Bayesian Validation of Fuzzy Clustering for Analysis of Yeast Cell Cycle Data*

Kyung-Joong Kim, Si-Ho Yoo, and Sung-Bae Cho

Dept. of Computer Science, Yonsei University
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea
{kjkim,bonanza,sbcho}@cs.yonsei.ac.kr

Abstract. Clustering for the analysis of the gene expression profiles has been used for identifying the functions of the genes and of unknown genes. Since the genes usually belong to multiple functional families, fuzzy clustering methods are more appropriate than the conventional hard clustering methods. However, it is still required to devise natural way to measure the quality of the cluster partitions that are obtained by fuzzy clustering. In this paper, a Bayesian validation method of selecting a fuzzy partition with the largest posterior probability given the dataset is proposed to evaluate the fuzzy partitions effectively. Analysis of yeast cell-cycle data follows to show the usefulness of the proposed method.

1 Introduction

Clustering groups thousands of genes by their similarity of expression levels and helps to analyze gene expression profiles. This organizes the patterns of genes into groups by the similarity of the dataset and has been used for identifying the functions of the genes in the cluster and analyzing the functions of unknown genes. Hard clustering, a hard partitioning method, assigns a sample to only one group. But the real world data like gene expression profiles do not have clear boundaries and they cannot be easily partitioned by hard clustering. Since some genes also belong to multiple functional families, analyzing the genes by hard clustering method has limitations. Fuzzy clustering, unlike the hard clustering, assigns a sample to multiple groups by their grade of membership values [1].

The most important matters that need to be addressed in any clustering method are how many clusters are actually in the dataset and how good the clusters are. Thus, it is necessary to validate each of the fuzzy partition and this evaluation is called cluster validity. Various investigations about these matters have been conducted. Partition coefficient (PC) and partition entropy (CE) were first proposed by Bezdeck [2]. These two cluster validity indexes produce optimal partition at maximum validity measures. Xie-Beni's index (XB) [3] and Fukuyama Sugeno index (FS) [4] are popular in the field of fuzzy clustering. The Xie-Beni index is a ratio of the within cluster sum of squared distances to the product of the number of elements and the minimum between cluster separations, and the Fukuyama Sugeno index measures the compactness and

* This research was supported by the Ubiquitous Computing Research Program funded by the Ministry of Information and Communication of Korea

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 777–784, 2005.

separation of the resulting fuzzy partition after a dataset has been separated into several clusters. However, since the conventional validity indexes are based on the distance between the clusters, we cannot fully represent the structure of the dataset [5].

In this paper, we propose a Bayesian validation method, which evaluates the result of clustering by posterior probability of the fuzzy partitions of given dataset. Unlike the conventional validity indexes, Bayesian validation method never uses the distance between the clusters. It selects the partition with the largest posterior probability in a given dataset. Yeast cell-cycle data is analyzed by the proposed method.

2 Backgrounds

Studies about cluster analysis of the DNA microarray data are summarized in Table 1. Yeung analyzed yeast cell-cycle data by k-means and single-linkage algorithm [6]. Bolshakova and Azuaje used SOM and hard k-means algorithm for clustering and Silhouette index for cluster validation [7]. Also, Eisen analyzed yeast cell-cycle data by fuzzy k-means algorithm and k-means algorithm [8]. Dembele and Kastner used fuzzy c-means algorithm to analyze serum and yeast cell-cycle data [9]. Most of validity indexes used in these researches is all based on the distance between the clusters or between the samples in a cluster: intra-cluster distance and inter-cluster distance.

