한국 퍼지 및 지능시스템 학회2007, Vol. 17, No. 3, pp. 397-402

Cluster Analysis of Incomplete Microarray Data with Fuzzy Clustering

Dae-Won Kim

School of Computer Science and Engineering, Chung-Ang University, Seoul 156-756, Korea

Abstract

In this paper, we present a method for clustering incomplete Microarray data using alternating optimization in which a prior imputation method is not required. To reduce the influence of imputation in preprocessing, we take an alternative optimization approach to find better estimates during iterative clustering process. This method improves the estimates of missing values by exploiting the cluster information such as cluster centroids and all available non-missing values in each iteration. The clustering results of the proposed method are more significantly relevant to the biological gene annotations than those of other methods, indicating its effectiveness and potential for clustering incomplete gene expression data.

Key words : Bioinformatics, fuzzy clustering, Microarray, missing value

1. Introduction

DNA microarray technology has allowed for the monitoring of the transcript abundance of thousand of genes in parallel under a variety of conditions. Since Eisen et al. first used the hierarchical clustering method to find groups of coexpressed genes [1], numerous methods have been studied for clustering gene expression data: self-organizing map [2], k-means clustering [3], graph-theoretic approach [4], mutual information approach [5], fuzzy c-means clustering [6], diametrical clustering [7], quantum clustering with singular value decomposition [8], bagged clustering [9], CLICK [10]. However, the analysis results obtained by clustering methods will be influenced by missing values in microarray experiments, and thus it is not always possible to correctly analyze the clustering results due to the incompleteness of data sets. The problem of missing values have various causes, including dust or scratches on the slide, image corruption, spotting problems [11, 12]. Ouyang et al. [13] pointed out that most of the microarray experiments contain some missing entries and more than 90 % of rows (genes) are affected.

To convert incomplete microarray experiments to a complete data matrix that is required as an input for a clustering method, we must handle the missing values before calculating clustering. To this end, typically we have either removed the genes with missing values or estimated the missing values using an imputation prior to cluster analysis. Of the methods proposed, several imputation methods have been demonstrating their effectiveness in building the complete matrix of clustering: missing values are replaced by zeros [14] or by the average expression value over the row (gene). Troyanskaya et al. [11] presented two correlation-based imputation methods: a singular value decomposition based method (SVDimpute) and weighted K-nearest neighbors (KNNimpute). Besides, a classical Expectation Maximization approach (EMimpute) exploits the maximum likelihood of the convariance of the data for estimating the missing values [12, 13]. However, a common limitation of existing approaches for clustering incomplete microarray data is that the estimation of missing values must be calculated in the preprocessing step of clustering. Once the estimates are found, they are not changed during the subsequent steps of clustering. Thus badly estimated missing values during data preprocessing can deteriorate the quality and reliability of clustering results, and therefore drive the clustering method to fall into a local minimum; it prevents missing values from being imputed by better estimates during the iterative clustering process. To minimize the influence of bad imputation, in the present study we developed a method for clustering incomplete microarray data, which iteratively finds better estimates of missing values during clustering process. Incomplete gene expression data is used as an input without any prior imputation. This method preserves the uncertainty inherent in the missing values for longer before final decisions are made, and is therefore less prone to falling into local optima in comparison to conventional imputation-based clustering methods. To achieve this, a method for measuring the distance between a cluster centroid and a row (a gene with missing values) is proposed, along with a method for

접수일자 : 2006년 11월 25일
운료일자 : 2007년 3월 15일

estimating the missing attributes using all available information in each iteration.

2. The Proposed Method

The objective of the proposed method is to classify a data set $X = \{x_1, x_2, \ldots, x_n\}$ in p-dimensional space into k disjoint and homogeneous clusters represented as $C = \{C_1, C_2, \ldots, C_k\}$. Here each data point $x_j = [x_{j1}, x_{j2}, \ldots, x_{jp}]$ $(1 \le j \le n)$ is the expression vector of the j-th gene over p-different environmental conditions or samples. A data point with some missing conditions or samples is referred to as an incomplete gene; a gene x_j is incomplete if x_{jl} is missing for $\exists 1 \le l \le p$, i.e., an incomplete gene $x_1 = [0.75, 0.73, ?, 0.21]$ where x_{13} is missing. A gene expression data set X is referred to as an incomplete data set if X contains at least one incomplete gene expression vector.

