Parameter Optimization Method for Gaussian Mixture Model with Data Evolution

Yu Yuecheng (於跃成)[1,2*], *Sheng Jiagen* (生佳根)[1], *Zou Xiaohua* (邹晓华)[1]

1. College of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang, 212003, China;
2. Information Technology Research Base of Civil Aviation Administration of China,
Civil Aviation University of China, Tianjin, 300300, China

(Received 21 September 2013; revised 26 November 2013; accepted 5 December 2013)

Abstract: To learn from evolutionary experimental data points effectively, an evolutionary Gaussian mixture model based on constraint consistency (EGMM) is proposed and the corresponding method of parameter optimization is presented. Here, the Gaussian mixture model (GMM) is adopted to describe the data points, and the differences between the posterior probabilities of pairwise points under the current parameters are introduced to measure the temporal smoothness. Then, parameter optimization of EGMM can be realized by evolutionary clustering. Compared with most of the existing data analysis methods by evolutionary clustering, both the whole features and individual differences of data points are considered in the clustering framework of EGMM. It decreases the algorithm sensitivity to noises and increases the robustness of evaluated parameters. Experimental result shows that the clustering sequence really reflects the shift of data distribution, and the proposed algorithm can provide better clustering quality and temporal smoothness.

Key words: evolutionary clustering; evolutionary Gaussian mixture model; temporal smoothness; parameter optimization

CLC number: TP391 **Document code**: A **Article ID**: 1005-1120(2014)04-0394-11

1 Introduction

Clustering analysis is one of the most important methods of data prediction and data analysis in the field of machine learning and data mining[1-2]. According to measured or perceived intrinsic characteristics, the data set is partitioned into several clusters by the clustering algorithm, where the data points from the same cluster should be similar as much as possible, and the data points from the different clusters should be dissimilar[3]. In general, data clustering has been used for the following three main purposes: disclosing the underlying structure of data set, realizing the natural classification of data points, and organizing the data or summarizing the data through cluster prototypes[4]. Essentially, the above applications are used to find the potential patterns or data structure by clustering.

The conventional clustering algorithms focus on static data set, and assume that all data obey an underlying distribution which will not evolve along time[5]. However, there are some that applications, such as dynamic social network[6-8], blog communities[9] and moving objects tracking[10], where the size of dataset or the data distribution may drift along time due to concept drifting or noise varying[5,11]. In this case, the potential patterns implicated in dynamic data set cannot

Foundation items: Supported by the National Natural Science Foundation of China (61202137); the Open Project Foundation of Information Technology Research Base of Civil Aviation Administration of China (CAAC-ITRB-201302); the University Natural Science Basic Research Project of Jiangsu Province (13KJB520004); and the Fundamental Research Funds for the Central Universities (NS2012134).

[*] **Corresponding author**: Yu Yuecheng, Associate Professor, E-mail: yuyuecheng@nuaa.edu.cn.

be accurately analyzed by using conventional clustering algorithms.

Evolutionary clustering is an extended application of clustering analysis. Its main purpose is to disclose the inherent patterns of data set and the evolutionary characteristic of evolving datasets. The existing evolutionary clustering algorithms typically outperform conventional static clustering algorithms by adding a temporal smoothness penalty to the cost function[12]. The result of evolutionary clustering is a sequence of clustering in a time-line. For the clustering result produced during a particular timestamp, two criteria including high snapshot quality and low history cost, should be met. This means that the sequence of clustering should have high-quality clustering at the current timestamp, and meet temporal smoothness in the successive timestamp[13]. Now, several frameworks of evolutionary clustering have been proposed[4,11-15]. Evolutionary *k*-means, proposed by Chakrabarti, et al is the first framework of evolutionary clustering. This framework adopts the objective function of *k*-means as the function of snapshot quality and the differences between all pairs of centroids in successive timestamp as the penalty of history cost. Its main weak point is that the stability of clustering cannot be guaranteed. As a result, the small perturbation on the centroids may cause drastic changes of clusters[11]. Based on spectral clustering, Yun Chi, et al[11] proposed the other two frameworks of evolutionary spectral clustering, namely, preserving cluster quality (PCQ) and preserving cluster membership (PCM), where two different measure strategies of temporal smoothness were integrated in the overall measure of clustering quality. In the PCQ framework, the current partition is applied to historic data and the resulting cluster quality determine the temporal smoothness cost. In the PCM framework, the current partition is directly compared with the historic partitions and the resulting differences determine the temporal smoothness cost. From the measure strategy of history cost, PCQ is similar to evolutionary *k*-means, but PCM adopts the similar idea of constrained clustering[11].

Penalized likelihood is often used to learn the parameters of Gaussian mixture model (GMM) and has been used in regression analysis, classification and clustering etc. In the field of machine learning, conditional entropy and Kullback-Leibler divergence are the main measure strategies used to evaluate the differences of information and distributions[16]. By introducing manifold learning into parameters evaluation of GMM, Laplacian regularized Gaussian mixture model (LapGMM) was proposed by He, et al[17]. According to the idea of LapGMM, the nearby data points along the geodesics on the manifold have the similar conditional probability density functions. Based on the assumption, Laplacian regularization is defined to penalize the likelihood function and the conditional probability distribution of the nearby data points can be smoothened. Based on the similar idea, locally consistent Gaussian mixture model (LCGMM) was proposed by Liu, et al[18], where locally consistent assumption was adopted.

To most of the existing evolutionary clustering algorithms, the overall properties of data points between the corresponding clusters in successive timestamps are integrated into the definition of history cost function. This is beneficial to avoid the obvious fluctuation from the noise data points[13-15]. However, the individual differences between data points are ignored. In fact, the different data points may have different effects on the evolutionary clustering when the data distributions drift. Inspired by LapGMM and LCGMM, evolutionary Gaussian mixture model based on constraint consistency (EGMM) is proposed from the point of view of constrained clustering. Using GMM as the model of data description, the snapshot quality function of EGMM is defined by means of the log-likelihood of complete data set, and the history cost function of EGMM is defined according to the differences between the posterior distributions. These distributions describe the statistic characteristics of all pairwise data points, which have the same cluster labels in the previous clustering results.

2 Notations and Related Work

The research objective of evolutionary clustering is to cluster the dynamic datasets, whereas evolutionary clustering also relates to the other research fields, such as data stream clustering, incremental clustering and constrained clustering[13,19]. In this section, evolutionary k-means and evolutionary spectral clustering will be introduced based on the unified forms of notation representation. Evolutionary k-means is the first framework of evolutionary clustering, and the relation between evolutionary and constrained clustering has been firstly described in the framework of evolutionary spectral clustering.

