ISSN 1682-296X (Print)
ISSN 1682-2978 (Online)

Bio Technology

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Biotechnology 7 (1): 59-65, 2008
ISSN 1682-296X

Autoregressive-Model Based Dynamic Fuzzy Clustering for Time-Course Gene Expression Data

Xu Hong-Lin, Liu Yu-Hong and Wang Shi-Tong
School of Information, Southern Yantze University, Wuxi, 214122, China

Abstract: In this study, a novel algorithm called Dynamic Fuzzy Clustering (DFC) is proposed for clustering time-course gene expression data. The proposed method combines Autoregressive (AR) model and conventional Fuzzy Clustering Algorithm (FCM). Under this approach, a time-course gene expression data can be analyzed as a set of dynamic time series with AR model in order to utilize the important dynamic information more efficiently and the forecast process in AR model can be adjusted using the corresponding fuzzy membership such that better clustering results can be obtained. Experiments performed on a synthetic and two real-world time-course gene expression datasets also indicates that this proposed approach can be more effective than some other conventional clustering algorithms such as FCM and simple dynamic model-based clustering algorithm.

Key words: Autoregressive model, fuzzy clustering; time-course gene expression, dynamic fuzzy clustering, self-relationship

INTRODUCTION

Time-course gene expression data are often defined as a series of values recorded in each time point according to the periodic transformation of cells (Carla and Möller-Levet, 2003). Conventional clustering algorithms, such as k-means clustering (Hartigan and Wang, 1978), Self-Organizing Maps (SOM) (Kohonen, 1997) and hierarchical clustering (Eisen *et al.*, 1998) cannot be appropriate for the time-course gene expression data because that these methods all ignore the important dynamic relationship in the data.

The problem of capturing the dynamic patterns in the particular case of gene expression time-course data has been recently addressed by several authors (Fraley and Raffery, 2002; Möller-Levet *et al.*, 2003). The common idea in these studies is to represent the gene expression time-series as continuous or piecewise continuous curves and then to perform clustering based on the estimated curves. The Fuzzy Short Time-Series (FSTS) clustering method proposed by Möller-Levet *et al.* (2003) is based on the incorporation of a new distance metric, which utilizes a piecewise linear model, into a standard fuzzy clustering scheme. This method is simple and fast but its underlying linear assumption may be an oversimplification in the type of problems encountered in real biological applications. One promising approach is to use a general multivariate Gaussian model to account for the correlation structure (Fraley and Raffery, 2002). However, such a model still ignores the time order of gene expression.

Other popular ways to exploit time dependences is the use of Hidden Markov Models (HMM) and autoregressive model to describe the time-course gene expression. An HMM can be viewed as a stochastic generalization of a finite-state automata and it provides a probabilistic description of temporal dependences. Although HMM have been widely used in many fields, such as speech recognition and digital communications, their application on the clustering of temporal gene expression profiles has not been widespread. One line of work utilizes HMM to devise model-based metrics for time-series. The idea is to generate an HMM for each sequence and then to compute the Log-Likelihood (LL) of each HMM for any of the sequences. This information is used to build a matrix of distances between sequences and then the data is clustered by applying a distance-based clustering method employing such a matrix (Bicego *et al.*, 2003; Smyth, 1997). Alternatively, the LLs can be directly utilized as features for later clustering processing (Panuccio *et al.*, 2002). However, there are several limitations in the LL's calculation, which may degrade the performance of the whole clustering analysis. First, in order to calculate the LLs, one HMM is trained for each sequence. Since reliable training requires long sequences, the reliability of the clustering result may be heavily degraded when dealing with short sequences of gene expression data. Second, since each HMM is trained separately and independently, the model lacks a global view on the overall distribution of the patterns in the data. Finally, this technique assumes that for each gene the

Corresponding Author: Wang Shi-Tong, School of Information, Southern Yangtze University, Wuxi, 214122, China
Tel: 0086-510-85912151 Fax: 0086-510-85912136

transitions between neighboring temporal observations follow the same stationary stochastic process. However, this time-invariance assumption does not usually hold in microarray data, especially when the expression measures are taken non-uniformly in time. The autoregressive model is a linear regression equation which is a more appropriate model for this application. The autoregressive model can make use of the self-relationship in the data by linking the current value of some variable to its value in the previous period and a constant term (Wu *et al*., 2005). However, the autoregressive model is limited by the requirement of time interval and selection of order p. So new errors may be induced in the forecast of time point data followed.

