Distribution of different attributes of Iris dataset. (a) Attribute 1. (b) Attribute 2. (c) Attribute 3. (d) Attribute 4.

more compact in each cluster. They should be more important and contribute much more than other two attributes in clustering, so that higher weights should be assigned to these two attributes. This can be verified by Table 2, in which attribute 3 and attribute 4 have higher weights than attribute 1 and attribute 2 for each attribute weighted clustering algorithm. Compared with the WFCM method and the WFKCA method, our EWFCM method and KEWFCM method emphasize more on the importance of these two attributes, so as to achieve higher clustering accuracy. This further verifies the physical meaning of the maximum entropy method for attribute weight assignment which is consistent with the available data (the detail is shown in the Appendix).

Another observation from Table 1 is that the kernel-based attribute-weighted algorithms (the WFKCA method and the

KEWFCM method) do not appear to provide the significant improvement in ACR and ANMI as compared to traditional attribute-weighted approaches (the WFCM method and the EWFCM method). This can be clarified by the spherical shape of the dataset in these experiments. For these ‘spherical’ data, the kernel-based clustering methods have not shown distinctive advantages.

Table 2
Attribute weight assignment of different weighted clustering algorithms on Iris dataset.

WFCM				
	Attribute 1	**Attribute 2**	**Attribute 3**	**Attribute 4**
Cluster1	0.042	0.014	**0.667**	**0.277**
Cluster2	0.103	0.132	**0.409**	**0.356**
Cluster3	0.102	0.206	**0.433**	**0.259**
EWFCM				
	Attribute 1	**Attribute 2**	**Attribute 3**	**Attribute 4**
Cluster1	0.011	0.018	**0.842**	**0.129**
Cluster2	0.019	0.030	**0.479**	**0.472**
Cluster3	0.033	0.139	**0.611**	**0.217**
KFKCA				
	Attribute 1	**Attribute 2**	**Attribute 3**	**Attribute 4**
Cluster1	0.028	0.009	**0.638**	**0.325**
Cluster2	0.076	0.081	**0.447**	**0.396**
Cluster3	0.122	0.107	**0.563**	**0.208**
KEWFCM				
	Attribute 1	**Attribute 2**	**Attribute 3**	**Attribute 4**
Cluster1	0.000	0.002	**0.856**	**0.142**
Cluster2	0.000	0.000	**0.585**	**0.415**
Cluster3	0.001	0.000	**0.677**	**0.322**

4.3. Synthetic datasets

Four synthetic datasets with ‘non-spherical’ shaped clusters are considered for further experiments [26]. Their plots have been provided in Fig. 2 with including Fuzzy ‘X’ ($N=640$, $K=2$, $M=2$), Parabolic ($N=960$, $K=2$, $M=2$), Ring ($N=750$, $K=2$, $M=2$), Zig-zag ($N=400$, $K=2$, $M=2$). The fuzzification coefficient α and the kernel function are set to be same as the UCI examples.

Table 3 illustrates the clustering results of different clustering algorithms on these four synthetic datasets. We can see that the KEWFCM algorithm is still the best method in our comparisons and demonstrate the capability for clustering the ‘non-spherical’ dataset. In addition, for most of the synthetic datasets, such as Fuzzy “X”, Ring, and Zig-zag dataset, the kernel-based clustering algorithms can obtain ideal clustering solutions in ACR and ANMI. For example, the KFCM algorithm, the KFKCA algorithm, and the KEWFCM algorithm all can achieve 100% in ACR (only 50% in ACR of the traditional approaches) on Ring dataset. All these results prove the efficiency of kernel method for the ‘non-spherical’ shaped data clustering. Similar to the results obtained in [26], the kernel-based clustering algorithms do not provide significant improvement over the traditional clustering methods on Parabolic dataset. This is possibly because of (1) selection of kernel function