To find better estimates of missing values and improve the clustering result during iterative clustering process, in each iteration we exploit the information of current clusters such as cluster centroids and all available non-missing values. For example, a missing value x_{jl} is estimated using the corresponding l-th attribute value of the cluster centroid to which x_j is closest in each iteration. To improve the estimates during each iteration, the proposed method attempts to optimize the objective function with respect to the missing values, which is often referred to as the alternating optimization (AO) scheme. The objective of the proposed method is obtained by minimizing the function J_m:

$$\min \left\{ J_m(U, V) = \sum_{i=1}^{k} \sum_{j=1}^{n} (\mu_{ij})^m D_{ij} \right\} \tag{1}$$

where

$$D_{ij} = \|x_j - v_i\|^2 \tag{2}$$

is the distance between x_j and v_i,

$$V = [v_1, v_2, \ldots, v_k] \tag{3}$$

is a vector of the centroids of the clusters $C_1, C_2, \ldots, C_k$,

$$U = [\mu_{ij}] = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{k1} & \mu_{k2} & \cdots & \mu_{kn} \end{bmatrix} \tag{4}$$

is a fuzzy partition matrix of X satisfying the following constraints,

$$\begin{aligned} &\mu_{ij} \in [0, 1], \quad 1 \le i \le k,\ 1 \le j \le n, \\ &\textstyle\sum_{i=1}^{k} \mu_{ij} = 1, \quad 1 \le j \le n, \\ &0 < \textstyle\sum_{j=1}^{n} \mu_{ij} < n, \quad 1 \le i \le k. \end{aligned} \tag{5}$$

and

$$m \in [1, \infty) \tag{6}$$

is a weighting exponent that controls the membership degree μ_{ij} of each data point x_j to the cluster C_i. As $m \to 1$, J_1 produces a hard partition where $\mu_{ij} \in \{0, 1\}$. As m approaches infinity, J_∞ produces a maximum fuzzy partition where $\mu_{ij} = 1/k$. This fuzzy k-means-type approach has advantages of differentiating how closely a gene belongs to each cluster [6] and being robust to the noise in microarray data [15] because it makes soft decisions in each iteration through the use of membership functions.

Under this formulation, missing values are regarded as optimization parameters over which the functional J_m is minimized. To obtain a feasible solution by minimizing Eq. 1, the distance D_{ij} between an incomplete gene x_j and a cluster centroid v_i must be calculated as:

$$D_{ij} = \frac{p}{\sum_{l=1}^{p} \omega_{jl}} \sum_{l=1}^{p} (x_{jl} - v_{il})^2 \omega_{jl} \tag{7}$$

where

$$\omega_{jl} = \begin{cases} 1 & \text{if } x_{jl} \text{ is non-missing} \\ 1 - \exp(-t/\tau) & \text{if } x_{jl} \text{ is missing} \end{cases} \tag{8}$$

We differentiate the missing attribute values from the non-missing values in calculating D_{ij}. The fraction part in Eq. 7 indicates that D_{ij} is inversely proportional to the number of non-missing attributes used where p is the number of attributes. ω_{jl} indicates the confidence degree with which l-th attribute of x_j contributes to D_{ij}; specifically, $\omega_{jl} = 1$ if x_{jl} is non-missing and $0 \le \omega_{jl} < 1$ otherwise. The exponential decay, $\exp(-t/\tau)$, represents the reciprocal of the influence of the missing attribute x_{jl} on discrete time t where τ is a time constant. At the initial iteration ($t = 0$), w_{jl} has a value of 0. As time t (i.e., the number of iterations) increases, the exponent part decreases fast, and thus w_{jl} approaches 1. Let us consider an incomplete data point $x_1 = [0.75, 0.73, ?, 0.21]$ where initially x_{13} is missing. Suppose that x_{13} is estimated as a value of 0.52 after two iterations; then x_1 has a vector of $[0.75, 0.73, 0.52, 0.21]$. From this vector, we see that x_{13} participates in calculating the distance to cluster centroids less than the other three values because it is now being estimated. Besides, the influence of x_{13} to D_{i1} is increased as the iteration continues because its estimate is improved by an iterative optimization.

