SIViP (2016) 10:359–368
DOI 10.1007/s11760-015-0749-5

ORIGINAL PAPER

A new clustering method of gene expression data based on multivariate Gaussian mixture models

Zhe Liu · Yu-qing Song · Cong-hua Xie · Zheng Tang

Received: 15 September 2013 / Revised: 16 December 2014 / Accepted: 21 January 2015 / Published online: 8 February 2015

Abstract Clustering gene expression data are an important problem in bioinformatics because understanding which genes behave similarly can lead to the discovery of important biological information. Many clustering methods have been used in the field of gene clustering. This paper proposed a new method for gene expression data clustering based on an improved expectation maximization(EM) method of multivariate Gaussian mixture models. To solve the problem of over-reliance on the initialization, we propose a remove and add initialization for the classical EM, and make a random perturbation on the solution before continuing EM iterations. The number of clusters is estimated with the Quasi Akaike's information criterion in this paper. The improved EM method is tested and compared with some other clustering methods; the performance of our clustering algorithm has been extensively compared over several simulated and real gene expression data sets. Our results indicated that improved EM clustering method is superior than other clustering algorithms and can be widely used for gene clustering.

Keywords Gene expression data · Clustering · Multivariate Gaussian mixture models · Expectation maximization · QAIC criterion

Z. Liu (✉) · Y. Song · Z. Tang
School of Computer Science and Telecommunication,
Jiangsu University, Room 522, Zhenjiang, Jiangsu Province,
People's Republic of China
e-mail: lxxc1016@gmail.com

Z. Liu
School of Computer Science, Jilin Nomal University, Sipin,
Jilin Province, People's Republic of China

C. Xie
School of Computer Science and Engineering, Changshu Institute
of Technology, Suzhou, Jiangsu Province, People's Republic of China

1 Introduction

Microarray technology has been widely applied in study of measuring gene expression levels for thousand of genes simultaneously. In this technology, cluster analysis is the most important method for gene expression data analysis [1–5]. By comparing the expression pattern of unknown genes to those known functions, one can predict the functions of unknown genes. This is the primary objective of the unsupervised cluster analysis of gene expression data.

Many clustering algorithms have been proposed for gene expression data analysis. The hierarchical clustering is one of the earlier methods applied to cluster the gene expression data [6,7]. K-means clustering methods are used in gene expression data analysis due to its high computational performances [8,9]. As one kind of neural network, self-organizing map (SOM) which presents high-dimensional data by the low-dimensional data has also been used for gene expression data clustering [10]. Other common clustering methods include CAST algorithm [11] , SVM clustering [12] and model-based clustering [13,14]. Gene expression data have a lot of noise, and large amounts of data behind many variables cannot be observed. The model-based clustering algorithm assumes that the data comply with the internal framework for probabilistic model, according to the different parameters, the observed data can be divided into different clusters. Successful application of the model-based clustering to gene expression data has been reported; the most well-known mixture models clustering is the MCLUST [15,16]. Yeung et al. [17] applied finite Gaussian mixture model to fit the yeast cell cycle data and showed good fitness of the model. Yi et al. [18] used the model-based algorithm in supervised clustering of gene expression data. Gaussian mixture models have received the bulk of the attention in the mixture modeling literature due to their mathematical tractability [19,20].

The model-based clustering algorithm is used to estimate the parameters using EM algorithm. EM algorithm is a popular way to estimate the parameters of mixture models. Unfortunately, its performance highly depends on the initialization. In order to solve this problem, we propose to utilize an improved EM algorithm to estimate the model parameters. In our proposed method, we use random swap K-means algorithm to initialize the multivariate Gaussian mixture models. The EM algorithm starts with some initial values of all unknowns and iteratively updates each parameter conditional on the parameter values in the previous round of the iteration. Without any prior knowledge, each gene may be assigned to each cluster with equal probability.

This paper is structured as follows. Sect. 2 reviews the multivariate Gaussian mixture models and EM algorithm. The improved EM algorithm based on multivariate Gaussian mixture models is given in Sect. 3. Section 4 provides experimental results and analysis using the simulated and real gene expression data sets. We then conclude the paper in Sect. 5.

2 Multivariate Gaussian mixture models and EM algorithm

2.1 Multivariate Gaussian mixture models

The gene expression data are arranged in an m*n matrix denoted by X, where n is the number of genes and m is the level of treatment. Suppose $X = \{x_1, x_2, \ldots, x_n\}$ is a random observation gene expression data set. All $x_i (1 \leq i \leq n)$ are mutually independent. Let $x_i = [x_{i1}, x_{i2}, \ldots, x_{im}]^T$ be the i-th column of matrix X^T, where x_{ij} be the expression level of the i-th gene in the j-th treatment, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. $f(x_i)$is the corresponding probability density function, in which $x \in R^d$ is a d-dimensional random variable value of x_i and θ_i is the parameter. With the finite multivariate mixture model, each x_i is assumed to follow an m-dimensional mixture of normal distributions. So the mixture distribution probability density $f(x_i)$ is

$$f(x_i) = \sum_{k=1}^{C} \alpha_k f_k(x_i \,|\theta_i) \tag{1}$$

with$\sum_{k=1}^{C} \alpha_k = 1, 0 \leq \alpha \leq 1$, here C is the number of components of the mixture models. α_1 is the proportion of mixture or weighing, and $\theta = (\alpha_1, \ldots, \alpha_C, \theta_1, \ldots, \theta_C)$is the parameter space of models.

If the distribution of the component density function $f_i(x\,|\theta_i\,)$can be determined, the model will become the mixture models of this component, and Eq. (1) can be called as mixture density function. Although m in Eq. (1) is usually treated as a fixed value, its real value is unknown in most applications. A common method of estimating the parameters of finite mixture models is the EM algorithm [21], which can be used to estimate maximum likelihood model parameters in incomplete data.