This study introduces a new Dynamic Fuzzy Clustering Method (DFC) based on autoregressive model for time-course gene expression data in which fuzzy partition clustering algorithms and autoregressive model are integrated. Present new method can overcome the disadvantage in the conventional clustering methods, in which self-relationship information is omitted, by introducing the fuzzy membership to adjust the results dynamically when forecasting time point data using autoregressive model. Compared with FCM, our clustering results can be better for time-course gene expression data.

MATERIALS AND METHODS

Fuzzy Clustering Algorithm (FCM): Fuzzy partition clustering algorithms are unsupervised learning methods in pattern recognition field. These algorithms can partition data into different groups automatically according to some distance measurement by machine learning. Due to most of real-life objects dose not have strict attributes, the membership degree chosen as only 1 or 0 in conventional hard partition cannot reflect the real relationship between samples and groups. Apparently, the description that a sample belonging to different groups with their membership degree has obvious advantage. Fuzzy C-Means algorithm (FCM) (Zhao and Xue, 2000) which is proposed by Dunn in 1974 and developed by Bezdek has been widely used to achieve this goal (Zhang and Yu, 2004).

In FCM, let x_i (i = 1, 2, ..., n) be a set consisting of n samples, where, c is the No. of clusters, m_i (i = 1, 2, ..., c) are the centers of ith cluster, $\mu_j(x_i)$ is the membership of the ith sample in the jth cluster. The objective function defined by the membership function can be defined as:

$$J_f = \sum_{j=1}^{c}\sum_{i=1}^{n}\left[\mu_j(x_i)\right]^b \left\|x_i - m_j\right\|^2 \qquad (1)$$

Where, $b(1<b<+\infty)$ represents the so-called fuzzy index (Zhao and Xue, 2000), $\mu_j(x_i) \in (0, 1)$,

$$\sum_{j=1}^{c}\mu_j(x_i) = 1 \qquad (2)$$

Correspondingly, the update rules can be found by minimizing Eq. 1 with Eq. 2, i.e.,

$$m_j = \frac{\sum_{i=1}^{n}\left[\mu_j(x_i)\right]^b x_i}{\sum_{i=1}^{n}\left[\mu_j(x_i)\right]^b},\ j = 1, 2, \ ..., c \qquad (3)$$

$$\mu_j(x_i) = \frac{\left(1/\left\|x_i - m_j\right\|^2\right)^{1/(b-1)}}{\sum_{k=1}^{c}\left(1/\left\|x_i - m_j\right\|^2\right)^{1/(b-1)}},\ i = 1, 2, ..., n,\ j = 1, 2, ..., c \qquad (4)$$

Autoregressive model: Autoregressive model is a more appropriate model for time-course gene expression data (Shu-Xin Zhang and Li-Xin Qi, 2003). In its simplest form, an autoregressive model is a linear regression equation which links the current value of some variable to its value in the previous period and a constant term.

Let x = {x_1, ..., x_m, ..., x_M} be a time series of continuous values with M equally-time-spaced observations. The time series follows an autoregressive model of order p, denoted by AR (p), if the value of the series at time m (m>p) is a linear function of the values of previous p observations plus a term representing error. More formally, an autoregressive model of order p may be written as:

$$x_m = a_1x_{m-1} + \cdots + a_p x_{m-p} + \varepsilon_m,\ m = p+1, ..., M \qquad (5)$$

Where, a_i (i = 1, ..., p) are the autoregressive coefficients and ε_m (m = p+1, ..., M) represents error. This study assumes that the error has a normal distribution independent of time with mean 0 and variance σ^2. Thus the conditional probability distributions of x_m(m = p+1, ..., M), with (x_{m-1}, ..., x_{m-p}) are normal with mean $a_1x_{m-1}+\ldots+a_px_{m-p}$ and varianceσ^2, in terms of (Wu *et al*., 2005). The probability distribution of x_m can be defined as:

$$p(x_m \mid \sigma^2, x_{m-1}, \cdots, x_{m-p}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{-(x_m - a_1x_{m-1} - \cdots - a_px_{m-p})^2}{2\sigma^2}, m = p+1, ..., M \qquad (6)$$