2.1 Definition of notations

To obtain the smooth clustering sequence, the clustering results and data distributions at $t-1$ timestamp should be integrated into the clustering process at t timestamp. So, the subscripts, namely "t" and "$t-1$", represent the corresponding information at t and $t-1$ timstamps, respectively. Let $X_t=\{x_{t,i}\}_{i=1}^{n}$ denote the data set at t timestamp, where n represents the number of data points. Correspondingly, let X_{t-1} denote the data set at $t-1$ timestamp. The current data set consists of $X_{t(\text{new})}$ and $X_{t(\text{old})}$, namely $X_t = X_{t(\text{new})} \cup X_{t(\text{old})}$, where $X_{t(\text{new})}$ denotes the newly increased data set at t timestamp and $X_{t(\text{old})}$ denotes the joint data set at t and $t-1$ timestamps. As the subset of X_t, $X_{t(\text{old})}$ may include all or part of X_{t-1}, namely $X_{t(\text{old})} \subseteq X_{t-1}$. This includes the following two cases. If $X_{t(\text{old})} = X_{t-1}$, it means that all data points at $t-1$ timestamp will appear at t timestamp. Otherwise, only part of data points at $t-1$ timestamp will appear at t timestamp, namely $X_{t(\text{old})} \subseteq X_{t-1}$. Among them, the former can be regarded as a special form of the latter case.

Assume that K_t represents the number of cluster, and $MC_t=\{\mu_{t,k}\}_{k=1}^{K_t}$ the set of centroids, where $\mu_{t,k}$ is the centroid of the kth cluster and the subscript "t" represents the t timestamp. We also use $C_{t,k}$ to represent the data set of the kth cluster at t timestamp. Correspondingly, we assume that $MC_{t-1}=\{\mu_{t-1,k'}\}_{k'=1}^{K_{t-1}}$ represents the set of centroids at $t-1$ timestamp, where $\mu_{t-1,k'}$ denotes the centroid of the k'th cluster and K_{t-1} represents the number of cluster at $t-1$ timestamp. The notation of $C_{t-1,k'}$ is used to represent the data set of the k' th cluster at the $t-1$ timestamp.

2.2 Evolutionary k-means

As the first framework of evolutionary clustering, the object of evolutionary k-means is evolving data points. Its goal is to obtain a smooth clustering sequence. In the framework, the clustering differences between adjacent time-stamps are measured by history cost function, denoted as $\text{hc}(MC_t, MC_{t-1})$, and the clustering quality of current data points is measured by the function of snapshot quality, denoted as $\text{sq}(MC_t)$. By minimizing the objective function Eq. (1), the clustering sequence with the optimal clustering quality can be guaranteed by performing the algorithm of evolutionary k-means[13].

$$\min \sum_{t=1}^{T} \text{sq}(MC_t) - \lambda \cdot \sum_{t=2}^{T} \text{hc}(MC_t, MC_{t-1}) \quad (1)$$

where λ is the tradeoff parameter of clustering quality and history cost defined by user.

On the online setting, the objective function of evolution k-means can be described

$$J_{\text{EKM}} = \min \text{sq}(MC_t) - \lambda \cdot \text{hc}(MC_t, MC_{t-1}) \quad (2)$$

where the objective function of standard k-means is the snapshot quality function[13], namely

$$\text{sq}(MC_t) = \min \sum_{k=1}^{K_t} \sum_{x_{t,i} \in C_{t,k}} \| \boldsymbol{x}_{t,i} - \boldsymbol{\mu}_{t,k} \|^2 \quad (3)$$

In the framework of evolutionary k-means, history cost is measured by the distance sum of the pair of centroids, where the most similar two clusters are respectively from the adjacent timestamps and the distance between the two centroids is the shortest. Let f: $MC_t \rightarrow MC_{t-1}$ be the mapping defined in the set of centroids. These centroids are either from t or $t-1$ timestamp.

Then, $\boldsymbol{\mu}_{t-1,f(k)} \in MC_{t-1}$ denotes the centroid that is the nearest centroid at the $t-1$ timestamp from $\boldsymbol{\mu}_{t,k}$. That is to say, for all $\boldsymbol{\mu}_{t-1,a} \in MC_{t-1}$, the equation of $d(\boldsymbol{\mu}_{t,k}, \boldsymbol{\mu}_{t-1,f(k)}) = \min d(\boldsymbol{\mu}_{t,k}, \boldsymbol{\mu}_{t-1,a})$ is true, where the notation of $d(\cdot)$ represents the Euclidean distance between two clusters. So, the pair of clusters, namely $C_{t-1,f(k)}$ and $C_{t,k}$, are the most similar clusters. The history cost function of evolutionary k-means can be defined as Eq. (4)[13]

$$\mathrm{hc}(MC_{t-1}, MC_t) = \min_{f: MC_t \to MC_{t-1}} \sum_{k=1}^{K_t} \| \boldsymbol{\mu}_{t,k} - \boldsymbol{\mu}_{t-1,f(k)} \|^2 \tag{4}$$

The algorithm of evolutionary k-means runs in an iterative manner adopted by standard k-means. Eq. (5) is the updating formula of centroids. Its essence is to adjust the position of cluster centers at current timestamp using the historic centroids, where the corresponding historic cluster is the most similar to the current cluster. Then, the cluster centers at t timestamp will lie in between the centroids suggested by standard k-means and its closest match from the previous timestamp. With this mechanism, the better clustering quality and smooth clustering sequence can be guaranteed.

$$\boldsymbol{\mu}_{t,k} = (1-\gamma) \cdot \lambda \cdot \boldsymbol{\mu}_{t-1,f(k)} + \gamma \cdot (1-\lambda) \sum_{\boldsymbol{x}_{t,i} \in C_{t,k}} \boldsymbol{x}_{t,i} \tag{5}$$

where $\gamma = |C_{t,k}| / (|C_{t,k}| + |C_{t-1,f(k)}|)$, $|C_{t,k}|$ denotes the data number in $C_{t,k}$ and $|C_{t-1,f(k)}|$ denotes the data number in $C_{t-1,f(k)}$.