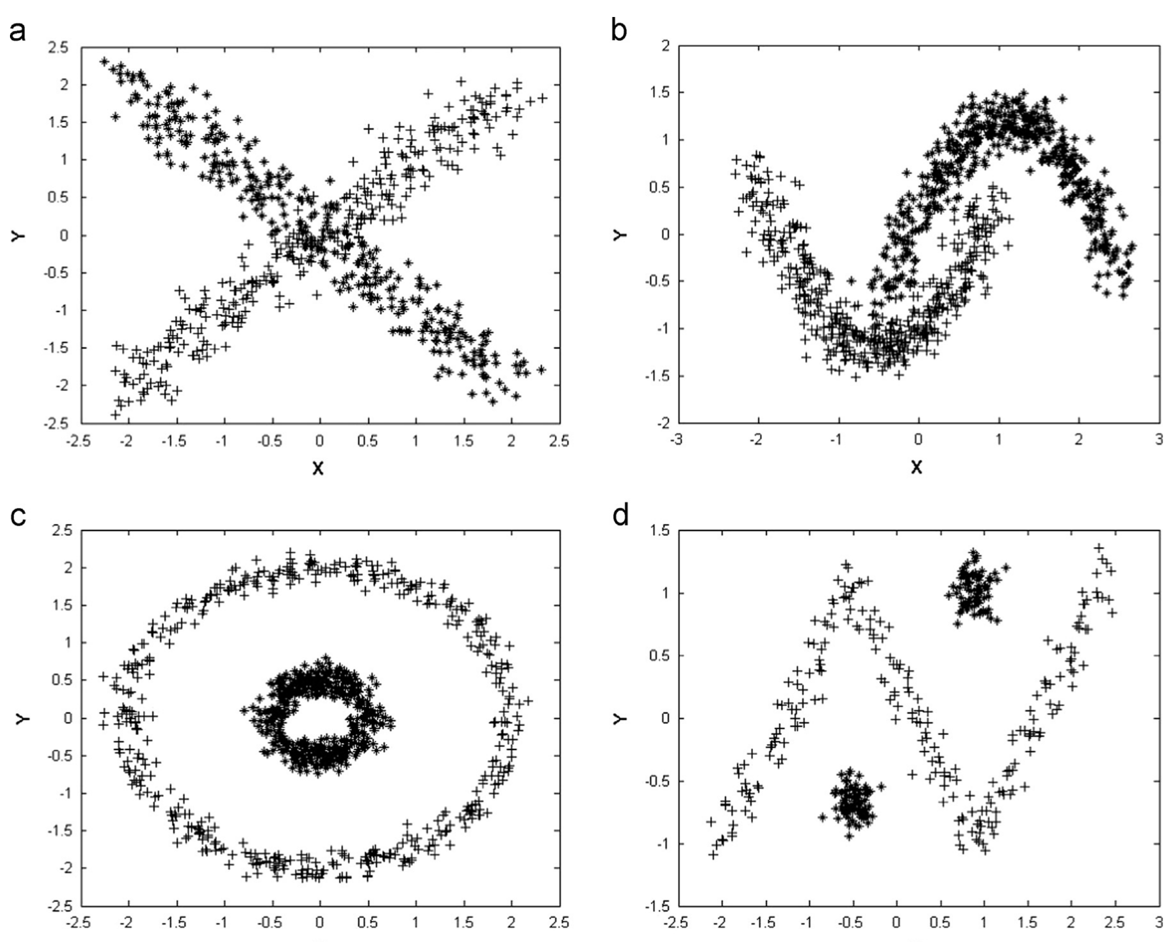

Fig. 2. Six synthetic datasets. (a) Fuzzy ‘X’. (b) Parabolic. (c) Ring. (d) Zig-zag.

Table 3
Statistics of different clustering algorithms on four synthetic datasets in terms of AIN, ACR, ANMI.

Dataset	FCM	WFCM	EWFCM	KFCM(σ^2)	KFKCA(σ^2)	KEWFCM(σ^2)
Fuzzy 'X'	43.4	55.8	**50.6**	*39.6(2^{-9})	45.1(2^{-6})	**59.2(2^{-8})**
	0.5016	0.5041	**0.5090**	0.6689	0.6691	***0.6706**
	0.0008	0.0012	**0.0034**	0.1958	0.1977	***0.2025**
Parabolic	*12.2	29.6	**36.6**	15.2(2^{-2})	14.7(2^{1})	**17.8(2^{0})**
	0.8813	0.8832	**0.8844**	0.8865	0.8879	***0.8885**
	0.4759	0.4813	**0.4845**	0.4911	0.4937	***0.4973**
Ring	40.0	51.2	**44.6**	*32.6(2^{-3})	40.7(2^{0})	**59.4(2^{-1})**
	0.5011	0.5041	**0.5208**	*1.00	*1.00	***1.00**
	0.0013	0.0015	**0.0022**	*1.00	*1.00	***1.00**
Zig-zag	*15.0	18.6	**27.2**	25.8(2^{-8})	29.1(2^{-7})	**37.4(2^{-8})**
	0.5368	0.5416	**0.5484**	0.8126	0.8197	***0.8373**
	0.0026	0.0044	**0.0050**	0.4165	0.4346	***0.4806**