So, the log-likelihood function that time series x is generated by an autoregressive model of order p with coefficients a_i(i = 1, ..., p) can be written directly as follows:

$$
\begin{aligned}
&L\left(x \mid a, \sigma^2, x_1, \cdots, x_p\right) \\
&= \log p(x_1, \cdots, x_p) + \sum_{m=p+1}^{M} \log p(x_m \mid \sigma^2, x_{m-1}, \cdots, x_{m-p}) \\
&= \log p(x_1, \cdots, x_p) - \frac{M-p}{2} \log(2\pi\sigma^2) \\
&\quad - \frac{1}{2\sigma_0^2} \sum_{m=p+1}^{M} (x_m - a_1 x_{m-1} - \cdots - a_p x_{m-p})^2
\end{aligned} \tag{7}
$$

Dynamic Fuzzy Clustering algorithm (DFC): In this subsection the Dynamic Fuzzy Clustering algorithm (DFC) is proposed based on the characteristic of the fuzzy membership function in the above fuzzy clustering algorithm and the above autoregressive model to combine the advantage of the two methods.

In FCM, the objective function is defined based on the distances between samples to measure the uncomparability. After autoregressive model is introduced, the distance in the objective function can be replaced with the maximum likelihood in order to measure the comparability in the samples. Thus, the objective function expression in algorithm DFC is designed as;

$$
J_d{}' = \sum_{j=1}^{c} \sum_{i=1}^{n} \left[\mu_j(x_i)\right]^b L(x_i \mid a_j, \sigma_j^2, x_{i1}, \cdots, x_{ip}) \tag{8}
$$

Accordingly, our purpose is to get the maximum of the above objective function $J_d{}'$.

By assuming the first p observations have a multivariate normal distribution with mean $u = (u_1, \ldots, u_M)$ and covariance matrix $\Sigma = \sigma_0 I_p$ ($I\ _p$represents the p*p identity matrix). The maximum likelihood distribution in the objective function can be defined as (Wu *et al.*, 2005):

$$
p(x_1, \cdots, x_p \mid \sigma^2, u) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^p \exp\left(\frac{-\sum_{i=1}^{p} (x_i - u_i)^2}{2\sigma_0^2}\right) \tag{9}
$$

Substitute Eq. 9 into Eq. 7 , we have

$$
\begin{aligned}
&L(x \mid u, \sigma_0^2, a, \sigma^2) \\
&= \log p(x \mid u, \sigma_0^2, a, \sigma^2) \\
&= -\frac{M-p}{2} \log(2\pi\sigma^2) \\
&\quad - \frac{p}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma_0^2} \sum_{i=1}^{p} (x_i - u_i)^2 \\
&\quad - \frac{1}{2\sigma_0^2} \sum_{m=p+1}^{M} (x_m - a_1 x_{m-1} - \cdots - a_p x_{m-p})^2
\end{aligned} \tag{10}
$$

As in Wu *et al.* (2005), let x_0 be the vector $[x_1, \ldots, x_m]^T$, y be the vector $[x_{p+1}, \ldots, x_M]^T$ and X be the (M-p*p regression matrix whose mth row is $(x_{m-1}, \ldots, x_{m-p})$ for m = p+1, …, M. By substituting into Eq. 10, we have:

$$
\begin{aligned}
&L(x \mid u, \sigma_0^2, a, \sigma^2) \\
&= \log p(x \mid u, \sigma_0^2, a, \sigma^2) \\
&= -\frac{M-p}{2} \log(2\pi\sigma^2) + \frac{p}{2} \log(2\pi\sigma^2) \\
&\quad - \frac{1}{2\sigma_0^2} (x_0 - u)^T (x_0 - u) \\
&\quad - \frac{1}{2\sigma_0^2} (y - Xa)^T (y - Xa)
\end{aligned} \tag{11}
$$

For a time-course gene expression data, each gene sample can be viewed as a single time series. Given a set of the observed time-course gene expression data $X = \{x_1, \ldots, x_i, \ldots, x_n\}$, where, $x_i (i = 1, \ldots, n)$ represents a time series. If the dataset X contains K clusters and the optimal partition of X can be assumed to be $C = \{C_{1, \ldots,} C_{, \ldots,} C_K\}$, then a gene sample x_i should be put in cluster j, if it can make J_{ij} to reach its maximum.