History cost in evolutionary k-means is measured by the distance between each of the pair of centroids at t timestamp to its nearest peer at $t-1$ timestamp. The sum of distance between all pairs of centroids is smaller, the differences of clustering results between adjacent timestamps are smaller too. Such a strategy of history cost has three weak points. First, the cluster number must be the same at adjacent timestamps. However, the dynamic data set will drift along time, this may cause the change of distributions. With in-depth understanding of the user, the concept may evolve too. These situations may result in changes of cluster number. Second, history cost is measured in the distance between the corresponding centroids. When the position of centroid changes slightly, the history cost may change drastically. This will affect the stability of clustering results[5,11,19]. In addition, only the overall differences of data points at adjacent timestamps are considered, but the different effects from individual data points are ignored during the process of evolutionary clustering.

2.3 Evolutionary spectral clustering

Since the clustering results of evolutionary k-means is not stable, two frameworks of evolutionary spectral clustering, namely PCQ and PCM, were proposed by Chi, et al[11]. To measure temporal smoothness, two different fitting methods are adopted, including the one indicating how the current clustering results are consistent with the characteristics of historic data points or the historic clustering results[11,19].

Assume that $V = \{v_i\}_{i=1}^n$ represents the set of n vertices, $\boldsymbol{W}$ is the matrix used to describe the similarity of all vertices in V, and $\{C_k\}_{k=1}^K$ represents a partition of the vertices in V, where K denotes the number of clusters. The partition of the vertices in V can be represented as a n-by-K matrix $\boldsymbol{Z}$ whose elements are in $\{0, 1\}$, where $Z(i,j)=1$ ($i=1,\cdots,n$; $k=1,\cdots,K$) if only if vertex v_i is partitioned to cluster C_k. Because matrix $\boldsymbol{Z}$ is an orthogonal matrix, matrix $\boldsymbol{Z}$ can be normalized in the following way: matrix $\hat{\boldsymbol{Z}}$ can be obtained by dividing the kth column of $\boldsymbol{Z}$, where $|C_k|$ represents the number of vertices in cluster C_k. On this basis, the negated average association, denoted as NA, can be defined as

$$NA = \mathrm{tr}(\boldsymbol{W}) - \mathrm{tr}(\hat{\boldsymbol{Z}}^{\mathrm{T}} \boldsymbol{W} \hat{\boldsymbol{Z}}) \tag{6}$$

where $\mathrm{tr}(\cdot)$ represents the trace of matrix. Assume that $\boldsymbol{W}_t$ and $\boldsymbol{W}_{t-1}$ represent the vertex similarity matrix at t and $t-1$ timestamps, respectively, then the total cost function of PCQ can be defined as Eq. (7)[19]

$$J_{\mathrm{PCQ}} = \lambda(\mathrm{tr}(\boldsymbol{W}_t) - \mathrm{tr}(\hat{\boldsymbol{Z}}_t^{\mathrm{T}} \boldsymbol{W}_t \hat{\boldsymbol{Z}}_t)) + (1-\lambda)(\mathrm{tr}(\boldsymbol{W}_{t-1}) - \mathrm{tr}(\hat{\boldsymbol{Z}}_t^{\mathrm{T}} \boldsymbol{W}_{t-1} \hat{\boldsymbol{Z}}_t)) =$$

$$\mathrm{tr}(\lambda \boldsymbol{W}_t + (1-\lambda)\boldsymbol{W}_{t-1}) - \mathrm{tr}(\hat{\boldsymbol{Z}}_t^{\mathrm{T}}(\lambda \boldsymbol{W}_t + (1-\lambda)\boldsymbol{W}_{t-1})\hat{\boldsymbol{Z}}_t) \quad (7)$$

where λ $(0 \leqslant \lambda \leqslant 1)$ is the tradeoff parameter of clustering quality and history cost defined by user, and $\hat{\boldsymbol{Z}}_t$ is the representation matrix of the partition at t timestamp.

For the objective function of PCQ, the first term, namely $\mathrm{tr}(\lambda \boldsymbol{W}_t + (1-\lambda)\ \boldsymbol{W}_{t-1})$, is a constant, which is independent of the clustering partitions, minimizing Eq. (7) is equivalent to maximizing the second term of the objective function. Directly solving the objective function of PCQ is an NP-hard problem. Similar to most spectral clustering methods, one solution is to relax matrix $\hat{\boldsymbol{Z}}_t$ to projection matrix $\boldsymbol{X}_t \in \mathbf{R}^{n \times k}$, where the element values of $\hat{\boldsymbol{Z}}_t$ are discrete and the element values of $\boldsymbol{X}_t$ are continuous. Then, the problem of maximizing matrix trace is converted into solving the K eigenvectors associated with the top-K eigenvalues of matrix $\lambda \boldsymbol{W}_t + (1-\lambda)\ \boldsymbol{W}_{t-1}$. In the framework of PCQ, the current clustering result is used to fit the historic data points to guarantee temporal smoothness, so the vertex similarity matrixes of the current time-stamp and the previous timestamp are contained in the objective function simultaneously. The algorithm steps of PCQ is similar to the general spectral clustering, except that the weighted sum of matrixes $\boldsymbol{W}_t$ and $\boldsymbol{W}_{t-1}$ must be calculated when using eigenvectors to construct the projection space of data points.

Unlike the PCQ model, PCM uses the differences of partitions between adjacent timestamps to guarantee temporal smoothness. To achieve this goal, the representation matrixes, including $\hat{\boldsymbol{Z}}_t$ and $\hat{\boldsymbol{Z}}_{t-1}$, are relaxed to two projection matrixes $\boldsymbol{X}_t \in \mathbf{R}^{n \times k}$ and $\boldsymbol{X}_{t-1} \in \mathbf{R}^{n \times k}$, respectively, whose elements are continuous. Then, the differences between two projection matrixes are norm.

$$\mathrm{dist}(\boldsymbol{X}_t, \boldsymbol{X}_{t-1}) = \frac{1}{2} \| \boldsymbol{X}_t \boldsymbol{X}_t^{\mathrm{T}} - \boldsymbol{X}_{t-1} \boldsymbol{X}_{t-1}^{\mathrm{T}} \|_2 \quad (8)$$