*The best performance among the group.

Table 4
Summary of four test subsets.

Subsets	Topics	Words number
S1	Earn, Wheat	115
S2	Ship, Wheat	158
S3	Earn, Ship, Wheat	136
S4	Corn, Crude, Interest, Trade	281

Table 5
Statistics of different clustering algorithms on four test subsets in terms of AIN, ACR, ANMI.

Dataset	FCM	WFCM	EWFCM	KFCM(σ^2)	KFKCA(σ^2)	KEWFCM(σ^2)
S1	14.4	*12.8	**20.4**	17.4(2^{3})	19.5(2^{-6})	**25.6(2^{-8})**
	0.8455	0.8650	**0.9714**	0.9150	0.9418	***0.9750**
	0.5222	0.5621	**0.8460**	0.6039	0.7944	***0.8558**
S2	*9.6	13.6	**27.2**	25.4(2^{4})	26.0(2^{-7})	**31.8(2^{-9})**
	0.8575	0.8941	**0.9350**	0.8950	0.9197	***0.9450**
	0.4705	0.5438	**0.6770**	0.5652	0.5844	***0.7330**
S3	*11.8	17.2	**32.4**	19.6(2^{4})	25.1(2^{-7})	**39.8(2^{-9})**
	0.6080	0.7333	**0.8433**	0.6566	0.7850	***0.8674**
	0.3750	0.4338	**0.5604**	0.4057	0.4766	***0.5816**
S4	*18.2	34.8	**46.6**	28.4(2^{5})	33.8(2^{-5})	**52.8(2^{-8})**
	0.5232	0.6450	**0.7275**	0.5725	0.7135	***0.7583**
	0.3244	0.4112	**0.4992**	0.3464	0.4527	***0.5037**

*The best performance among the group.

and/or (2) optimization of kernel parameters as discussed in [26,33], which is still a good direction of our future work.

4.4. Reuters transcribed text dataset

In this example, the Reuters Transcribed text dataset is used to demonstrate the efficiency of the proposed algorithms on feature extraction. This dataset is created by selecting 200 files from the 10 largest classes in the Reuters-21578 collection, which is available from ⟨http://www.ics.uci.edu/ml/⟩ [32]. Ten clusters are labeled by the topic name, with each contains 20 files of transcriptions. These data are first preprocessed using the Bow toolkit [34] to eliminate the stop words, stem words and the words that occur in less than cluster number. Then the remaining words in each document are valued by the standard $tf \cdot idf$ as data attributes. Four test subsets are built with different topic/cluster numbers and word/attribute numbers as shown in Table 4.

Table 5 shows the clustering results of six different clustering algorithms on four test subsets. Compared with the other approaches, the EWFCM algorithm and KEWFCM algorithm achieve excellent clustering results in ACR and ANMI. Moreover, they can efficiently extract the important features for data classification. Fig. 3 illustrates all words weight computed by the EWFCM algorithm in each topic of subset S1. Evidently, most of the words with higher weight extracted by the EWFCM algorithm are closely related to the topics (The top 20 words are listed in Table 6).

5. Conclusions

This study focuses on the attribute-weighted fuzzy clustering problem and proposes a maximum-entropy-regularized weighted fuzzy c-means (EWFCM) algorithm. Based on the attribute-weight-entropy regularization, a new objective function is developed to simultaneously minimize the within cluster dispersion and maximize the attribute-weight-entropy. Thus, the optimal clustering results have been yielded and the important attributes are extracted according to the optimal distribution of attribute weight for clustering. Then the kernel-based EWFCM clustering algorithm is presented for clustering the data with 'non-spherical' shaped clusters. Compared with traditional clustering algorithms, the proposed algorithms show the superior performance either in the classification rate or in the normalized mutual information.