$$
J_{ij} = \left[\mu_j(x_i)\right]^b L(x_i \mid u_j, \sigma_{0j}^2, a_j, \sigma_j^2) \tag{12}
$$

Where, b∈(1,3) is used in this study. In terms of Eq. 8, we have

$$
\begin{aligned}
J_d &= \sum_{j=1}^{c} \sum_{i=1}^{n} \left[\mu_j(x_i)\right]^b L(x_i \mid u_j, \sigma_{0j}^2, a_j, \sigma_j^2) \\
&= \left[\mu_j(x_i)\right]^b \Bigg(-\sum_{j=1}^{K} |C_j| \left(\frac{M-p}{2} \log(2\pi\sigma_j{}^2) + \frac{p}{2} \log(2\pi\sigma_{0j}^{\;2})\right) \\
&\quad - \sum_{j=1}^{K} \frac{1}{2\sigma_{0j}^2} \sum_{x \in C_j} (x_0 - u_j)^T (x_0 - u_j) \\
&\quad - \sum_{j=1}^{K} \frac{1}{2\sigma_j^2} \sum_{x \in C_j} (y - Xa_j)^T (y - Xa_j)\Bigg)
\end{aligned} \tag{13}
$$

Where:

$$
\sum_{j=1}^{c} \mu_j(x_i) = 1
$$

$|C_j|$ represents the number of time series in cluster C_j,

$$
\sum_{j=1}^{K} |C_j| = N \cdot
$$

(1) Divide dataset into its initial partition with initial K clusters by using K-means clustering algorithm.
(2) Estimate $\mu_j(x_i), u_j, \hat{\sigma}_{0k}^2, \hat{a}_k, \hat{\sigma}_k^2$ using Eq. 14-18.
(3) Assign sample x_i to cluster j for which J_{ij} is maximal, i = 1, 2, ..., n, $j_j \in \{1, 2, ..., K\}$.
(4) Stop if J_d becomes a given threshold or the maximum difference of J_{ij} in this consecutive partitions is less than another given threshold. Otherwise go to Eq. 2.

Fig. 1: Algorithm for DFC

(1) Randomly divide the original dataset into two non-overlapping sets, a learning set L and a test set T.
(2) Apply the evaluated method to the learning set L to obtain a partition P_L.
(3) Construct a predictor C using the cluster labels from the obtained partition P_L.
(4) Apply the predictor C to the test set T to get the predicted partition P_T'.
(5) Apply the evaluated method to the test set T to obtain a partition P_T.
(6) Calculate ARI for partitions P_T' and P_T.

Fig. 2: The procedure for estimating ARI

By maximizing J_d we can obtain the following update rules about $\mu_j(x_i), u_j, \hat{\sigma}_{0k}^2, \hat{a}_k, \hat{\sigma}_k^2$

$$\mu_j(x_i) = \frac{\left(1/\|x_i - u_j\|^2\right)^{1/(b-1)}}{\sum_{k=1}^{c}\left(1/\|x_i - u_j\|^2\right)^{1/(b-1)}} \tag{14}$$

$$u_j = \frac{\sum_{i=1}^{n}\left[\mu_j(x_i)\right]^b x_i}{\sum_{i=1}^{n}\left[\mu_j(x_i)\right]^b} \tag{15}$$

$$\hat{\sigma}_{0k}^2 = \frac{\sum_{i=1}^{n}\left[\mu_j(x_i)\right]^b (x_0 - \hat{u}_k)^T(x_0 - \hat{u}_k)}{p\sum_{i=1}^{n}\left[\mu_j(x_i)\right]^b} \tag{16}$$

$$\hat{a}_k = (\sum_{i=1}^{n}\left[\mu_j(x_i)\right]^b X^T X)^{-1}\sum_{i=1}^{n}\left[\mu_j(x_i)\right]^b X^T y \tag{17}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^{n}\left[\mu_j(x_i)\right]^b (y - X\hat{a}_k)^T(y - X\hat{a}_k)}{|C_k|(M-p)\sum_{i=1}^{n}\left[\mu_j(x_i)\right]^b} \tag{18}$$

Where, k = 1, ..., K.

According to the above update rules, the proposed algorithm DFC can be summarized as Fig. 1.