Multivariate Gaussian mixture model is a common mixture density model, which is used to describe the distribution of spatial data. Gaussian hypotheses are generally made (i.e., $\theta = (\alpha_1, \mu_1, \sum_1, \alpha_2, \mu_2, \sum_2, \ldots, \alpha_k, \mu_k, \sum_k)$, where $\mu_k(\text{a}\, m \times 1\text{vector})$ is the mean of model k, $\sum_k(\text{an}\, m \times m$ matrix) is the covariance of model k. The multivariate Gaussian mixture probability density is expressed as

$$f(x_i \,|\mu_k, \Sigma_k\,) = \sum_{k=1}^{C} \alpha_k f_k(x_i \,|\mu_k, \Sigma_k\,) \tag{2}$$

2.2 EM algorithm

The EM algorithm [22–24] is an iterative technique for finding maximum likelihood estimates when there are, or are assumed to be, missing data. Without any prior knowledge, each gene may be assigned to each cluster with equal probability. The EM algorithm iterates between an expectation(E) step and a maximization(M) step. In the E step, we compute the expectation of the log likelihood of complete data with respect to latent variables given the current parameter estimates. In the M step, we maximize the expected log likelihood of complete data. Therefore, the EM algorithm transfers the problem of maximizing the original log likelihood to the problem of maximizing the expected log likelihood of complete data, which usually much easier to deal with. The EM iteration is described in the following steps:

(1) Initialization: The parameters needed for the next step are initialized in the following way:

$$P_{ik}^{0} = 1/C \qquad \forall i = 1, \ldots, n;\ k = 1, \ldots, C \tag{3}$$

$$\alpha_k^{0} = 1/C \qquad \forall k = 1, \ldots, C \tag{4}$$

(2) Step E: this step consists of the estimation of the posterior probability, i.e., $P_{ik}^{t}(x_i)$for the gene expression data k belonging to the class k at the t-th iteration:

$$P_{ik}^{t}(x_i) = \frac{\alpha_k^t f(x_i\,|\theta_k^t\,)}{\sum_{k=1}^{C} \alpha_k^t f(x_i\,|\theta_k^t\,)} \tag{5}$$

(3) Step M: the parameters needed for the next step are estimated in the following way:

$$\alpha_k^{t+1} = \frac{1}{n}\sum_{i=1}^{n} P_{ik}^{t}(x_i) \tag{6}$$

$$\mu_k^{t+1} = \frac{\sum_{i=1}^{n} x_i P_{ik}^{t}(x_i)}{\sum_{i=1}^{n} P_{ik}^{t}(x_i)} \tag{7}$$

$$\Sigma_k^{t+1} = \frac{\sum_{i=1}^{n} P_{ik}^{t}(x_i)(x_i - \mu_k^{t+1})(x_i - \mu^{t+1})^T}{\sum_{i=1}^{n} P_{ik}^{t}(x_i)} \tag{8}$$

(4) Turn back to step 2 or stop until convergence.

3 The improved EM algorithm based on multivariate Gaussian mixture models

The idea of the improved EM algorithm is to alternate between simple perturbation to the solution by remove–add and convergence toward nearest optimum by the EM algorithm. The initialization is performed as in the K-means method, after the solution has been initialized, we perform remove and add operations.

Our method is outlined as follows:

(1) Initialization step: the number of class C is assumed to be known. An initial solution of the parameters (μ_k^0, Σ_k^0) of the multivariate mixture models is extracted by the K-means. The parameter (α_k^0) of mixture weight can be initialized as following:

$$\alpha_k^0 = \frac{N_k^0}{N} \quad \forall k = 1, \ldots, C \tag{9}$$

(2) Remove and Add step: the remove operation is done by selecting a component randomly and adds a component. The location of the new component is decided by selecting one data point, and setting it as the mean vector of the new component. The new component is therefore more likely to be placed in areas of high point density, such as cluster centers. At the s-th iteration when remove and add operation, the parameters of the new component are as follows:

$$\mu_r^s = x_p \quad r = random(1, C) \tag{10}$$

$$\alpha_r^s = \sum_{l=1, l\neq r}^{C}\left(\sum_{i=1}^{N} P_{il}^{s-1}\right)\alpha_r^{s-1} \tag{11}$$

$$\Sigma_r^s = \sum_{l=1, l\neq r}^{C}\left(\sum_{i=1}^{N} P_{il}^{s-1}\right)\Sigma_r^{s-1} \tag{12}$$

(3) Step E: this step consists of the estimation of the posterior probability, i.e., $P_{ik}^{t}(x_i)$for the gene expression data x_i belonging to the class k at the t th iteration:

$$P_{ik}^{t}(x_i) = \frac{\alpha_k^t f(x_i \,|\theta_k^t)}{\sum_{k=1}^{C}\alpha_k^t f(x_i \,|\theta_k^t)} \tag{13}$$

(4) Step M: the parameters needed for the next step are estimated in the following way:

$$\mu_k^{t+1} = \frac{\sum_{i=1}^{n} x_i P_{ik}^{t}(x_i)}{\sum_{i=1}^{n} P_{ik}^{t}(x_i)} \tag{14}$$

$$\alpha_k^{t+1} = \frac{1}{n}\sum\nolimits_{i=1}^{n} P_{ik}^{t}(x_i) \tag{15}$$

$$\Sigma_k^{t+1} = \frac{\sum_{i=1}^{n} P_{ik}^{t}(x_i)(x_i - \mu_k^{t+1})(x_i - \mu^{t+1})^T}{\sum_{i=1}^{n} P_{ik}^{t}(x_i)} \tag{16}$$

(5) Turn back to step 2 or stop until convergence.

(6) Turn back to step 2 or stop when the number of remove and add operation is large enough.

This improved EM algorithm iteration scheme is robust and well behaved. The convergence speed is also reasonably fast.

4 Result and analysis

In this section, in order to evaluate the gene expression data clustering algorithm based on multivariate Gaussian mixture models, both synthetic and real data sets used in the clustering experiments are introduced and analyzed. All the experiments were programmed by C++. Lenovo PC with CPU—Intel (2) nl (TM), speed—1.5GHz, memory—1Gb, hard disk—120Gb is used in this experiment.

4.1 Evaluation measures

The adjusted rand index (ARI) evaluates the degree of agreement between two partitions of the same set of objects [25]. Suppose C is the true clustering of a gene expression data set based on domain knowledge, and C' is clustering result given by some clustering algorithm. Let a represent the number of pairs that are members of the same cluster in both C and C', b represent the number of pairs belonging to the same cluster in C but to different clusters in C', c represent the number of pairs belonging to different cluster in C but to the same clusters in C' and d represent the number of pairs belonging to different clusters in both C and C'. The ARI (C, C') is defined by

$$\text{ARI(C,C)} = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)} \tag{17}$$

The higher the value of ARI indicates that when C is more similar to C', the better the clustering performance.