Using D_{ij} in Eq. 7, the saddle point of J_m is obtained by considering the constraint Eq. 5 as the Lagrange multipliers:

$$\begin{aligned} &\nabla J_m(U, V, \lambda) \\ &\quad = \sum_{i=1}^{k} \sum_{j=1}^{n} (\mu_{ij})^m D_{ij} + \sum_{j=1}^{n} \lambda_j \left[\sum_{i=1}^{k} \mu_{ij} - 1 \right] \end{aligned} \tag{9}$$

and by setting $\nabla J_m = 0$. If $D_{ij} > 0$ for all i, j and $m > 1$, then (U, V) may minimize J_m only if,

$$\mu_{ij} = \left[\sum_{z=1}^{k} \left(\frac{D_{ij}}{D_{iz}}\right)^{2/(m-1)}\right]^{-1}, \quad 1 \le i \le k; 1 \le j \le n, \tag{10}$$

and

$$v_i = \frac{\sum_{j=1}^{n} (\mu_{ij})^m x_j}{\sum_{j=1}^{n} (\mu_{ij})^m}, \quad 1 \le i \le k. \tag{11}$$

This solution also satisfies the remaining constraints of Eq. 5. Along with the optimization of the cluster centroids and membership degrees in Eqs. 10 and 11, missing values are optimized during each iteration to minimize the functional J_m. In this study, we optimize the missing values by minimizing the function $J(x_j)$ presented by [16]:

$$J(x_j) = \sum_{i=1}^{k} (\mu_{ij})^m \|x_j - v_i\|_A^2 \tag{12}$$

By setting $\nabla J = 0$ with respect to the missing attributes of x_j, a missing value x_{jl} is calculated as:

$$x_{jl} = \frac{\sum_{i=1}^{k} (\mu_{ij})^m v_{il}}{\sum_{i=1}^{k} (\mu_{ij})^m}, \quad 1 \le i \le k. \tag{13}$$

By Eq. 13, x_{jl} is estimated by the weighted mean of all cluster centroids in each iteration. At the initial iteration, x_{jl} is initialized with the corresponding attribute of the cluster centroid to which x_j has the highest membership degree.

This method iteratively improves a sequence of sets of clusters until no further improvement in $J_m(U, V)$ is possible. It loops through the estimates for $V_t \to U_{t+1} \to V_{t+1}$ and terminates on $\|V_{t+1} - V_t\| \le \epsilon$. Equivalently, the initialization of the algorithm can be done on U_0, and the iterates become $U_t \to V_{t+1} \to U_{t+1}$, with the termination criterion $\|U_{t+1} - U_t\| \le \epsilon$. This way of alternating optimization using membership computation makes the present method be less prone to falling into local minima than conventional clustering methods.

3. Experimental results

To test the effectiveness with which the proposed method clusters incomplete microarray data, we applied the proposed method and conventional imputation-based clustering methods to the well-known yeast sporulation data set of Chu et al. [17]., and compared the performance of each method. The Chu data set consists of the expression levels of the yeast genes measured at seven time points during sporulation. Of the 6,116 gene expressions analyzed by Eisen et al. [1], 3,020 significant genes obtained through two-fold change were used. The data set was preprocessed for the test by randomly removing 5–25% (5, 10, 15, and 25) of the data in order to create incomplete matrices.

To cluster these incomplete data sets with conventional methods, we first estimated the missing values using the widely used KNNimpute [11] and EMimpute [12, 13]. For the estimated matrices yielded by each imputation method, we used EXPANDER [10] software that implements many clustering methods, of which we investigated the results of the k-means method. In these experiments, the parameters used in the proposed method were $\epsilon = 0.001$,$m = 3.0$, and the KNNimpute was tested with $K = 20$; these values were chosen because they have been overwhelmingly favored in previous studies [11]. In the tests reported here, we analyzed the performance of each approach at the number of clusters of $k = 5$.