Furthermore, the total cost function of PCM can be defined as Eq. (9)[19]

$$J_{\mathrm{PCM}} = \lambda \cdot (\mathrm{tr}(\boldsymbol{W}_t) - \mathrm{tr}(\boldsymbol{X}_t^{\mathrm{T}} \boldsymbol{W}_t \boldsymbol{X}_t)) + \frac{1}{2}(1-\lambda) \cdot \| \boldsymbol{X}_t \boldsymbol{X}_t^{\mathrm{T}} - \boldsymbol{X}_{t-1} \boldsymbol{X}_{t-1}^{\mathrm{T}} \|_2 = \lambda \cdot \mathrm{tr}(\boldsymbol{W}_t) + (1-\lambda) \cdot K - \mathrm{tr}(\boldsymbol{X}_t^{\mathrm{T}}(\lambda \boldsymbol{W}_t + (1-\lambda)\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^{\mathrm{T}})\boldsymbol{X}_t) \quad (9)$$

Similar to the PCQ model, PCM has the same steps of algorithm with the general spectral clustering, but the different matrixes are used during the process of constructing projection space by calculating the eigenvectors. By solving the K eigenvectors associated with the top-K eigenvalues of matrix $\lambda \boldsymbol{W}_t + (1-\lambda)\ \boldsymbol{X}_{t-1}\ \boldsymbol{X}_{t-1}{}^{\mathrm{T}}$ to construct a vector space, the data points are projected into the space and evolutionary spectral clustering can be realized. Because the partition differences between adjacent timestamps need be compared, the projection matrix at the t timestamp includes the partition information from the projection matrix at the $t-1$ timestamp in the framework of PCM.

3 Model of EGMM

3.1 Gaussian mixture model

To effectively deal with the evolving data set generated from independent and identically distributed (IID) samples from one underlying distribution, we can assume that all data points at each timestamp are from special distributions of GMM. Thus, from the perspective of model-based clustering, evolutionary clustering becomes a mixture-density parameter estimation problem of dynamic GMM. We assume that the data set $X_t = \{\boldsymbol{x}_{t,i}\}_{i=1}^{n}$ is generated from the parametric GMM model, where n is the data number of X_t. The probabilistic model can be defined as Eq. (10)[16]

$$p(\boldsymbol{x}_{t,i} \mid \Theta_t) = \sum_{k=1}^{K_t} \pi_{t,k} p_{t,k}(\boldsymbol{x}_{t,i} \mid \theta_{t,k})$$
$$\text{s.t.} \quad \pi_{t,k} \geqslant 0;\ k = 1, \cdots, K_t \quad \text{and} \quad \sum_{k=1}^{K_t} \pi_{t,k} = 1 \quad (10)$$

where $\boldsymbol{x}_{t,i}$ denotes the i th data point at the t timestamp, $\Theta_t = \{\pi_{t,1}, \cdots, \pi_{t,K_t}, \theta_{t,1}, \cdots, \theta_{t,K_t}\}$ the parameter set of GMM, K_t the expected number of clusters at t timestamp, as well as the number of

Gaussian components of GMM, $\pi_{t,k}$ the priori probability of data points which represents the weight of the kth Gaussian component; and $p_{t,k}(\boldsymbol{x}_{t,i}|\theta_{t,k})$ a probability density function parameterized by parameter $\theta_{t,k}=\{\boldsymbol{\mu}_{t,k},\boldsymbol{\Sigma}_{t,k})\}$, here, $\boldsymbol{\mu}_{t,k},\boldsymbol{\Sigma}_{t,k}$ represent the mean and covariance matrixes of the corresponding Gaussian component, respectively.

In evolutionary clustering scenario, the data point at the t timestamp should appear with the greatest probability in the current model from the perspective of model-based clustering. This method can guarantee the better clustering quality to be obtained. Simultaneously, from the perspective of constrained clustering, pair of data points, generated from the same Gaussian component at the $t-1$ timestamp, should be partitioned into the same cluster with the greatest probability under the current parameters of GMM. This strategy will produce a smooth clustering sequence. In EGMM, better clustering quality can be obtained through the maximization likelihood method, and smaller history cost can be guaranteed depending on the adjustment of regularization operator. Regarding the clustering results at the $t-1$ timestamp as prior, the regularization operator can be defined to adjust the parameters estimations of EGMM at the t timestamp. As a result, the pair of data points meeting a prior constraint can be assigned the same cluster with a greater probability.

3.2 Snapshot quality function

Assume that $\boldsymbol{Y}_t=\{y_{t,i}\}_{i=1}^{n}$ is the set of unobserved data, whose values inform us which component density "generated" each data in $\boldsymbol{X}_t$. That is to say, $y_{t,i}\in\{1,\cdots,K_t\}$ represents that each data points $x_{t,i}$ is generated from the $y_{t,i}$th Gaussian component. Thus, log-likelihood function of complete data set, namely log $(L(\Theta_t\,|\,\boldsymbol{X}_t,\boldsymbol{Y}_t))$, can be defined as[20]

$$\log(L(\Theta_t \mid \boldsymbol{X}_t,\boldsymbol{Y}_t))=\log(P(\boldsymbol{X}_t,\boldsymbol{Y}_t \mid \Theta_t))=\sum_{i=1}^{n}(\log\pi_{t,k}p_{t,k}(\boldsymbol{x}_{t,i} \mid \theta_{t,k})) \qquad (11)$$

Essentially, the expectation of log-likelihood should be maximized to realize the evolutionary clustering in EGMM. The clustering quality at the t timestamp depends on the values of parameters estimations by maximizing the log-likelihood. Therefore, in Eq. (12) the expectation of $\log(L(\Theta_t\mid\boldsymbol{X}_t,\boldsymbol{Y}_t))$ is the function of snapshot quality of EGMM, where Θ_t^g represents the current parameters of EGMM at the t timestamp and $\theta_{t,k}^g$ the parameters of the kth Gaussian component at current iteration step. Correspondingly, the notations of $\pi_{t,k}$ and $\theta_{t,k}$ represent the parameters of the kth Gaussian component, which will be updated at current iteration step.

$$\mathrm{sq}_{\mathrm{EGMM}}(\Theta_t)=E(\log(P(\boldsymbol{X}_t,\boldsymbol{Y}_t \mid \Theta_t)) \mid \boldsymbol{X}_t,\Theta_t^g)=\sum_{i=1}^{n}\sum_{k=1}^{K_t}(\log\pi_{t,k}+\log p_{t,k}(\boldsymbol{x}_{t,i}|\theta_{t,k}))\cdot p(k \mid \boldsymbol{x}_{t,i},\theta_{t,k}^g) \qquad (12)$$