The quantities, Precision (P) and Recall (R), are methods of performance evaluation. Precision is a measure of how much noise there is in the output of the detector, while recall is a measure of how much of the ground truth is detected. Precision is the fraction of detections that are true positives rather than false positives, and Recall is the fraction of true positives that are detected rather than missed. F-measure that combines precision (precision) and recall (recall) evaluate the clustering results

$$F = \frac{2PR}{P + R} \tag{18}$$

The higher the value of F, the better the clustering performance.

4.2 The number of clusters

A good estimation of the number of the clusters is very important for clustering. The number of clusters may be treated as another parameter and inferred from the data. The Akaike's information criterion (AIC) or Bayesian inference criterion (BIC) is used to estimate the optimal number of clusters [26,27]. The AIC is

$$\mathrm{AIC}(G) = -2\mathrm{In}L(\theta(G)) + 2P(G) \tag{19}$$

Where P is the number of parameters to be estimated in the model, $L(\theta(G))$ is the likelihood value evaluated at $\theta(G)$,the vector of maximum likelihood estimate of the parameters, and G is the Gaussian density function. The number of clusters(C) that has the minimum AIC value is the estimated C.

Bozdgoan proposed the modified AIC called AIC_3 and AIC_4, and their expression as follows:

$$\mathrm{AIC}_3(G) = -2\mathrm{In}L(\theta(G)) + 3P(G) \tag{20}$$

$$\mathrm{AIC}_4(G) = -2\mathrm{In}L(\theta(G)) + 4P(G) \tag{21}$$

In order to adapt to the over-dispersed data, Lebreton [28] proposed the modified AIC criteria as QAIC and the expression is:

$$\mathrm{QAIC}(G) = -2^{*}\ln L(\theta(G))/c + 2P(G) \tag{22}$$

where c is a single variance inflation factor, which can be estimated from the goodness-of-fit Chi-square statistic χ^2 of the global model and its degrees of freedom,

$$c = \chi^2/df \tag{23}$$

We compare and analyze the experiment results based on the simulated datasets, consisting of 81 data, each data has coordinate attributes (x, y), as we can depict in Fig. 1. Figure 2 illustrate the results of density function and the Gaussian mixture density estimation when the clustering number $C = 2, 3, 4, 5$, singular covariance is existed when set $C = 6$, so the maximum value of the clustering number $C = 5$. From Fig. 2, the result of the mixture density function is smoothest when $C = 3$. How can the number C of mixture models be quantitative analyzed? We have done a lot of experiments by using the methods of the AIC, AIC_3, AIC_4, BIC and QAIC, the results illustrated in Figs. 3 and 4.

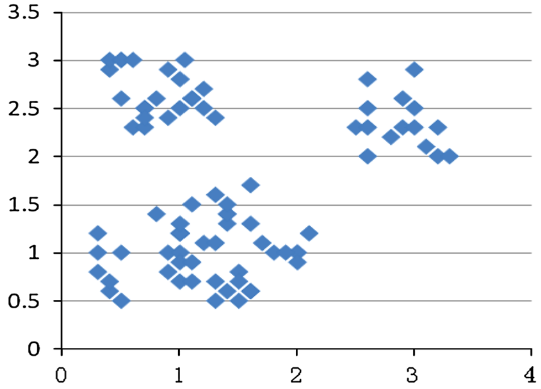
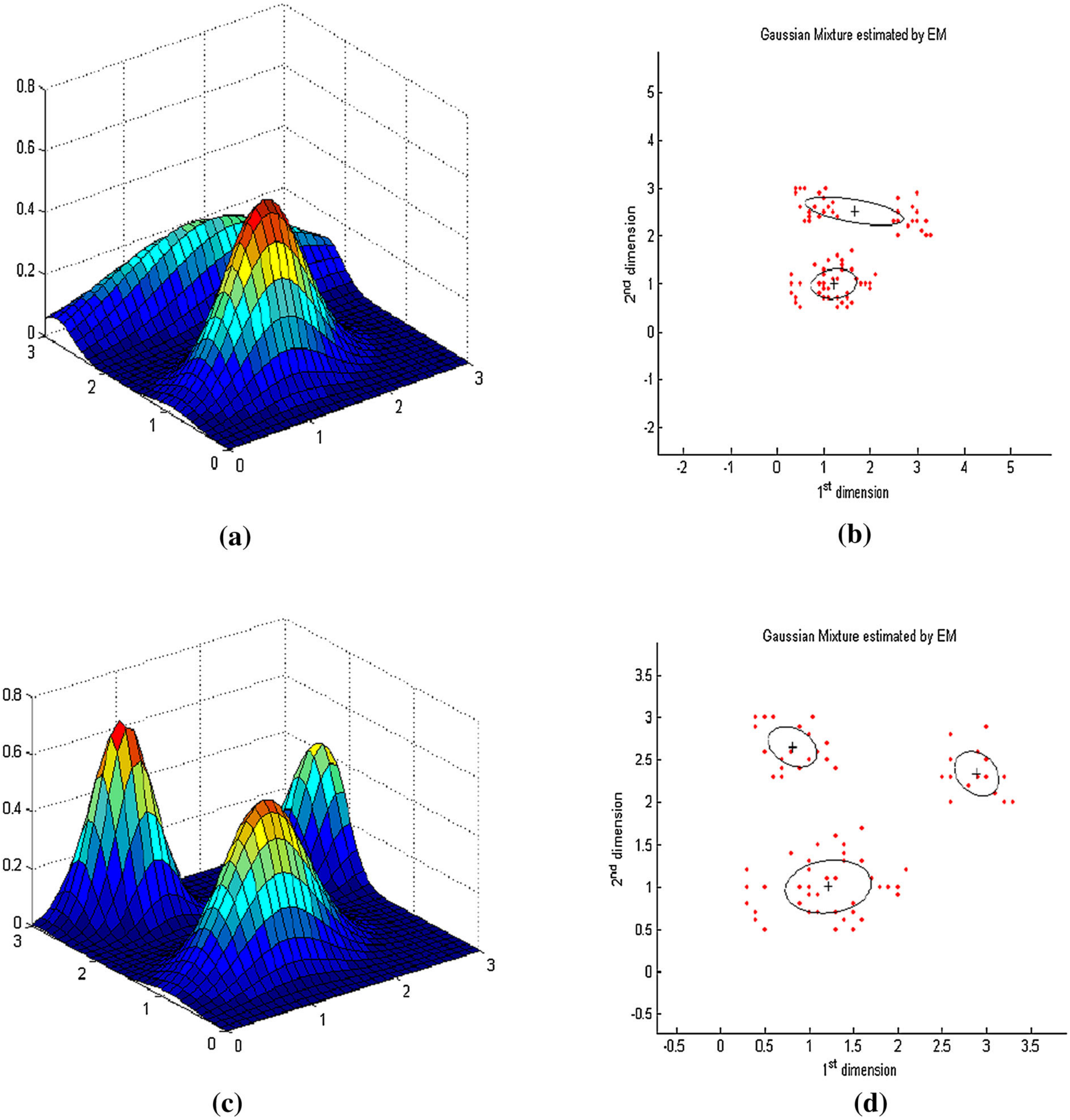