Validity measures: The validity process explores whether the clustering algorithm with the specified parameters (number of clusters, similarity measure, model, etc.) can identify the underlying patterns of the considered dataset (Hoppner *et al*., 1999). In order to solve this problem, several cluster quality or validity measures have been proposed in literature. For gene expression data, the best and simplest method is comparing the clustering result with the correct labels. But it is not possible to get all the labels of the gene expression data because of the noises in the real gene expression data and some important information can be achieved by existing techniques. Since a clustering result can be considered as a partition of objects into a number of groups, for evaluating a clustering method it is necessary to define a measure of agreement between two partitions of the same dataset. We uses the Adjusted Rand Index (ARI) (Dudoit *et al*., 2002) to evaluate the quality of the clustering results in this paper.

Consider two partitions of N objects, the R-cluster partition U = {u_1, ..., u_r} and the S-cluster partition V = {v_1,...,v_s}. One may construct a contingency matrix, where, n_{ij} denotes the number of objects that are both in cluster u_i and v_j, i = 1,..., r, j = 1, ..., s. Let:

$$n_i = \sum_{j=1}^{s} n_{ij}$$

and

$$n_j = \sum_{i=1}^{r} n_{ij}$$

denote the sum of row i(i = 1, ..., r) and the sum of column j(j = 1, ..., s) in the contingency matrix, respectively and let:

$$Z = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}^2$$

and $V = \binom{N}{2} = N\ (N-1)/2$ (the number of pairs of N objects). Based on the contingency matrix of two partitions, the ARI is defined as (Dudoit *et al*., 2002);

$$ARI = \frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\binom{n_{ij}}{2} - \frac{1}{V}\sum_{i=1}^{r}\binom{n_{i.}}{2}\sum_{j=1}^{s}\binom{n_{.j}}{2}}{\frac{1}{2}[\sum_{i=1}^{r}\binom{n_{i.}}{2} + \sum_{j=1}^{s}\binom{n_{.j}}{2}] - \frac{1}{V}\sum_{i=1}^{r}\binom{n_{i.}}{2}\sum_{j=1}^{s}\binom{n_{.j}}{2}} \tag{19}$$

The ARI is an adjusted Rand index in that its expected value is 1 when they matched perfectly and 0 when the two partitions are selected at random.

The procedure for estimating ARI can be described in Fig. 2.

In order to make the results fair in this study, we run the above procedure 10 times and evaluate the quality of the obtained clustering results by taking the average value of the obtained ARIs for the given number of clusters K. ARI ranges from -1 to 1 and so does AARI. Accordingly, the larger the AARI, the better the quality of the obtained clustering results is.

RESULTS AND DISCUSSION

Synthetic dataset: The synthetic dataset can be generated by the sine function to model cyclic behavior of genes in (Yeung *et al.*, 2001). Let x_{ij} be the simulated expression level of gene i at time point j in the dataset and be modeled by $x_{ij} = \delta_j + \lambda_j * (\alpha_i + \beta_j \phi(i,j))$, where, $\phi(i,j) = \sin(2\pi j/8 - \omega_{k(i)} + \varepsilon)$. α_i represents the average expression level of gene i, which is chosen according to the standard normal distribution. β_j is the amplitude control for gene i, which is chosen according to the normal distribution with mean 3 and standard deviation 0.5. λ_j is the amplitude control at time j, which is chosen according to the normal distribution with mean 3 and standard deviation 0.5. δ_j represents additive experimental error at time point j, which is chosen according to the normal distribution with mean 0 and standard deviation 2. $\phi(i,j)$ models the cyclic behavior of genes. Each cycle is assumed to span eight time points. Different clusters are represented by different phase shifts and $\omega_{k(i)}$ represents a phase shift for gene i in cluster k, which is chosen according to the uniform distribution on interval (0,2 π). The random variable ε represents the noise of gene synchronization, which is chosen according to the standard normal distribution. Using the model above, a synthetic dataset is generated consisting of expression levels of 500 genes at 24 equally spaced time points.

The dataset generated above can be descript in Fig. 3, in which the x-axis represents the observational values of one sample and the y-axis represents the identification code of each sample. The curve of each sample observed at different time points are described in the same coordinate.