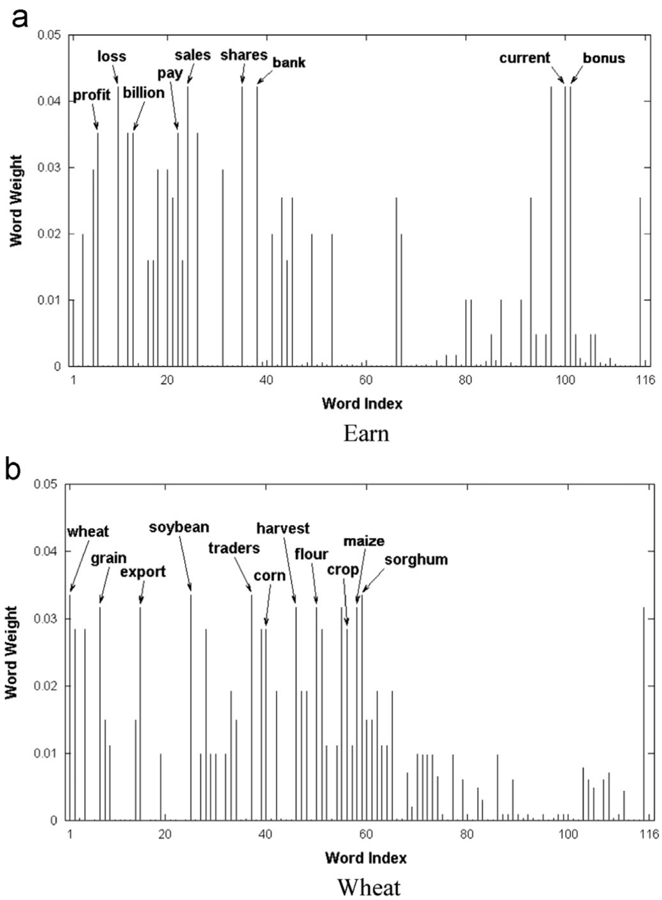

Fig. 3. Word weight obtained by EWFCM algorithm on subset S1. (a) Earn. (b) Wheat.

Table 6
Top 20 words extracted by EWFCM algorithm on subset S1.

Topic	Extracted words
Earn	**Loss**, **sales**, **shares**, **bank**, **products**, **current**, **bonus**, **profit**, statement, **billion**, **pay**, **dividend**, net, **company**, services, expects, industries, **paid**, reuter, reported
Wheat	**Wheat**, **soybean**, **traders**, **sorghum**, **export**, **flour**, **grain**, **harvest**, **maize**, season, reuter, tonnes, **agriculture**, **shipment**, stocks, **corn**, French, **crop**, March, soft

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant no.61572540, 61573166 and 61572231, the UM Multiyear Research Grants, and the Science and Technology Planning Project for University of Shandong Province under Grant no. J12LN18.

Appendix

The first, we will show that the attribute weighted fuzzy clustering problem is just a maximum entropy inference. A loss function can be defined as (20):

$$L(\boldsymbol{U},\boldsymbol{C},\boldsymbol{W}) = \sum_{k=1}^{K}\sum_{n=1}^{N} u_{kn}^{\alpha} \sum_{m=1}^{M} w_{km}(x_{nm}-c_{km})^2 \tag{20}$$

where $\boldsymbol{U}=[u_{kn}]$ is a K-by-N matrix, u_{kn} denotes the degree of membership of the n-th object belonging to the k-th cluster, $\boldsymbol{C}=[c_{km}]$ is a K-by-M matrix, c_{km} denotes cluster center of the k-th cluster defined by u_{kn}, $\boldsymbol{W}=[w_{km}]$ is a K-by-M matrix, w_{km} indicates the attribute weight of the m-th object dimension in the k-th cluster, and w_{km} satisfies the following condition:

$$\sum_{m=1}^{M} w_{km}=1, \quad 0 \le w_{km} \le 1 \tag{21}$$

Then we can see that the weighted fuzzy clustering can be considered as maximum entropy inference problem of finding an attribute weight probability assignment which maximizes the entropy $\left(-\sum_{k=1}^{K}\sum_{m=1}^{M} w_{km}\log(w_{km})\right)$, while subjecting to minimizing the loss function (20) and constraint (21). Let E is the minimum of the loss function $L(\boldsymbol{U},\boldsymbol{C},\boldsymbol{W})$, formally, this problem can be written as:

$$\max -\sum_{k=1}^{K}\sum_{m=1}^{M} w_{km}\log(w_{km}) \tag{22}$$

Subject to $\sum_{m=1}^{M} w_{km}=1$, $L(\boldsymbol{U},\boldsymbol{C},\boldsymbol{W})=E$ where $L(\boldsymbol{U},\boldsymbol{C},\boldsymbol{W})$ is given by (20) and E is a parameter.

We can use the Lagrangian multiplier technique to solve the above constrained maximization problem.

$$G(\boldsymbol{W},\gamma,\lambda) = -\sum_{k=1}^{K}\sum_{m=1}^{M} w_{km}\log(w_{km}) - \gamma(L(\boldsymbol{U},\boldsymbol{C},\boldsymbol{W})-E) - \lambda\left(\sum_{m=1}^{M} w_{km}-1\right) \tag{23}$$

where γ and λ are two Lagrange multipliers corresponding to the constraints.

By setting the gradient of $G(\boldsymbol{W},\gamma,\lambda)$ to zero with respect to γ, λ and w_{il}, we obtain

$$\frac{\partial G(\boldsymbol{W},\gamma,\lambda)}{\partial \gamma} = -(L(\boldsymbol{U},\boldsymbol{C},\boldsymbol{W})-E)=0 \tag{24}$$

$$\frac{\partial G(\boldsymbol{W},\gamma,\lambda)}{\partial \lambda} = -\left(\sum_{m=1}^{M} w_{km}-1\right)=0 \tag{25}$$

$$\frac{\partial G(\boldsymbol{W},\gamma,\lambda)}{\partial w_{km}} = -(\log\ w_{km}+1) - \gamma\sum_{n=1}^{N} u_{kn}^{\alpha}(x_{nm}-c_{km})^2 - \lambda = 0 \tag{26}$$

From (26), we obtain

$$\begin{aligned} w_{km} &= \exp\left(-\gamma\sum_{n=1}^{N} u_{kn}^{\alpha}(x_{nm}-c_{nm})^2-\lambda-1\right) \\ &= \exp\left(-\gamma\sum_{n=1}^{N} u_{kn}^{\alpha}(x_{nm}-c_{nm})^2\right)\exp(-\lambda-1) \end{aligned} \tag{27}$$

Substituting (27) into (25), we have

$$\begin{aligned} \sum_{s=1}^{M} w_{ks} &= \sum_{s=1}^{M}\exp\left(-\gamma\sum_{n=1}^{N} u_{kn}^{\alpha}(x_{ns}-c_{ks})^2\right)\exp(-\lambda-1) \\ &= \exp(-\lambda-1)\sum_{s=1}^{M}\exp\left(-\gamma\sum_{n=1}^{N} u_{kn}^{\alpha}(x_{ns}-c_{ks})^2\right) \\ &= 1 \end{aligned} \tag{28}$$

It follows that

$$\exp(-\lambda-1) = \frac{1}{\sum_{s=1}^{M}\exp\left(-\gamma\sum_{n=1}^{N} u_{kn}^{\alpha}(x_{ns}-c_{ks})^2\right)} \tag{29}$$

Substituting (29) into (27), we obtain

$$w_{km} = \frac{\exp\left(-\gamma \sum_{n=1}^{N} u_{kn}^{\alpha}(x_{nm} - c_{km})^2\right)}{\sum_{s=1}^{m} \exp\left(-\gamma \sum_{n=1}^{N} u_{kn}^{\alpha}(x_{ns} - c_{ks})^2\right)} \tag{30}$$

The maximum entropy method has clear physical meaning for attribute weight assignment, however, the parameter E is difficult to determine beforehand. At the same time, it is easily seen that the solution (30) of this problem is given by the same formula of (4). Thus, actually the EWFCM method of minimizing (1) can be considered as a regularized problem which is equivalent to the above maximum entropy approach of maximizing (22).