3.3 History cost function

Assume that $Z_t(\boldsymbol{x}_{t,i})$ and $Z_{t-1}(\boldsymbol{x}_{t,i})$ represent the cluster assignment at t and $t-1$ timestamps respectively. If $Z_{t-1}(\boldsymbol{x}_{t,i})=Z_{t-1}(\boldsymbol{x}_{t,j})$ is true, we can say that pairwise data points, denoted as $(\boldsymbol{x}_{t,i},\ \boldsymbol{x}_{t,j})$, are generated from the same Gaussian distribution. According to the idea of constrained clustering, all pairwise data points can be regard as a prior knowledge at the t timestamp to smoothen the clustering. Such knowledge shows that pairwise data points, namely $(\boldsymbol{x}_{t,i},\ \boldsymbol{x}_{t,j})$, may meet the constraints of "must link". This means that for each $\boldsymbol{x}_{t,i},\ \boldsymbol{x}_{t,j}\in\boldsymbol{X}_{t(\mathrm{old})}$, pairwise data points $(\boldsymbol{x}_{t,i},\boldsymbol{x}_{t,j})$ should satisfy the constraint of "must link" if $\boldsymbol{x}_{t,i},\ \boldsymbol{x}_{t,j}\in C_{t-1,k'}$ is true. Thus, $C_{t-1,k'}$ becomes the data set composed of pairwise data points. Since the equation $\boldsymbol{X}_{t(\mathrm{old})}=C_{t-1,k'}$ is true, the other equation $M_t=\boldsymbol{X}_{t(\mathrm{old})}$ will be true, where M_t is the constraints set of "must link".

Assume that evolutionary data set obeys the IID assumption, and each data point of $\boldsymbol{X}_t$ is sampled according to the marginal distribution of Gaussian component. Since the correlation between marginal distribution and conditional distribution is based on the assumption of cluster consistency, if pairwise data points $(\boldsymbol{x}_{t,i},\ \boldsymbol{x}_{t,j})$ meet the constraint of "must link", the corresponding conditional distribution of $\boldsymbol{x}_{t,i}$ and $\boldsymbol{x}_{t,j}$, i. e. , $p(k|\boldsymbol{x}_{t,i},\Theta_t^g)$ and $p(k|\boldsymbol{x}_{t,j},\Theta_t^g)$, should be a great similarity[21]. For simplicity, we use $p_{t,i}(k)$ and $p_{t,j}(k)$ as the substitutes for $p(k \mid \boldsymbol{x}_{t,i},\Theta_t^g)$ and $p(k|\boldsymbol{x}_{t,j},\Theta_t^g)$, respectively. To measure the difference between $p_{t,i}(k)$ and $p_{t,j}(k)$, we use $\hat{D}(p_{t,i}(k) \parallel p_{t,j}(k))$ defined in Eq. (13) as a substitute for Kullback-Leibler (KL) divergence[18]

$$\hat{D}(p_{t,i}(k) \parallel p_{t,j}(k)) = \frac{1}{2}(D(p_{t,i}(k) \parallel p_{t,j}(k)) + D(p_{t,j}(k) \parallel p_{t,i}(k))) \tag{13}$$

In Eq. (14), $D(p_{t,i}(k) \parallel p_{t,j}(k))$ represents the KL divergence about $p_{t,i}(k)$ and $p_{t,j}(k)$[16]

$$D(p_{t,i}(k) \parallel p_{t,j}(k)) = \sum_k p_{t,i}(k)\log\frac{p_{t,i}(k)}{p_{t,j}(k)} \tag{14}$$

To smoothen the results of evolutionary clustering, pairwise data points that have the same cluster labels at the $t-1$ timestamp should be partitioned into the same group as much as possible. To realize the goal, $\hat{D}(p_{t,i}(k) \parallel p_{t,j}(k))$ defined in Eq. (13), is used to measure the differences between posterior distributions of pairwise data points during the process of evaluating parameters of EGMM. Intuitively, if $(\boldsymbol{x}_{t,i}, \boldsymbol{x}_{t,j}) \in M_t$, the value of $\hat{D}(p_{t,i}(k) \parallel p_{t,j}(k))$ should be smaller, which can smoothen the corresponding posterior distributions and increase the probability that $\boldsymbol{x}_{t,i}$ and $\boldsymbol{x}_{t,j}$ are assigned to the same Gaussian component at the t timestamp. Thus, the cluster labels of all pairwise data points in M_t can be used as constraint information. Then we define the function of history cost in Eq. (15). It minimizes the differences of posterior distributions of pairwise data points in M_t, and leads to the end that constraint violation occurs with a minimal probability in the estimation model at the t timestamp.

$$\mathrm{hc}_{\mathrm{EGMM}}(MC_{t-1}, MC_t) = \sum_{(\boldsymbol{x}_{t,i},\boldsymbol{x}_{t,j})\in M_t}\sum_{k=1}^{K_t}\hat{D}(p_{t,i}(k) \parallel p_{t,j}(k)) = \frac{1}{2}\sum_{(\boldsymbol{x}_{t,i},\boldsymbol{x}_{t,j})\in M_t}\sum_{k=1}^{K_t}(D(p_{t,i}(k) \parallel p_{t,j}(k)) + D(p_{t,j}(k) \parallel p_{t,i}(k))) \tag{15}$$

3.4 Objective function

According to Eqs. (12, 15), the objective function of EGMM can be defined as

$$J_{\mathrm{EGMM}} = \mathrm{sq}_{\mathrm{EGMM}}(\Theta_t) - \lambda \cdot \mathrm{hc}_{\mathrm{EGMM}}(MC_{t-1}, MC_t) = \sum_{i=1}^{n}\sum_{k=1}^{K_t}\Big(\log\pi_{t,k} + \log p_{t,k}(\boldsymbol{x}_{t,i} \mid \theta_{t,k})\Big) \cdot p(k \mid \boldsymbol{x}_{t,i}, \theta_{t,k}^g) - \frac{\lambda}{2} \cdot \sum_{(\boldsymbol{x}_{t,i},\boldsymbol{x}_{t,j})\in M_t}\sum_{k=1}^{K_t}\Big(D(p_{t,i}(k) \parallel p_{t,j}(k)) + D(p_{t,j}(k) \parallel p_{t,i}(k))\Big) \tag{16}$$

As an alternative form of KL divergence, $\hat{\boldsymbol{D}}(p_{t,i}(k) \parallel p_{t,j}(k))$ is nonnegative too. Maximizing the objective function is equivalent to performing the following two operations simultaneously, including maximizing the expectation of log-likelihood of the complete data set at the t timestamp and minimizing the differences between posterior distributions of pairwise data points in M_t. Thus, the clustering quality can be guaranteed by using the model of GMM to fit the current data points. Moreover, using the information of cluster labels from the previous clustering results, the clustering results between adjacent timestamps can be smoothened.