Real dataset: The real datasets of time-course gene expression data and the distribution figures used in the experiment can be achieved from http://genome-www.stanford.edu/SVD. The dataset comes from the circulation experiment for the yeast gene and consists of expression level of cell-cycle regulated genes at equally-spaced time points. The 396 samples in original data Yeast (Fig. 2) and the 1000 samples of standardization data Sort_Elutriation (Fig. 3) are chosen to conduct the experiment (Alter *et al.*, 2001). Each sample consists of 18 and 14 observation values at equally-spaced time points, respectively.

The real dataset used in the experiment can be descript in Fig. 4 and 5, in which the x-axis represents the observational values of one sample and the y-axis represents the identification code of each sample. The clusters of the time-course gene expression data can be observed in the figures immediately, because the distributions of different clusters are not in accordance with the identification code of each sample.

The original data Yeast has been standardized before experiment in order to facilitate the assumption in the study that data should obey the normal distribution with mean 0 and standard deviation 1.

Synthetic dataset experiment: In this experiment, the synthetic dataset has been analyzed using algorithm DFC, FCM and the dynamic model-based clustering algorithm (Wu *et al.*, 2005) with the order p be 1, 2 and 3, respectively. We partition the synthetic dataset with 500 samples into 2 to 10 clusters. As reported in (Wu *et al.*, 2005), the best results of the dynamic model-based clustering algorithm can be achieved with the order p = 1 and the average results of FCM in 10 times are also presented in. Figure 6 shows the contrastive effect about the clustering results obtained these three algorithms, in which the x-axis represents the number of clusters and the y-axis represents the corresponding AARI.

By comparing the obtained optimal results for the synthetic dataset, we can see that the proposed algorithm DFC is better than the dynamic model-based clustering algorithm and FCM and DFC reaches its best results when p = 2, which is different from p = 1 in (Wu *et al.*, 2005).

Real dataset experiment: In this example we take the real dataset yeast. The same execution strategy about the DFC, FCM and the dynamic model-based clustering algorithm is adopted. Figure 7 shows the obtained contrastive effects on these three algorithms.

Obviously, the proposed algorithm DFC and the dynamic model-based clustering algorithm have better quality of clustering than FCM who has ignored the self-relationship on the time points for partial Yeast dataset. The self-relationship in the time-course gene expression data has been essentially testified before the experiment.