References

[1] J.W. Han, M. Kamber, Data Mining: Concept and Techniques, Morgan Kanfmann, San Mateo, CA, 2001.
[2] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[3] J. MacQueen, Some methods for classification and analysis of multivariate observation, in: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1, 1967, pp.281–297.
[4] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, New York, 1981.
[5] T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, A.Y. Wu, An efficient k-means clustering algorithm: analysis and implementation, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 881–892.
[6] P.F. Huang, D.Q. Zhang, Locality sensitive C-means clustering algorithms, Neurocomputing 73 (2010) 2935–2943.
[7] L. Zhu, F.L. Chung, S. Wang, Generalized fuzzy c-means clustering algorithm with improved fuzzy partitions, IEEE Trans. Syst. Man Cybern. B: Cybern. 39 (2009) 578–591.
[8] Y. Liu, F. Tian, Z. Hu, C. DeLisi, Evaluation and integration of cancer gene classifiers: identification and ranking of plausible drivers, Nat. Sci. Rep. 5 (2015) 1–15.
[9] D.S. Modha, W.S. Spangler, Feature weighting in k-means clustering, Mach. Learn. 52 (2003) 217–237.
[10] L. Parsons, E. Haque, H. Liu, Subspace clustering for high dimensional data: a review, SIGKDD Explor. 6 (2004) 90–105.
[11] W.S. Desarbo, J.D. Carroll, L.A. Clark, P.E. Green, Synthesized clustering: a method for amalgamating clustering bases with differential weighting variables, Psychometrika 49 (1984) 57–78.
[12] J.H. Friedman, J.J. Meulman, Clustering objects on subsets of attributes, J. R. Stat. Soc. 66 (2004) 815–849.
[13] L.P. Jing, M.K. Ng, J.Z. Huang, An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data, IEEE Trans. Knowl. Data Eng. 19 (2007) 1026–1041.
[14] C.Y. Tsai, C.C. Chiu, Developing a feature weight self-adjustment mechanism for a k-means clustering algorithm, Comput. Stat. Data Anal. 52 (2008) 4658–4672.
[15] R. Xu, D. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (2005) 645–678.
[16] X.Z. Wang, Y.D. Wang, L.J. Wang, Improving fuzzy c-means clustering based on feature-weight learning, Pattern Recognit. Lett. 25 (2004) 1123–1132.
[17] C. Borgelt, Feature weighting and feature selection in fuzzy clustering, Proc. IEEE Conf. Fuzzy Syst. 1 (2008) 838–844.
[18] A. Keller, F. Klawonn, Fuzzy clustering with weighting of data variables, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 8 (2000) 735–746.
[19] H. Frigui, O. Nasraoui, Unsupervised learning of prototypes and attribute weights, Pattern Recognit. 37 (2004) 567–581.
[20] Z. Deng, K.S. Choi, F.L. Chung, S. Wang, Enhanced soft subspace clustering integrating within-cluster and between-cluster information, Pattern Recognit. 43 (2010) 767–781.
[21] T.F. Ng, T.D. Pham, X. Jia, Feature interaction in subspace clustering using the Choquet integral, Pattern Recognit. 45 (2012) 2645–2660.
[22] C.L. Tang, S.G. Wang, W. Xu, New fuzzy c-means clustering model based on the data weighted approach, Data Knowl. Eng. 69 (2010) 881–900.
[23] E.T. Jaynes, Information theory and statistical mechanics, Phys. Rev. 106 (1957) 620–630.
[24] J.C. Bezdek, A convergence theorem for the fuzzy ISODATA clustering algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 2 (1980) 1–8.
[25] J.H. Chiang, P.Y. Hao, A new kernel-based fuzzy clustering approach: support vector clustering with cell growing, IEEE Trans. Fuzzy Syst. 11 (2003) 518–527.
[26] D. Graves, W. Pedrycz, Kernel-based fuzzy clustering and fuzzy clustering: a comparative experimental study, Fuzzy Sets Syst. 161 (2010) 522–543.
[27] L. Chen, C.L. Philip Chen, M. Lu, A multiple-kernel fuzzy c-means algorithm for image segmentation, IEEE Trans. Syst. Man Cybern. B: Cybern. 41 (2011) 1263–1274.
[28] R. Herbrich, Learning Kernel Classifiers, MIT Press, Cambridge, MA, 2002.
[29] K.R. Muller, S. Mika, G. Ratsch, K. Tsuda, B. Scholkopf, An introduction to kernel-based learning algorithms, IEEE Trans. Neural Netw. 12 (2001) 181–201.
[30] H. Shen, J. Yang, S. Wang, X. Liu, Attribute weighted mercer kernel based fuzzy clustering algorithm for general non-spherical datasets, Soft Comput. 10 (2006) 1061–1073.
[31] A. Strehl, J. Ghosh, Cluster ensembles — a knowledge reuse framework for combining multiple partitions, J. Mach. Learn. Res. 3 (2003) 583–617.
[32] D.J. Newman, S. Hettich, C.L. Blake, C.J. Merz, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 1987, (Online). Available⟨http://archive.ics.uci.edu/ml/⟩⟩.
[33] N.R. Pal, K. Sarkar, What and when can we gain from the kernel versions of c-means algorithm? IEEE Trans. Fuzzy Syst., Print. (2013), in print.
[34] A.K. McCallum, Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering, (Online). Available: ⟨⟨http://www.cs.cmu.edu/~mccallum/bow/⟩⟩, 1998.