3.1 Comparison of clustering performance

To show the performance of imputation, most of imputation methods proposed to date, including KNNimpute and EMimpute, have examined the the root mean squared error (RMSE) between the true values and the imputed values. However, as Bo et al. pointed out [12], the RMSE is limited to study the impact of missing value imputation on cluster analysis. To make this study more informative regarding how large an impact the imputation method has on cluster analysis, in the present work the clustering results obtained using the alternative imputations were evaluated by comparing gene annotations using the z-score [19, 12]. The z-score is calculated by investigating the relation between a clustering result and the functional annotation of the genes in the cluster. To achieve this, this score uses the *Saccharomyces* Genome Database (SGD) annotation of the yeast genes, along with the gene ontology developed by the Gene Ontology Consortium [20, 21]. A higher score of z indicates that genes are better clustered by function, indicating a more biologically significant clustering result.

Table 1 shows the clustering performance of the KNNimpute/EMimpute-based clustering methods and proposed method for the yeast sporulation data set. The z-score of each method is listed with respect to the percentages of missing values (5-25%). The k-means method using KNNimpute gave z-scores from 38.5% to 49.1%. The z-scores of the k-means using EMimpute were ranged from 38.9 to 49.5. In comparison to these methods, it is evident that the proposed clustering method shows markedly better performance, giving z-scores of more than 48.5 for all missing values; it provided significantly better clustering performance than other methods, giving $z = 55.0$ at 5% and $z = 51.9$ at 10%.

The proposed method has a parameter, τ, a time constant. We investigated the influence of the choice of τ on the clustering results in Table. 2. The proposed method

Table 1: Comparison of the clustering performance (z-scores) of the imputation-based clustering methods and the proposed method for the yeast sporulation data set of [17]. The k-means method was tested on the data obtained by KNNimpute and EMimpute.

Method \ %missing	5%	10%	15%	25%
KNNimpute+k-means	47.90	49.10	45.70	38.50
EMimpute+k-means	49.50	44.60	45.40	38.90
Proposed (τ=100)	55.06	51.97	50.87	48.59

Table 2: Comparison of the clustering performance (z-scores) of the proposed method for different τ values

Method \ %missing	5%	10%	15%	25%
Proposed (τ=10)	54.70	50.47	50.42	46.73
Proposed (τ=50)	52.72	51.98	51.36	46.46
Proposed (τ=100)	55.06	51.97	50.87	48.59
Proposed (τ=500)	54.88	53.40	48.57	46.94

with different $\tau = 10, 50, 100$, and 500 values showed similar performances over 5-15% missing values. We observe that the performance of the proposed is less insensitive to the choice of τ.

Table 3 shows the comparison of RMSE of the imputation methods and the proposed method for the incomplete data sets. From the comparison results for the sporulation data, the KNNimpute gave better RMSE at lower missing values whereas the proposed method gave better RMSE at higher missing values. The EMimpute shows the most ineffective of the methods considered. We see that RMSE of each method increases as the missing value increases. However, as mentioned in earlier, RMSE is limited to investigate the impact of the both imputation and clustering together, indicating that better RMSE does not necessarily lead to better z-scores.

The results of the comparison tests indicate that the proposed method gave markedly better clustering performance than the other imputation-based methods considered, highlighting the effectiveness and potential of the proposed method.

3.2 Functional enrichment

To investigate the functional enrichment of the clustered genes, we applied the proposed method to the yeast cell-cycle data set of Cho et al. [18], which has been extensively studied to reveal the gene functions of the yeast. The Cho data set contains the expression profiles of 6,200 yeast genes measured at 17 time points over two complete cell cycles. We used the same selection of 2,945 genes made by Tavazoie et al. [3] in which the data for two time points (90 and 100 min) were removed.