Fig. 1 Simulated dataset D1

But the log-likelihood function values of AIC, AIC_3, AIC_4 are larger than those of $P(G)$, so the trend of decline is shown in Fig. 3a–c. If we want to obtain the minimum function value and the best fitting graphics, the value of G is always large and can result in over-fitting. The method of BIC increased the value of the penalty function, with more obvious trend of increasing first and decreasing later. When the value of G is 3, we obtain the minimum function value and the best fitting.

For the method of QAIC, the value of C is set to 2–9 corresponding to Fig. 4a–h, 11. From the trend of the figure: when $c = 2$, the function values of the QAIC method decrease from large to small; when $c = 3, 4, 5$, the values decrease first and increase gradually, and the monotonous decreasing amplitude is large while the monotonous increasing amplitude is small; when $c = 6, 7, 8$, the values of QAIC function subject to monotonous decrease first and monotonous increase later, and the increase changes from small to large, at this time the monotonous decreasing amplitude is small while the monotonous increasing amplitude is large, and the line of symmetry moves from right to left, only when $C = 5$, the minimum value of function is the closest to the axis of symmetry, so $C = 5$ and $G = 3$ is the most reasonable fitting.

From the analysis above, considering G values which is based on information criterion, we need to consider the relationship between the value of likelihood function and penalty function value, which means specific data set is needed to be modified constantly.



(a) (b)

(c) (d)

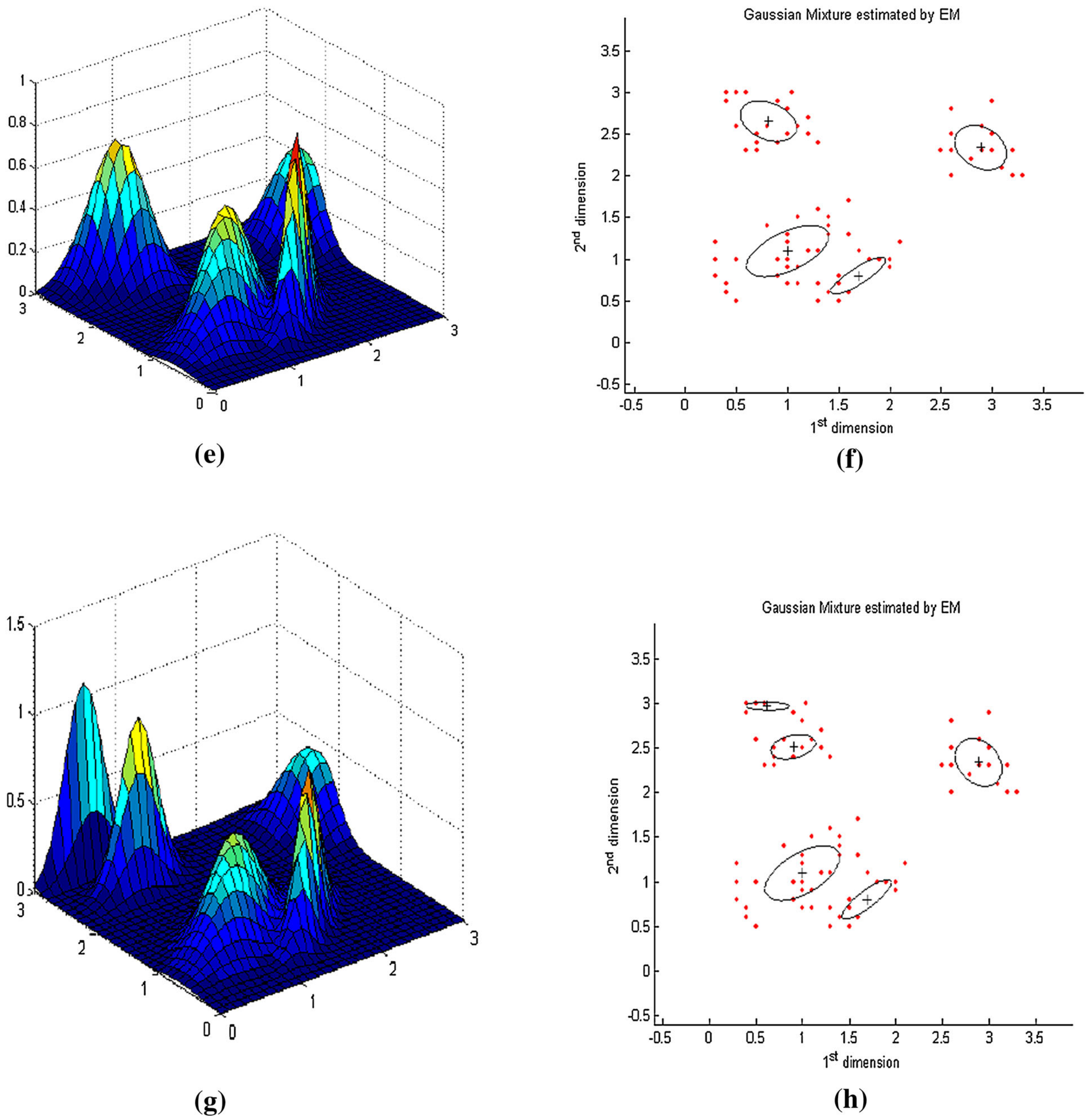
Fig. 2 Results of the different density function and the Gaussian mixture estimation of the dataset D1 **a** The density function of dataset D1 $G = 2$) **b** The Gaussian mixture estimation of the dataset D1 ($G = 2$). **c** The density function of dataset D1 ($G = 3$). **d** The Gaussian mixture estimation of the dataset D1 ($G = 3$). **e** The density function of dataset D1 ($G = 4$). **f** The Gaussian mixture estimation of the dataset D1 ($G = 4$). **g** The density function of dataset D1 ($G = 5$). **h** Estimation of the dataset D1 ($G = 5$)