Jin Zhou received the B.S. degree in computer science and technology from Shandong University, Jinan, China, in 1998, the M.S. degree in software engineering from Shandong University, Jinan, China, in 2001, and the Ph. D. degree in software engineering from the University of Macau, Macau, China, in 2014. He is currently an associate professor with the School of Information Science and Engineering, University of Jinan, Shandong, China. His current research interests include computational intelligence, machine learning techniques and their applications.

Long Chen received the M.S.E. degree from Institute of Automation, Chinese Academy of Sciences in 2003, and the Ph.D. degree in electrical engineering from the University of Texas at San Antonio, USA, in 2010. He currently is an Assistant Professor of the department of Computer and Information Science, University of Macau. His current research interests include computational intelligence, Bayesian methods, and other machine learning techniques and their applications.

C. L. Philip Chen received the M.S. degree in electrical engineering from the University of Michigan, in 1985 and the Ph.D. degree in electrical engineering from Purdue University, in 1988. He is currently a Chair Professor of the Department of Computer and Information Science and Dean of the Faculty of Science and Technology, University of Macau. He is a Fellow of the AAAS & IEEE. He is the Jr. Past President of the IEEE Systems, Man, and Cybernetics Society, and the Editor-in-Chief of IEEE Transactions on Systems, Man, and Cybernetics: Systems. He is an Accreditation Board of Engineering and Technology Education Program Evaluator for Computer Engineering, Electrical Engineering, and Software Engineering programs.

Yuan Zhang received the M.S. degree in Communication Systems and Ph.D. degree in Control Theory & Engineering both from Shandong University, China, in 2003 and 2011 respectively. He is currently an Associate Professor at University of Jinan (UJN), China. Dr. Zhang was a visiting scholar at Computer Science Department, Georgia State University, USA, in 2014. As the first author he has published more than 20 peer reviewed papers in international journals and conference proceedings, 1 book chapters, and 5 patents in the areas of wireless networking. He has served as Corresponding Guest Editor for two special issues of Springer and InderScience publications. His research interests are in wireless networking and mobile computing, currently focusing on smartphone sensing and wireless sensor networks. He is a senior member of IEEE.

Han-Xiong Li received B.E. from National University of Defence Technology, China, M.E. from Delft University of Technology, The Netherlands, and Ph.D from the University of Auckland, New Zealand. He is a professor in Department of Systems Engineering and Engineering Management, City University of Hong Kong. His current research interests are in intelligent modeling and control, integrated process design and control, intelligent learning and decision informatics. He published over 130 SCI journal papers with h-index 25. He serves as Associate Editor of IEEE Transactions on Cybernetics, and IEEE Transactions on Industrial Electronics. He is a fellow of the IEEE.