The enriched functional categories for each cluster obtained by the proposed method on the yeast cell-cycle data set are listed in Table 4. The enrichment of each GO category in each of the clusters was calculated by its p-value. To compute the p-value, we employed the hypergeometric distribution that was used by [3] and [6] in order to obtain the probability of observing the number of genes from a specific GO functional category within each cluster. More detailed explanation on this p-value can be found in [3] and [6]. A low p-value indicates that the genes belonging to the enriched functional categories are biologically significant in the corresponding clusters. In the present study, only functional categories with p-value less than 5.0×10^{-15} are reported.

Of the five clusters obtained for the yeast cell-cycle data set (Table 4), the cluster C_2 contains several enriched categories on 'ribosome'. The highly enriched category in cluster C_2 is the 'ribosome' with p-value of 2.32×10^{-37}. The GO categories 'cytosol' and 'protein biosysthesis' are also highly enriched in this cluster with p-values of 1.25×10^{-36} and 1.27×10^{-32} respectively. The cluster C_3 contains the yeast genes corresponding to the DNA replication-involved GO biological process. The highly enriched categories in cluster C_3 are the 'cell cycle' with p-value of 3.16×10^{-26} and the 'DNA replication and chromosome cycle' with p-value of 2.38×10^{-22}. From the results of Table 4, we see that the cluster obtained by the proposed method shows a high enrichment of functional categories.

4. Conclusion

Conventional clustering methods have required a complete data matrix as input even if many microarray data sets are incomplete due to the problem of missing values.

Table 3: Comparison of RMSE of the imputation methods and the proposed method for the yeast sporulation data set.

Method \ %missing	5%	10%	15%	25%
KNNimpute	0.11	0.20	0.28	0.51
EMimpute	0.17	0.30	0.38	0.60
Proposed	0.15	0.22	0.27	0.37

Table 4: Enrichment of GO categories in each of the clusters obtained by the proposed method for the yeast cell-cycle data set of [18]. The number of clusters is five. Only functional categories with p-values less than 5.0E-15 are reported.

Cluster	No. of genes	GO category	GO number	% of genes	p-value
C_1	656	nucleolus	GO:0005730	8.23	6.45E-19
		RNA metabolism	GO:0016070	11.13	7.05E-19
		ribosome biogenesis	GO:0007046	7.01	7.63E-17
		RNA processing	GO:0006396	9.91	1.39E-16
C_2	850	ribosome	GO:0005840	11.88	2.32E-37
		cytosol	GO:0005829	13.76	1.25E-36
		protein biosynthesis	GO:0006412	14.82	1.27E-32
		ribonucleoprotein complex	GO:0030529	13.76	2.66E-30
		cytosolic ribosome (sensu Eukarya)	GO:0005830	8.59	1.01E-27
		large ribosomal subunit	GO:0015934	6.12	1.88E-19
		small ribosomal subunit	GO:0015935	4.94	5.40E-17
		cytosolic large ribosomal subunit (sensu Eukarya)	GO:0005842	4.59	2.62E-15
C_3	523	cell cycle	GO:0007049	19.69	3.16E-26
		DNA replication and chromosome cycle	GO:0000067	12.43	2.38E-22
		chromosome	GO:0005694	10.71	1.13E-19
		mitotic cell cycle	GO:0000278	12.43	1.90E-16
		DNA metabolism	GO:0006259	14.53	2.80E-16
C_4	618	carbohydrate metabolism	GO:0005975	7.77	5.96E-15

In such cases, typically either genes with missing values have been removed or the missing values have been estimated using imputation methods prior to the cluster analysis. In the present study, we focused on the bad influence of the earlier imputation on the subsequent cluster analysis. To address this problem, we have presented the proposed method of clustering incomplete gene expression data. By taking the alternative optimization approach, the missing values are considered as additional parameters for optimization. The evaluation results based on gene annotations have shown that the proposed method is the superior and effective method for clustering incomplete gene expression data. Besides the issues mentioned in present work, we initialized missing values with the corresponding attributes of the cluster centroid to which the incomplete data point is closest. Although this way of initialization is considered appropriate, further work examining the impact of different initializations on clustering performance is needed.

Acknowledgement. This Research was supported by the Chung-Ang University Research Grants in 2006.