When the function AIC(G) obtains the minimum value, G is most reasonable number of the mixture density model. But to search the minimum value for AIC function, we need to be set up a larger M in advance, and find a minimum value of function AIC in the interval of 1 to M. But M should be more than the minimum parameters G at least, but it is difficult to determine; at the same time, it is difficult to guarantee efficiency when M is larger and computational time is too much.

According to the analysis of the above experiments, we used this QAIC method for determining the number of components, the criterion function expression:

$$\mathrm{QAIC}(G) = -2^{*}\mathrm{In}L(\theta(G))/c + 2P(G) \tag{24}$$

where $c = 5$. We can determine the number of components from this expression.



(e)

(f)

(g)

(h)

Fig. 2 continued

4.3 Synthetic data experiment

We tested the algorithms using synthetic data sets on different simulated multivariate Gaussian mixture distributions. We demonstrate the Gaussian models estimated from EM, and our improved EM method on synthetic data sets in Fig. 5. The data sets contain more than 10,000 points, and the models are displayed as ellipses. The experiment is conducted by 20 repetitions.

Figure 5a shows the initialization effect of the data sets. Figure 5b is the result of EM clustering and classification compared to the initial solution, while the clustering result of our improved EM method in Fig. 5c is clearly better in parameter estimation than initial solution and EM. For the standard EM has the shortcoming of getting stuck in a local maximum and our proposed improved EM algorithm overcome it well.

The number of add and remove operations is a key parameter in our improved EM method .To ensure a good solution, the number of iterations for remove and add has been set large enough. We observed that the effect of a bad initialization is diminished when remove–add operation is used. We there-

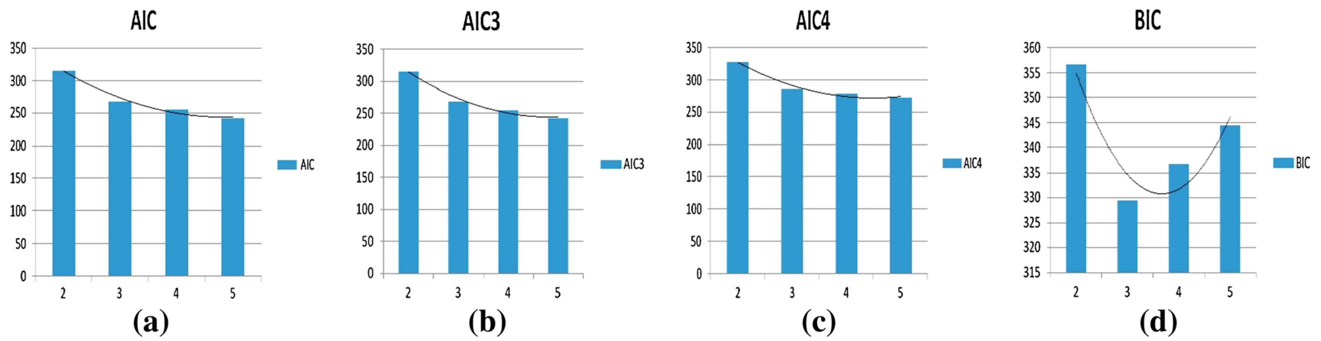


Fig. 3 The values of *AIC(G)*, *AIC(G)*$_3$, *AIC(G)*$_4$ and *BIC*

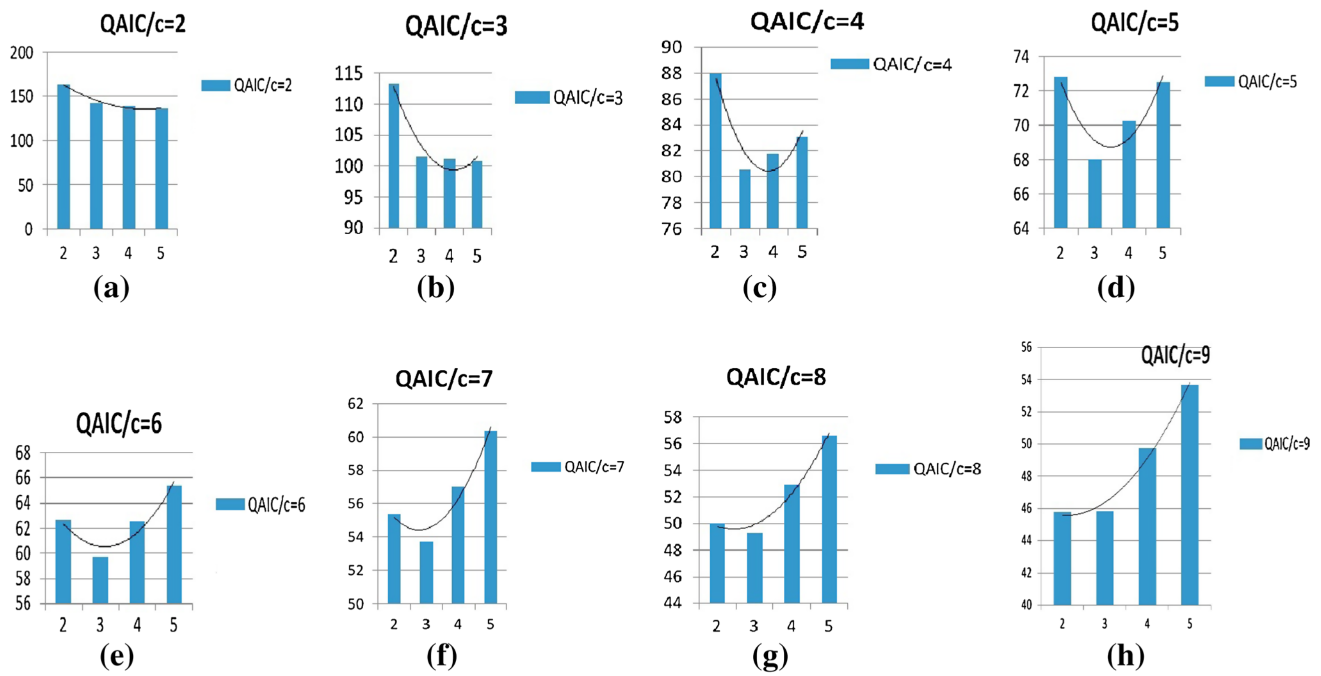


Fig. 4 The values of QAIC(G) when $c = 2, c = 3, c = 4, c = 5, c = 6, c = 7, c = 8, c = 9$