3.5 Model parameter fitting

The objective function of EGMM is a combined optimization function. Its maximization can be iteratively solved using the framework of expectation maximization (EM) algorithm. Similar to the standard EM algorithm, the posterior probability of $\boldsymbol{x}_{t,i}$ generated from the kth Gaussian component, denoted as $p_{t,k}(k \mid \boldsymbol{x}_{t,i}, \Theta_t^g)$, will be calculated according to the Eq. (17) in the E step. In the M step, Eq. (16) will be maximized, and the two goals, including maximizing the expectation of log-likelihood and minimizing the difference of posterior distributions, will be performed simultaneously.

$$p_{t,k}(k \mid \boldsymbol{x}_{t,i}, \Theta_t^g) = \frac{\pi_{t,k}^g p_{t,k}(\boldsymbol{x}_{t,i} \mid \theta_{t,k}^g)}{\sum_{r=1}^{K_t}\pi_{t,r}^g p_{t,r}(\boldsymbol{x}_{t,i} \mid \theta_{t,r}^g)} \tag{17}$$

To maximize Eq. (16) in the M step, it needs to take the partial derivative of the combined function with respect to each parameter, and set it to be zero, then the iterative equations used to estimate the parameters can be obtained. The equations from Eq. (18) to Eq. (20) are the corresponding updating equations of $\pi_{t,k}$, $\boldsymbol{\mu}_{t,k}$ and $\boldsymbol{\Sigma}_{t,k}$ of EGMM, where $\pi_{t,k}$ denotes the prior probability of data point generated from the kth Gaussian component at the t timestamp, $\boldsymbol{\mu}_{t,k}$ the sample mean, and $\boldsymbol{\Sigma}_{t,k}$ the covariance matrix.

$$\pi_{t,k}=\frac{1}{n}\sum\nolimits_{i=1}^{n}p_{t,k}(k\mid \boldsymbol{x}_{t,i},\boldsymbol{\mu}_{t,k}^{g},\boldsymbol{\Sigma}_{t,k}^{g}) \tag{18}$$

$$\boldsymbol{\mu}_{t,k}=\frac{\sum\nolimits_{i=1}^{n}\boldsymbol{x}_{t,i}p_{t,k}(k\mid \boldsymbol{x}_{t,i},\theta_{t,k}^{g})}{\sum\nolimits_{i=1}^{n}p_{t,k}(k\mid \boldsymbol{x}_{t,i},\theta_{t,k}^{g})}-\frac{\lambda}{2}\frac{\sum\nolimits_{(\boldsymbol{x}_{t,i},\boldsymbol{x}_{t,j})\in M_t}(\boldsymbol{x}_{t,i}-\boldsymbol{x}_{t,j})(p_{t,k}(k\mid \boldsymbol{x}_{t,i},\theta_{t,k}^{g})-p_{t,k}(k\mid \boldsymbol{x}_{t,j},\theta_{t,k}^{g}))}{\sum\nolimits_{i=1}^{n}p_{t,k}(k\mid \boldsymbol{x}_{t,i},\theta_{t,k}^{g})} \tag{19}$$

$$\boldsymbol{\Sigma}_{t,k}=\frac{\sum\nolimits_{i=1}^{n}p_{t,k}(k\mid \boldsymbol{x}_{t,i},\theta_{t,k}^{g})\boldsymbol{S}_{t,i,k}}{\sum\nolimits_{i=1}^{n}p_{t,k}(k\mid \boldsymbol{x}_{t,i},\theta_{t,k}^{g})}+\frac{\lambda}{2}\frac{\sum\limits_{(\boldsymbol{x}_{t,i}\boldsymbol{x}_{t,j})\in M_t}(p_{t,k}(k\mid \boldsymbol{x}_{t,i},\theta_{t,k}^{g})-p_{t,k}(k\mid \boldsymbol{x}_{t,j},\theta_{t,k}^{g}))(\boldsymbol{S}_{t,j,k}-\boldsymbol{S}_{t,i,k})}{\sum\nolimits_{i=1}^{n}p_{t,k}(k\mid \boldsymbol{x}_{t,i},\theta_{t,k}^{g})} \tag{20}$$

where $\boldsymbol{S}_{t,i,k}=(\boldsymbol{x}_{t,i}-\boldsymbol{\mu}_{t,k})^{\mathrm{T}}(\boldsymbol{x}_{t,i}-\boldsymbol{\mu}_{t,k})$, $\boldsymbol{S}_{t,j,k}=(\boldsymbol{x}_{t,j}-\boldsymbol{\mu}_{t,k})^{\mathrm{T}}(\boldsymbol{x}_{t,j}-\boldsymbol{\mu}_{t,k})$.

The following is the description of EGMM algorithm:

Input Input data set $\boldsymbol{X}_t$ and the expected number of clusters, denoted as K_t, at the t timestamp; data set $\{C_{t-1},k'\}_{k'=1}^{K_{t-1}}$ from $\boldsymbol{X}_{t(\mathrm{old})}$ with the same cluster labels at the $t-1$ timestamp and the corresponding parameter set $\{\pi_{t-1,k'}, \boldsymbol{\mu}_{t-1,k'}, \boldsymbol{\Sigma}_{t-1,k'}\}_{k'=1}^{K_{t-1}}$; regularization parameter λ.

Outpt Output parameter set $\Theta_t=\{\pi_{t,k}, \boldsymbol{\mu}_{t,k}, \boldsymbol{\Sigma}_{t,k}\}_{k=1}^{K_t}$ and the value of objective function described by Eq. (16).

Method

(1) Initialization: In the 0th iteration step, if $K_{t-1}\geqslant K_t$, randomly select K_t parameters from $\{\pi_{t-1,k'}, \boldsymbol{\mu}_{t-1,k'}, \boldsymbol{\Sigma}_{t-1,k'}\}_{k'=1}^{K_{t-1}}$ as the initial parameter set, namely $\Theta_t^{(0)}=\{\pi_{t,k}^{(0)},\boldsymbol{\mu}_{t,k}^{(0)},\boldsymbol{\Sigma}_{t,k}^{(0)}\}_{k=1}^{K_t}$; if $K_{t-1}<K_t$, initialize $\Theta_t^{(0)}=\{\pi_{t,k}^{(0)},\boldsymbol{\mu}_{t,k}^{(0)},\boldsymbol{\Sigma}_{t,k}^{(0)}\}_{k=1}^{K_t}$ by performing k-means algorithm.