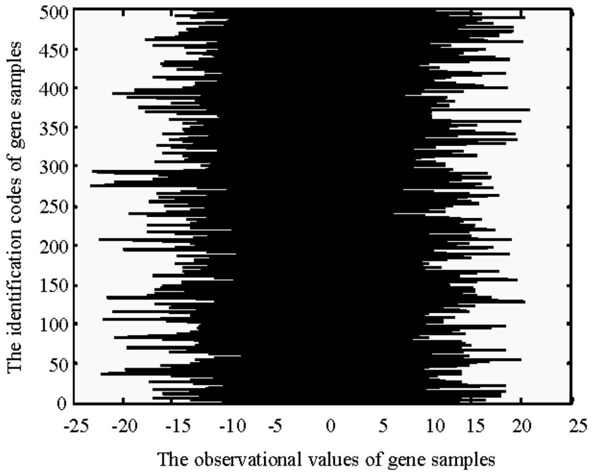

Fig. 3: The distribution of the synthetic dataset

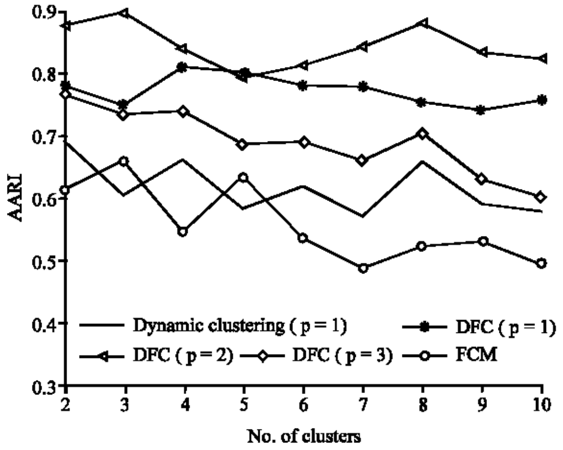

Fig. 6: The contrastive effect for the synthetic dataset

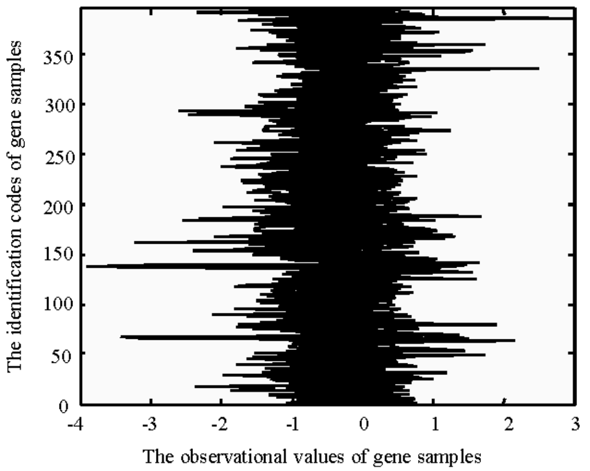

Fig. 4: The distribution of a part of yeast

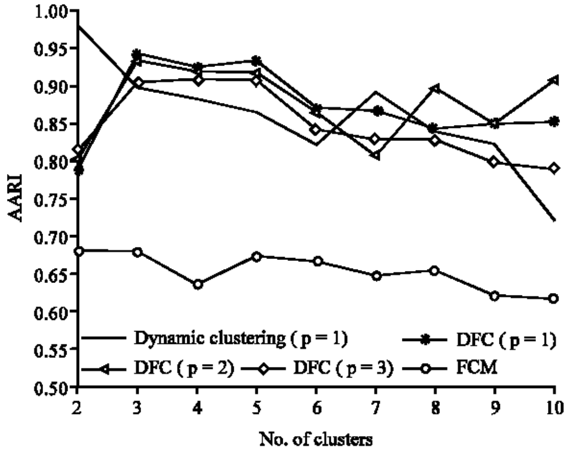

Fig. 7: The contrastive effect for the yeast dataset

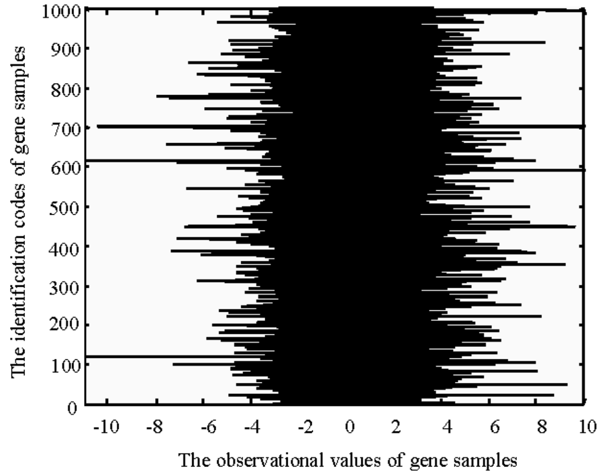

Fig. 5: The distribution of a part of sort_elutriation

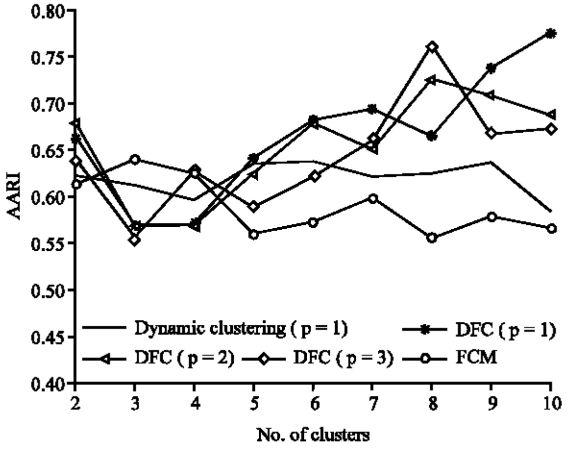

Fig. 8: The contrastive effect for the sort_elutriation dataset

As a whole, the clustering results from the proposed algorithm DFC are better than those from the dynamic clustering algorithm and the clustering quality is not strongly affected by the order p of the autoregressive model.

Also, we tested another real dataset Sort_Elutriation for these three algorithms. Figure 8 shows the obtained results. Obviously, it indicates that the same conclusion still holds.