fore expect remove and add operation to yield good results with Gaussian mixture models. For a good remove and add to occur, a badly placed component must be chosen and a location from the area where the component needs to move must also be chosen.

To compare the random swap EM algorithm with the standard EM algorithm, we calculated the quantitative accuracy of the results are shown in Table 1.

From Table 1, we can observe that RSEM got higher precision (P) and recall (R) value compared with EM. According to column (3), our improved EM method undoubtedly reached a higher value of F, which means the boundary representation is relatively better. The higher value of ARI of our improved EM method indicates that C is more similar to C', so the clustering performance of our improved EM method is better. With the random perturbation on the solution before continuing EM iterations and the simple parameter-setting of the number of add and remove operation, the random swap EM is shown to be much simpler and more efficient than EM.

4.4 Real data sets experiment

In this section, we compare performance of our improved EM algorithm with the classical EM, EM algorithm based on multivariate distributions [29], two standard clustering algorithm [30], and K-means [31] on the four real gene expression data sets.

The rat CNS data set was obtained by reversing transcription-coupled PCR to study the expression levels of 112 genes during rat central nervous system development over nine time points. These 112 genes were selected from four gene families according to prior knowledge of biology by Wen et al.

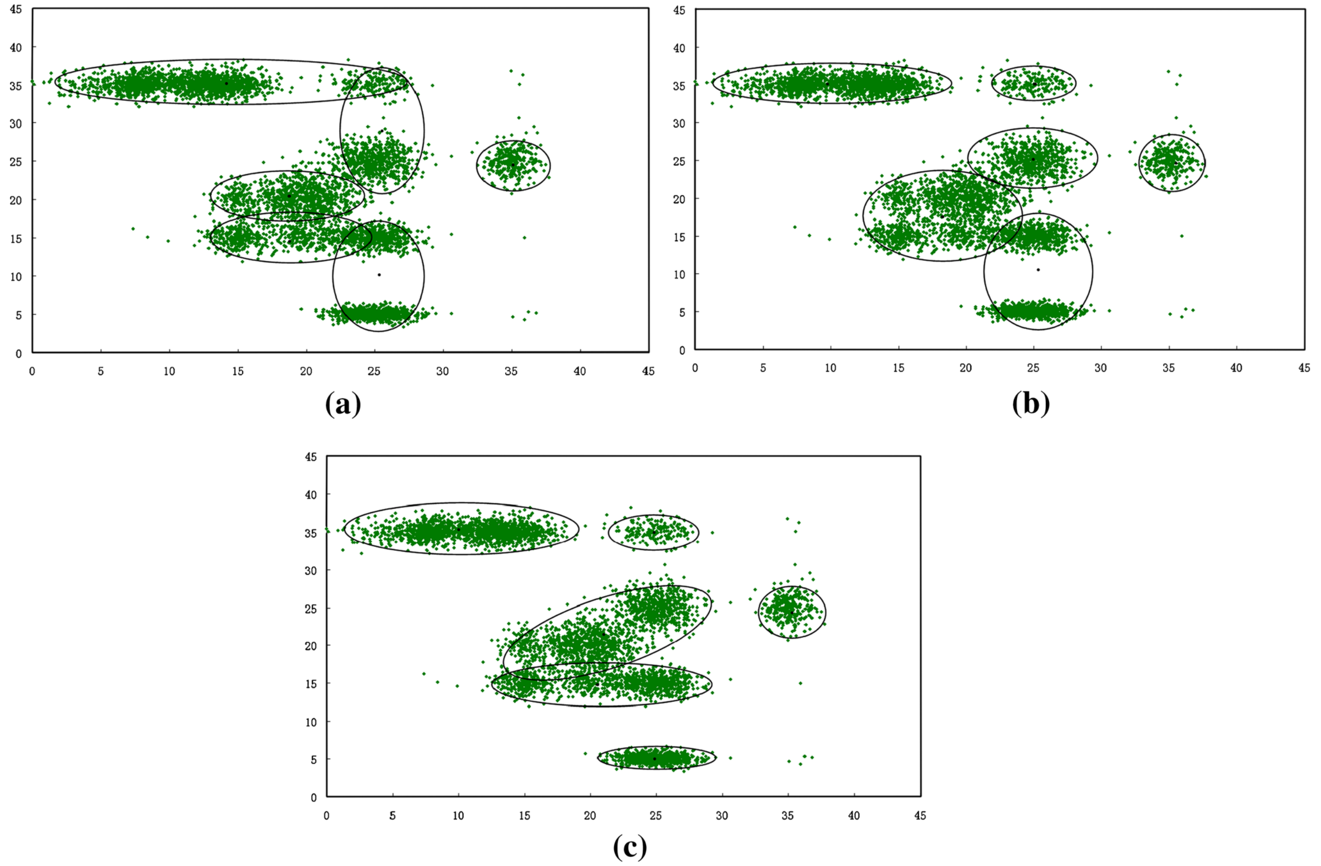


Fig. 5 Gaussian models estimated from initial solution, EM and Our method. **a** Initialization. **b** EM. **c** Our improved EM method

Table 1 P's, R's, and F's for the EM algorithm and RSEM algorithm

Methods	*P*	*R*	*F*	ARI
EM	0.8423	0.9274	0.8828	0.9248
Our method	0.9318	0.9853	0.9578	0.9892

[32] we used this 112 × 9 microarray as one of our experimental data sets and took these four functional categories as external standard categories.