(2) To loop iteration until convergence:

(a) E step: In the lth iteration step, using the current parameter set $\Theta_t^{(l)}=\{\pi_{t,k}^{(l)},\boldsymbol{\mu}_{t,k}^{(l)},\boldsymbol{\Sigma}_{t,k}^{(l)}\}_{k=1}^{K_t}$, calculate the posterior probability $p_{t,k}(k\mid \boldsymbol{x}_{t,i},\Theta_t^{(l)})$ of each data point;

(b) M step: Sequentially update the parameter set $\Theta_t^{(l+1)}=\{\pi_{t,k}^{(l+1)},\boldsymbol{\mu}_{t,k}^{(l+1)},\boldsymbol{\Sigma}_{t,k}^{(l+1)}\}_{k=1}^{K_t}$ according to Eqs. (18−20) to prepare the $(l+1)$th step;

(c) If the convergence criteria are not met, go to (a); otherwise, terminate the loop.

To estimate the parameters of EGMM, the algorithm is performed by way of clustering evolutionary data, which is similar to the parameter estimation method of GMM. The difference between GMM and EGMM is that the previous clustering result is used to adjust the parameter estimation of EGMM. It is inspired from constrained clustering. However, it should be noted that constraint information used in EGMM comes from the previous clustering result rather than user provision.

In the E step, the same method is adopted by EGMM and EM algorithms to solve log-likelihood expectation. Hence, they have the same time complexity, namely $O(nK_tD^3)$, where n is the number of data points, K_t the number of Gaussian components of GMM, and D the dimension of data points. Because EGMM algorithm needs the previous clustering results to smoothen the estimation of current parameters, the corresponding time complexities to estimate each parameter in each step are $O(nK_t)$, $O((n+|\boldsymbol{X}_{t(\mathrm{old})}|)K_t)$ and $O((n+|\boldsymbol{X}_{t(\mathrm{old})}|)K_tD)$, respectively.

4 Experimental Analysis

4.1 Datasets and measure criteria

To verify the effectiveness of EGMM algorithm, synthetic data sets and real data set, named the Columbia University Image Library (COIL), are adopted in this paper. For synthetic data set, the data are two-dimensional ones generated from GMM. According to the evolutionary timestamps, the data sets are generated individually and contain 15 timestamps altogether. Furthermore, the data set at different timestamps will vary in the size and data distribution. In addition, some noise data points are added. The dataset of COIL contains 1 044 objects and each object is an image described by a 1 024-dimentional vector. During the process of the experiment, these objects are randomly divided into five subsets and each subset contains a number of objects of the previous timestamp.

The evolutionary clustering result is a time-varying sequence of clusters. At each timestamp, the clustering result needs to satisfy the two criteria: clustering quality and temporal smoothness. In this paper, clustering accuracy is adopted to assess the clustering quality. To evaluate the temporal smoothness of the clustering sequence, the fluctuation of history cost is compared.

4.2 Synthetic datasets

As shown in Table 1, the clustering accuracy of three algorithms including EM, evolutionary *k*-means and EGMM are compared, which are performed on synthetic datasets at six timestamps, respectively. The data sets with noise data are generated from GMM, therefore the EM algorithm has the higher and more stable clustering accuracy. For this special synthetic datasets, its clustering accuracy is even higher than the clustering of evolutionary *k*-means. The foundation of evolutionary *k*-means is *k*-means algorithm, so satisfactory clustering results are difficult to be obtained for non-spherical datasets by running evolutionary *k*-means. From the stability of clustering results, a greater fluctuation often arises for the impact of noise data. As the evolutionary clustering version of EM algorithm, EGMM is still suitable for handing non-spherical datasets. By using the previous clustering results to adjust the parameters estimations, EGMM has the better performance to deal with the evolving datasets. Hence, the clustering quality is more stable than the other two algorithms and the algorithm is less sensitive to noise data. The results running on synthetic data sets show that the clustering results tend to be stable within a short time.

Table 1 Comparison of clustering accuracy on synthetic datasets (Acc)

Timestamp	T_1	T_2	T_3	T_4	T_5	T_6
EM	0.83±5%	0.87±5%	0.88±5%	0.84±5%	0.85±5%	0.84±5%
Evolutionary *k*-means	0.67±5%	0.75±5%	0.83±5%	0.85±5%	0.80±5%	0.82±5%
EGMM	0.83±5%	0.91±5%	0.90±5%	0.91±5%	0.92±5%	0.91±5%

The history cost of EGMM at each timestamp is demonstrated in Fig. 1. The results running on synthetic datasets show that the history cost fluctuates within a short range, which means the clustering sequence of EGMM has better performance of temporal smoothness.

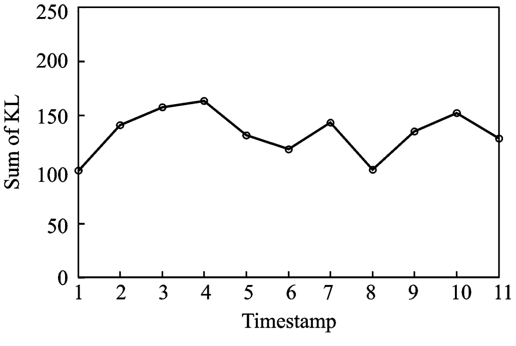

Fig. 1 History cost of EGMM on synthetic datasets

4.3 Real datasets

To further validate the performance of EGMM, the dataset of COIL is divided into five subsets, then an evolutionary dataset that may drift at five timestamps is successfully constructed. It should be noted that all subsets at each timestamp contain a number of same or different samples to simulate the evolution scenario better. In Table 2, the clustering quality of three algorithms including EM algorithm, evolutionary *k*-means and EGMM algorithm, is demonstrated, where the three algorithms are performed on the dataset of COIL at five timestamps. Compared with EM algorithm, the clustering results of EGMM better reflect the evolutionary features of

Table 2 Comparison of clustering accuracy on COIL dataset (Acc)

Timestamp	T_1	T_2	T_3	T_4	T_5
EM	0.58±5%	0.64±5%	0.61±5%	0.63±5%	0.66±5%
Evolutionary *k*-means	0.53±5%	0.62±5%	0.59±5%	0.64±5%	0.61±5%
EGMM	0.58±5%	0.67±5%	0.71±5%	0.73±5%	0.72±5%

data points by introducing historic clustering results; and compared with the evolutionary *k*-means, EGMM can better fit the data features. The experimental results show that the evolutionary clustering framework of EGMM can obtain better clustering quality and realize the better data fitting in a short time.