References

[1] M. Eisen, P.T. Spellman, P.O. Brown, et al., *Cluster analysis and display of genome-wide expression patterns*, Proc. Natl. Acad. Sci. USA **95** (1998) 14863–14868.

[2] P. Tamayo, D. Slonim, J. Mesirov, et al., *Interpreting patters of gene expression with self-organizing maps - methods and application to hematopoietic differentiation*, Proc. Natl. Acad. Sci. USA **96** (1999) 2907–2912.

[3] S. Tavazoie, J.D. Hughes, M.J. Campbell, et al., *Systematic determination of genetic network architecture*, Nat. Genet. **22** (1999) 281–285.

[4] Y. Xu, V. Olman, D. Xu, *Clustering gene expression data using a graph-theoretic approach - an application of minimum spanning trees*, Bioinformatics **17** (2001) 309–318.

[5] R. Steuer, J. Kurths, C.O. Daub, et al., *The mutual information: Detecting and evaluating dependencies between variables*, Bioinformatics **18** (2002) S231-S240.

[6] D. Dembele, P. Kastner, *Fuzzy c-means method for clustering microarray data*, Bioinformatics **19** (2003) 973–980.

[7] I.S. Dhilon, E.M. Marcotte, U. Roshan, *Diametrical clustering for identifying anti-correlated gene clusters*, Bioinformatics **19** (2003) 1612–1619.

[8] D. Horn, I.Axel, *Novel clustering algorithm for microarray expression data in a truncated SVD space*, Bioinformatics **19** (2003) 1110–1115.

[9] S. Dudoit, J. Fridlyand, *Bagging to improve the accuracy of a clustering procedure* Bioinformatics **19** (2003) 1090–1099.

[10] R. Sharan, A. Maron-Katz, R. Shamir, *CLICK and EXPANDER: a system for clustering and visualizing gene expression data*, Bioinformatics **19** (2003) 1787–1799.

[11] O. Troyanskaya, M. Cantor, G. Sherlock, et al., *Missing value estimation methods for DNA microarrays*, Bioinformatics **17** (2001) 520–525.

[12] T.H. Bo, B. Dysvik, I. Jonassen, *LSimpute: accurate estimation of missing values in microarray data with least square methods*, Nucleic Acids Research **32** (2004) e34.

[13] M. Ouyang, W.J. Welsh, P. Georgopoulos, *Guassian mixture clustering and imputation of microarray data*, Bioinformatics **20** (2004) 917–923.

[14] A.A. Alizadeh, M.B. Eisen, R.E. David et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling*, Nature **403** (2000) 503–511.

[15] M.E. Fuschik, *Methods for Knowledge Discovery in Microarray Data*, Ph.D. Thesis, University of Otago (2003).

[16] R.J. Hathaway, J.C. Bezdek, *Fuzzy c-means clustering of incomplete data*, IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics **31** (2001) 735–744.

[17] S. Chu, J. DeRish, M. Eisen, et al., *The transcriptional program of sporulation in budding yeast*, Science **282** (1998) 699–705.

[18] R.J.Cho, M.J. Campbell, E.A. Winzeler, et al., *A genome-wide transcriptional analysis of the mitotic cell cycle*, Mol. Cell **2** (1998) 65–73.

[19] F.D. Gibbons, F.P. Roth, *Judging the quality of gene expression-based clustering methods using gene annotation*, Genome Res. **12** (2002) 1574–1581.

[20] M. Ashburner, C.A. Ball, J.A. Blake, et al., *Gene Ontology: tool for the unification of biology*, Nat. Genet. **25** (2000) 25–29.

[21] L. Issel-Tarver, K.R. Christie, K. Dolinski, et al., *Saccharomyces genome database*. Methods Enzymol **350** (2002) 329–346.

[22] K. Yeung, D.R. Haynor, W.L. Ruzzo, *Validating clustering for gene expression data*, Bioinformatics **17** (2001) 309–318.

저 자 소 개

Dae-Won Kim

한국 퍼지 및 지능시스템학회 이사
현재 중앙대학교 컴퓨터공학부 부교수
제 16권 3호(2006년 6월호) 참조
E-mail : dwkim@cau.ac.kr