The human fibroblasts serum data set contains the expression levels of 8613 human genes [33]. The data set has 13 dimensions corresponding to 12 time points (0,0.25,0.5,1,2, 4,6,8,12,16,20, and 24h) and one unsynchronized sample. A subset of 517 genes used in our experiments whose expression levels changed substantially across the time points have been chosen [34].

The yeast cell cycle data set is from http://cellcycle-www.stanford.edu. In the study by Yeung et al. [17], a subset of 384 genes was used ($n=84$). These genes had expreddion levels peaking at different times corresponding to the five ($C=5$) phases of the cell cycle, including Early G1, Late G1, S, G2 and *M*. This subset of the data is available at http://www.cs.washington.edu/homes/kayee/model. For preprocessing, we removed the data corresponding to the 90- and 100-min time points, because the two time point, points were reported to be unreliable [35]. After the deletion, the total number of treatment became 15. We then standardized each gene expression profile by subtracting the mean expression from the original value and dividing the difference by the SD so that the transformed expression level has a 0 mean and variance of 1. All the 384 genes were assigned to one of the five clusters by the original investigators [17].

The HAHrma data consist of time course responses of human bronchial cell line A549 to Interleukin 13(IL13), a protein coded by the IL13 gene. Il13 is known to up-regulate CD23 and MHC class II expression, promote switching of the IgE isotype in a special kind of white blood cells known as B cells, and down-regulate the production of pro-inflammatory cytokines and chemokines that aid in the defense mechanism for white blood cells [36]. The human bronchial cells were exposed to IL13, and measurements on the expression levels of the 22,283 genes were taken at 0, 4, 12 and 24 hours after exposure by hybirdization with Affymetrix U133a chips.

For all the competitor algorithms, we used this QAIC method for determining the number of components based on four real data sets. We analyzed and compared these clus-

Table 2 Comparison of the clustering results of various classification methods on the rat CNS data

Methods	*P*	*R*	*F*	ARI
EM	0.8194	0.8296	0.8244	0.8261
EM based on multivariate *t*-distributions	0.7349	0.8017	0.7716	0.8170
FCM	0.7089	0.7802	0.7429	0.7728
K-means	0.6213	0.6913	0.6544	0.6781
Improved EM	**0.9031**	**0.9425**	**0.9224**	**0.9713**

Table 3 Comparison of the clustering results of various classification methods on the human fibroblasts serum data

Methods	*P*	*R*	*F*	ARI
EM	0.8865	0.8912	0.8888	0.8749
EM based on multivariate *t*-distributions	0.8727	0.8619	0.8672	0.8678
FCM	0.8023	0.8502	0.8255	0.8418
K-means	0.6914	0.7029	0.6971	0.7920
Improved EM	**0.9325**	**0.9536**	**0.9382**	**0.9430**

Table 4 Comparison of the clustering results of various classification methods on the yeast cell cycle microarray data

Methods	*P*	*R*	*F*	ARI
EM	0.7895	0.8976	0.8401	0.8496
EM based on multivariate *t*-distributions	0.7743	0.8812	0.8386	0.8370
FCM	0.7121	0.8200	0.7529	0.7498
K-means	0.6510	0.7219	0.6438	
Improved EM	**0.9125**	**0.9654**	**0.9382**	**0.9875**

tering algorithms on the quantities Precision (*P*), Recall (*R*), F-measure and ARI for algorithms.

Table 2 shows the comparative results of the six clustering algorithms implemented in this experiment on the rat CNS data set T. The values of the quantities Precision (*P*) , Recall (*R*) and F-measure are listed in Table 2 as the criteria to evaluate the respective clustering algorithm. The values of ARI were also considered in our comparison of the clustering algorithms. The best solution in each case has been shown in bold.

Tables 3, 4 and 5 compare the quality of the six clustering algorithms on human fibroblasts serum, yeast cell cycle and HAHrma data sets. The values of the quantities Precision (P), Recall (R),F-measure and ARI were also as the criterion to evaluate the performance of the clustering algorithms.

Table 5 Comparison of the clustering results of various classification methods on the HAHrma data

Methods	*P*	*R*	*F*	ARI
EM	0.8096	0.8471	0.8279	0.8586
EM based on multivariate *t*-distributions	0.7613	0.7819	0.7186	0.8374
FCM	0.7497	0.8302	0.7879	0.7791
K-means	0.6715	0.7117	0.6910	0.6983
Improved EM	**0.9227**	**0.9579**	**0.9141**	**0.9843**

From Tables 2, 3, 4 and 5, it can be seen that our method of improved EM had overall better performance than the other four competitive clustering algorithms, from the precision and recall. For the ARI of the algorithms, the value of our method is higher than other clustering algorithms. It is because that our method can solve the problem of over-reliance on the initialization better.

5 Conclusions

An improved EM method based on multivariate Gaussian mixture models of gene expression data clustering was presented in this paper. The proposed method is used to estimate the coefficients of the multivariate Gaussian mixture models and the weight of the model. In order to find the number of clusters that fits the data best, we have compared and analyzed the results of the experiments by the methods of the AIC,AIC$_3$,AIC$_4$ BIC and QAIC, the QAIC method is used in the paper. To estimate the performance of the improved EM method, we compared our new method with other clustering algorithm based on gene expression data analysis. One artificial and real gene expression data set was selected to implement the experiments. Experimental results show that the improved EM method was better than the other clustering algorithm.

Finally, data normalization is the prerequisite of gene expression data analysis. The model-based method developed here based on the normal assumption but have the problem of model mismatch between the real distribution of gene expression data and the assumption. Therefore, the method may be more sensitive to any departure from normality. Future research should focus on overcoming the model mismatch.

Acknowledgments This work was supported by the National Natural Science Foundation of China (No. 61402204), the Jiangsu Province Natural Science Foundation (Nos. BK20130529, BK2012209), the Research Fund for the Doctoral Program of Higher Education of China (No. 20113227110010), the research foundation for talented scholars, Jiangsu University (No. 14JDG141), and the science and technology program of Zhenjiang city (SH20140110).