In EGMM model, λ is an adaptive parameter used to adjust the confidence of a priori. That is to say, different values of λ reflect different degrees of effects coming from the previous clustering results during the process of parameter estimation. In this paper, we adopt the approach by dynamically and adaptively selecting λ. Experiments show that several factors can affect the value of λ, including the discrepancy between two distributions, the size of datasets and the number of joint data points at adjacent timestamps.

5 Conclusions

EGMM based on constraint consistency is proposed, which is closely related to the two technologies, i. e. , the constrained clustering and the evolutionary clustering. The basic idea comes from the constrained clustering, and the optimization method of model parameter needs to perform evolutionary clustering. From the feature of dataset, data points will evolve along time; and from the obtained clustering results, a clustering sequence with temporal smoothness must be guaranteed. The method thus belongs to evolutionary clustering. To guarantee the clustering quality at each timestamp and the temporal smoothness of clustering sequence, the previous clustering results are used as a priori constraints to adjust the parameter estimation and the assignment of data points. From this perspective, the method belongs to constrained clustering.

Compared with the existing constrained clustering algorithm, the a priori constraints come from the previous clustering results rather than the labeled data points. These a priori constraints can be obtained in a more natural way, and can reflect the internal structure of the data more truely. As the evolutionary algorithm, the algorithm of EGMM uses the overall differences and individual differences simultaneously to smoothen clustering results. Therefore the fluctuation of history cost becomes more stable.

As many other existing approaches, EGMM still has some limitations. For example, the adaptive parameter λ is used to adjust the confidence of a priori, therefore, how to set it clearly according to previous clustering results is an interesting topic. To evolutionary clustering, the number of clusters may be changed with the evolution of data distributions, and then how to determine the number of Gaussian components becomes another interesting topic. Moreover, the method of EGMM can be used to handle various evolving data which evolves along time. Now, we are engaged in the analysis of air passenger behavior, which gradually evolves with region and age.

References:

[1] Hand D, Mannila H, Smyth P. Principles of data mining[M]. USA: MIT Press, 2001.

[2] Wang Lina, Wang Jiandong, Jiang Jian. New shadowed c-means clustering with feature weights[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2012, 29(3): 273-283.

[3] Backer E, Jain A. A clustering performance measure based on fuzzy set decomposition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1981,3(1):66-75.

[4] Jain A K. Data clustering: 50 years beyond K-means [J]. Pattern Recognition Letters, 2010,31(8): 651-666.

[5] Zhang J, Song Y, Chen G, et al. On-line evolutionary exponential family mixture[C]// Proceeding of the 21st international joint conference on Artificial in-

telligence. [S. l.]: Morgan Kaufmann Publishers, 2009:1610-1615.

[6] Falkowski T, Bartelheimer J, Spiliopoulou M, et al. Mining and visualizing the evolution of subgroups in social networks[C]// Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence. [S. l.]: IEEE Computer Society, 2006:52-58.

[7] Zhang Bo, Xiang Yang, Huang Zhenhua. Recommended Trust computation method between individuals in social network site[J]. Journal of Nanjing University of Aeronautics and Astronautics, 2013, 45(4): 563-568.

[8] Falkowski T, Bartelheimer J, Spiliopoulou M. Mining and visualizing the evolution of subgroups in social networks[C]// Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence. Hong Kong, China: IEEE Computer Society, 2006: 52-58,.

[9] Ning H, Xu W, Chi Y, et al. Incremental spectral clustering with application to monitoring of evolving blog communities[C]// Proceedings of the 7th SIAM International Conference on Data Mining (SDM 2007). Minneapolis, Minnesota, USA: SIAM, 2007:261-272.

[10] Zhang Chao, Wang Daobo, Farooq M. Real-time tracking for fast moving object on complex background[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2010, 27(4): 321-325.

[11] Chi Y, Song X, Zhou D, et al. Evolutionary spectral clustering by incorporating temporal smoothness[C] // Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007). [S. l.]: ACM Press, 2007: 153-162.

[12] Xu K S, Kliger M, Hero A O. Adaptive evolutionary clustering[J]. Data Mining and Knowledge Discovery, 2014,28:304-336.

[13] Chakrabarti D, Kumar R, Tomkins A. Evolutionary clustering[C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery And Data Mining (KDD 2006). [S. l.]: ACM Press, 2006:554-560.

[14] Xu Tianbing, Zhang Zhongfei, Yu Philip, et al. Evolutionary clustering by hierarchical Dirichlet process with hidden Markov state[C]// Proceedings of 8th International Conference on Data Mining (ICDM 2008). [S. l.]: IEEE Computer Society, 2008: 658-667.

[15] Ravi Shankar, Kiran G V R, Vikram Pudi. Evolutionary clustering using frequent item sets[C]// 14th International Conference of Knowledge-Based and Intelligent Information and Engineering Systems (KES 2010). [S. l.]:Springer, 2010:11-20.

[16] Bishop Christopher M. Pattern recognition and machine learning [M]. Singapore: Springer, 2006.

[17] He Xiaofei, Cai Deng, Shao Yuanlong, et al. Laplacian regularized Gaussian mixture model for data clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2011,23(9):1406-1418.

[18] Liu Jialu, Cai Deng, He Xiaofei, et al. Gaussian mixture model with local consistency[C]// Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI 2010). USA: AAAI Press, 2010: 512-517.

[19] Chi Yun, Song Xiaodan, Zhou Dengyong, et al. Evolutionary spectral clustering[J]. ACM Transactions on Knowledge Discovery from Data, 2009,3(4):1-30.

[20] Bilmes J A. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov mode[R]. TR-97-021 U. C. Berkeley, 1998.

[21] Zhou Dengyong, Bousquet O, Navin Lal T, et al. Learning with local and global consistency[C]// Advances in Neural Information Processing Systems 16 (NIPS 2003). [S. l.]:MIT Press,2003:321-328.

(Executive editor: Zhang Tong)