Table 1. Related works on DNA microarray data

Author	Algorithm	Validity index	Data
Yeung et al. (2001)	K-means Single-linkage	Figure of Merits	Yeast cell-cycle
Bolshakova and Azuaje (2002)	SOM K-means	Dunn's based Index Silhouette Index	Leukemia Lymphoma
Gasch and Eisen (2002)	Fuzzy k-means	N/A	Yeast cell-cycle
Dembele and Kastner (2003)	Fuzzy c-means	Silhouette index	Serum Yeast cell-cycle Human cancer

3 Bayesian Validation Method

All the previous indexes including PC, CE, FS and XB focused on only the compactness and the variation within cluster. However, those indexes lack to provide a correct representation of fuzzy partition in the data since the separation is simply computed by considering only the distance between cluster centroids.

$$\lim_{c \to n} \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} \left\| x_j - v_i \right\|^2 = 0 \tag{1}$$

As shown in Eq (1), if the number of clusters c approaches to the number of samples n, the distance between the cluster centroid and a sample becomes 0. Thus, the traditional indexes lose their ability to validate fuzzy partition for large values of c [5]. Bayesian validation method is a probability-based approach, selecting a fuzzy partition with the largest posterior probability given the dataset. It chooses a partition which has maximum posterior probability given the dataset as an optimal cluster partition. Using Bayes's theorem, the posterior probability given the *Dataset*=$\{d_1, d_2, \ldots, d_N\}$, could be obtained by multiplication rule and independence rule as follows:

$$P(Cluster \mid Dataset) = \frac{P(Cluster)P(Dataset \mid Cluster)}{P(Dataset)} \tag{2}$$

$$\begin{aligned} P(Cluster \mid Dataset) &= P(Cluster \mid d_1, d_2, ..., d_N) \\ &= P(Cluster \mid d_1) \times P(Cluster \mid d_2) \times ... \times P(Cluster \mid d_N) \end{aligned} \tag{3}$$

The sum of $P(Cluster|Dataset)$ for all c is calculated using Eq (4) and Eq (5) and this value is defined as Bayesian Score (BS). This score indicates how well the fuzzy partition represents the dataset by the posterior probability. Larger value of BS means better cluster partition.

$$\begin{aligned} BS &= \frac{\sum_{i=1}^{c} P(C_i \mid D_i)}{c} = \frac{\sum_{i=1}^{c} P(C_i \mid d_{i1}, d_{i2}, ..., d_{iN})}{c} = \frac{\sum_{i=1}^{c} P(C_i \mid d_{i1}) P(C_i \mid d_{i2}) ... P(C_i \mid d_{iN})}{c} \\ &= \frac{\sum_{i=1}^{c} \prod_{j=1}^{N_i} P(C_i) P(d_{ij} \mid C_i) / P(d_{ij})}{c}, \qquad D_i = \{ d_{ij} \mid u_{ij} > \alpha,\ 1 \le j \le n \},\ N_i = \mathrm{n}(D_i) \end{aligned} \tag{4}$$

In Eq (4), d_{ij} is the jth sample which belongs to the ith cluster. $\mathrm{n}(D_i)$ is the number of D_i's and we select only a sample which has larger membership value (u_{ij}) than certain threshold α for calculation. Since the fuzzy clustering aims mainly to analyze the samples which belong to multiple classes, evaluating the partition with samples whose membership values are larger than certain threshold is more appropriate to group samples by fuzzy clustering method. This threshold is defined as α-cut. Since each membership value u_{ij} represents the belongness of a data x_i to certain cluster c, u_{ij} can be substituted for $P(d_{ij}|C_i)$. $P(C_i)$ and $P(d_{ij})$ are calculated as follows:

$$P(C_i) = \sum_{j=1, u_{ij} > \alpha}^{n} u_{ij} \Big/ \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}, \quad P(d_{ij}) = \sum_{i=1}^{c} P(C_i) P(d_{ij}) = \sum_{i=1}^{c} P(C_i) u_{ij} \tag{5}$$

Figure 1 shows the outline of the proposed method. D_1 includes the samples in cluster C_1 whose membership values are larger than α. Finally, BS is obtained and used to select the optimal fuzzy partition.