References

1. Pirim, H., Ekşioğlu, B., Perkins, A.D., Yüceer, Ç.: Clustering of high throughput gene expression data. Comput. Op. Res. **39**(12), 3046–3061 (2012)
2. Sun, J., Chen, W., Fang, W., Wun, X.J., Xu, W.B.: Gene expression data analysis with the clustering method based on an improved quantum-behaved Particle Swarm Optimization. Eng. Appl. Artif. Intell. **25**(2), 376–391 (2012)
3. Mukhopadhyay, A., Maulik, U.: Towards improving fuzzy clustering using support vector machine: application to gene expression data. Pattern Recognit. **42**(11), 2744–2763 (2009)
4. Zhang, W.F., Liu, C.C., Yan, H.: Clustering of temporal gene expression data by regularized spline regression and an energy based similarity measure. Pattern Recognit. **43**(12), 3969–3976 (2010)
5. Kerr, G., Ruskin, H.J., Crane, M., Doolan, P.: Techniques for clustering gene expression data. Comput. Biol. Med. **38**(3), 283–293 (2008)
6. Seal, S., Komarina, S., Aluru, S.: An optimal hierarchical clustering algorithm for gene expression data. Inform. Process Lett. **93**(3), 143–147 (2005)
7. Szeto, L.K., Wee-Chung Liew, A., Yan, Hong, Tang, Sy-sen: Gene expression data clustering and visualization based on a binary hierarchical clustering framework. J. Visual. Lang. Comput. **14**(4), 341–362 (2003)
8. Chan, Zeke S.H., Lesley Collins, Kasabov, N.: An efficient greedy K-means algorithm for global gene trajectory clustering. Expert Syst. Appl. **30**(1), 137–141 (2006)
9. Lam, Yau King, Tsang, Peter W.M.: Exploratory K-Means: a new simple and efficient algorithm for gene clustering. Appl. Soft Comput. **12**(3), 1149–1157 (2012)
10. Ghouila, Amel, Yahia, Sadok Ben, Malouche, Dhafer, et al.: Application of Multi-SOM clustering approach to macrophage gene expression analysis. Infect. Genet. Evol. **9**(3), 328–336 (2009)
11. Niciura, Simone Cristina Méo, Ibelli, Adriana Mércia Guaratini, Gouveia, Gisele Veneroni: Polymorphism and parent-of-origin effects on gene expression of CAST, leptin and DGAT1 in cattle. Meat Sci. **90**(2), 507–510 (2012)
12. Saha, Indrajit, Maulik, Ujjwal, Bandyopadhyay, Sanghamitra, Plewczynski, Dariusz: Improvement of new automatic differential fuzzy clustering using SVM classifier for microarray analysis. Expert Syst. Appl. **38**(12), 15122–15133 (2011)
13. Zeng, Y.J., Javier, G.F.: A novel HMM-based clustering algorithm for the analysis of gene expression time-course data. Comput. Stat. Data Anal. **50**(9), 2472–2494 (2006)
14. McNicholas, Paul D., Subedi, Sanjeena: Clustering gene expression time course data using mixtures of multivariate t-distributions. J. Stat. Plan. Inference **142**(5), 1114–1127 (2012)
15. Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognit. **28**, 781–793 (1995)
16. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. **97**(458), 611–631 (2002)
17. Yeung, K.Y., Fraley, C., Murua, A., et al.: Model-based clustering and data transformations for gene expression data. Bioinformatics **17**, 977–987 (2001)
18. Qu, Y., Xu, S.Z.: Supervised cluster analysis for microarray data based on multivariate Gaussian mixture. Bioinformatics **20**(12), 1905–1913 (2004)
19. Bouveyron, C., Girard, S., Schmid, C.: High-dimensional data clustering. Comput. Stat. Data Anal. **52**(1), 502–519 (2007)
20. McNicholas, P.D.: Model-based classification using latent Gaussian mixture models. J. Stat. Plan. Inference **140**(5), 1175–1181 (2010)
21. Yao, W.: A note on EM algorithm for mixture models. Stat. Probabil. Lett. **83**(2), 519–526 (2013)
22. Lee, G., Scott, C.: EM algorithms for multivariate Gaussian mixture models with truncated and censored data. Comput. Stat. Data Anal. **56**(9), 2816–2829 (2012)
23. Yang, M., Lai, C., Lin, C.: A robust EM clustering algorithm for Gaussian mixture models. Pattern. Recognit. **45**(11), 3950–3961 (2012)
24. Jacques, J., Preda, C.: Model-based clustering for multivariate functional data. Comput. Stat. Data. Anal. **71**, 92–106 (2014)
25. Maraziotis, I.A.: A semi-supervised fuzzy clustering algorithm applied to gene expression data. Pattern Recognit. **45**(1), 637–648 (2012)
26. Akaike, H.: A new look at statistical model identification. IEEE Trans. Autom. Control. **19**, 716–723 (1974)
27. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 2907–2912 (1978)
28. Lebreton, J.D., Burnham, K.P., Clobert, J., Anderson, D.R.: Modelling survival and testing biological hypotheses using marked animals:a unified approach with case studies. Ecol. Monogr. **62**, 67–118 (1992)
29. McNicholas, P.D., Subedi, S.: Clustering gene expression time course data using mixtures of multivariate t-distributions. J. Stat. Plan. Inference **142**, 1114–1127 (2012)
30. Dembele, D., Kastner, P.: Fuzzy C-means method for clustering microarray data. Bioinformatics **19**, 973–980 (2003)
31. Tavazoie, S., Hughes, J.D., Campbell, M.J., et al.: Systematic determination of genetic network architecture. Nat. Genet. **22**, 281–285 (1999)
32. Wen, X.L., Fuhman, S., Michaels, G.S., et al.: Larger-scale temporal gene expression mapping of central nervous system development. Proc. Natl. Acad. Sci. USA **95**(1), 334–339 (1998)
33. Iyer, V.R., et al.: The transcriptional program in the response of the human fibroblasts to serum. Science **283**, 83–87 (1999)
34. Eisen, M.B., Spellman, P.T., Brown, P.O., et al.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95**(14), 863–14868 (1998)
35. Tavazoie, S., Hughes, J.D., Campbell, M.J., et al.: Systematic determination of genetic network architecture. Nat. Genet. **22**, 218–285 (1999)
36. Weizmann Institute of Science, GeneCards: The Human Gene Compendium. Accessed February 9, 2011. (1996)