CONCLUSIONS

Recently, the time-course gene expression data are employed widely in bioinformatics and the conventional clustering algorithms are not fit for such data. It is necessary for us to suggest a mixed method to fit the requirement of such analysis.

We explore the proposed clustering algorithm DFC, which incorporates fuzzy clustering with a autoregressive model well in this study. Our experimental results demonstrate that the proposed algorithm DFC is better than FCM and the dynamic model-based clustering algorithm. Another obvious feature of DFC exist in that its clustering results seem not be affect by the order p. In other words, we can adjust p to avoid the localization issue existing in the dynamic model-based clustering algorithm in (Wu *et al*., 2005). Future study includes exploring its robust version of DFC such that it can be very fit for noisy time-course gene expression data.

ACKNOWLEDGMENTS

This research is supported by National 973 Key project (Grant No. 2006CB705700), National 863 Project (Grant No. 2006AA10Z313), National Science Foundation of China (grant No.60225015), JiangSu Natural Science Foundation (grant No. BK2003017), New_century Outstanding Young Scholar Grant of Ministry of Education of China (NCET-04-0496), National KeySoft Lab. at Nanjing University, The Key Lab. of Computer Science at Insitute of Software, CAS, China, The Key Lab. of Computer Information Technologies at JiangSu Province, the 2004 key project of Ministry of Education of China and National Key Lab. Of Pattern Recognition at Institute of Automation, CAS, China.

REFERENCES

Alter, O., P.O. Brown and D. Botstein, 2001. Processing and Modeling Genome-Wide Expression Data Using Singular Value Decomposition. In: Optical Technologies and Informatics. Bittner, M.L., Y. Chen and A.N. Dorsel *et al*. (Eds.), Bellingham, pp: 171-186.

Bicego, M., V. Murino and M. Figueiredo, 2003. Similarity-based clustering of sequences using Hidden Markov Models. MLDM, LNAI, 2734: 86-95.

Carla, S. and Möller-Levet, 2003. Clustering of gene expression time-series. Manchester M60 1QD, UK: UMIST.

Dudoit, S. and J. Fridlyland, 2002. A prediction-based re-sampling method for estimating the number of clustering in a dataset. Genome. Biol., 3: 0036.1-0036.21.

Eisen, M.B. *et al*., 1998. Cluster analysis and display of genome-wide expression patterns proc. Natl. Acad. Sci., USA., 95: 14863-14868.

Fraley, C. and A.E. Raftery, 2002. Model-based clustering, discriminant analysis and density estimation. Am. Stat. Assoc., 97: 611-631 .

Hartigan, J.A. and M.A. Wong, 1978. A K-means Clustering Algorithm. Applied Stat., 28: 100-108.

Hoppner *et al*., 1999. Fuzzy Cluster Analysis. Jonh Wiley and Sons, Ltd., New York.

Kohonen, T., 1997. Self-Organizing Maps. Springer, New York.

Moller-Levet, C.S., F. Klawonn, K.H. Cho and O. Wolkenhauer, 2003. Clustering of unevenly sampled gene expression time-series data. Department of Computer Science, University of Rostock, Rostock, Germany.

Panuccio, A., M. Bicego and V. Murino, 2002. A Hidden Markov Model-Based Approach to Sequential Data Clustering. Lecture Notes in Computer Science. Berlin: Springer, pp: 734-742.

Shu-Xin Zhang and Li-Xin Qi, 2003. Concise course for time-course analyse, in Chinese. Tsinghua University Press, Beijing, pp: 24-58.

Smyth, P., 1997. Clustering sequences with hidden Markov Models. Adv. Neural Inform. Proc. Syst., 9: 648-654.

Wu, F.X., W.J. Zhang, J. Anthony and Kusalik, 2005. Dynamic model-based clustering for time-course gene expression data. J. Bioinformat. Comput. Biol., 3: 821-836.

Yeung, K.Y. *et al*., 2001. Model-based clustering and data trans-formations for gene expression data. Bioinformatics, 17: 977-987.

Zhang, M. and J. Yu, 2004. Fuzzy partitional clustering algorithms. J. Software, Chinese, 15: 858-868.

Zhao-Qi Bian and Xue-Gong Zhang, 2000. Patten Recognition, in Chinese. Tsinghua University Press, Beijing, pp: 273-283.