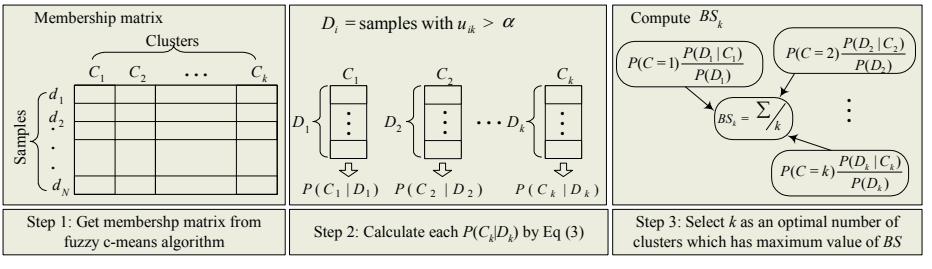


Fig. 1. Process of Bayesian validation

The algorithm of Bayesian validation method is as follow:

- Step 1: Compute the membership matrix u_{ij}
- Step 2: Construct D_i by selecting samples ($u_{ij} > \alpha$) in each cluster
- Step 3: Compute $P(D_j|C_j)$, $P(D_j)$, and $P(C_j)$ of D_i
- Step 4: Compute Bayesian Score using the calculated values at step 2
- Step 5: Evaluate the fuzzy partition with the maximum value of BS as optimal one

4 Experimental Results

Yeast cell-cycle data is analyzed with the proposed method. This set contains time-course expression profiles for more than 6000 genes, with 17 time points for each gene taken at 10-min intervals covering nearly two yeast cell cycles (160min). This dataset is very attractive because a large number of genes contained in it are biologically characterized and have been assigned to different phases of the cell cycle. 421 genes are extracted and used for experiments because they are known as informative genes in clustering [10].

Figure 2 shows the results of all the validation methods including the proposed one, where x axis represents the number of clusters and y axis represents the evaluation value of each validation method. PC and CE have determined the optimal fuzzy partition at c=5, FS at c=35, XB at c=13, and DI at c=7 respectively. Unlike the other methods, BS leads to the optimal value at c=29. All validity measures show different results and we analyzed biological functions of the cluster partition and its members (genes) which belong to multiple clusters.

We have compared the result of BS which produces the optimal fuzzy partition at c=29 with biological knowledge of yeast cell-cycle data [11]. Yeast cell-cycle data represents expression levels of the genes in each of the five cell cycles (Early G_1 – Late G_1 - S - G_2 - M). Each cell cycle includes the genes that show higher expression levels at that cycle time than other cycle times.

By finding clusters that show high peak point in expression levels at certain time in the cycle, we have assigned the cluster to that cycle. Table 2 shows the assigned cluster number and the cycles which they belong to. Clusters that have high expression levels at certain cycle time show low expression level at the other cycle times. Genes assigned between the cycles (intercourse) play a role in regulating the genes that lie in the next cell cycle.

The next step of the analysis is to verify known biological information that the proposed method is indeed able to extract correct information that corresponds to different phases of the yeast cell-cycle data.

Table 3 arranges the genes whose biological functions are known and their cluster number in bracket. Each cycle includes the detailed function groups like DNA replication, biosynthesis, mating pathway and so on. We have confirmed that the results produced by the proposed method are reliable according to the biological knowledge of the genes.

We have chosen special genes whose 1st membership values lie between 0.35 and 0.7, and 2nd membership values are larger than 0.3. These fuzzy genes are belonged to multiple clusters and they provide useful information in gene analysis. Figure 3 shows these fuzzy genes and their biological descriptions with cluster numbers which they belong to. We have classified 4 categories of genes by using the discovered knowledge from Table 7. The genes in cluster 3, cluster 10, cluster 20, and cluster 21 are related to Early G_1 phase. For example, YNL078W belongs to cluster 3 (0.4316) and cluster 19 (0.313888) simultaneously. Actually cluster 3 is related to mating pathway and cluster 19 is related to glycolysis respiration in the same Early G_1 cycle. YNL078W plays multiple roles in Early G_1 cycle. YPR019W, YHR113W, and YHR038W are also fuzzy genes that have multiple functions in cell's life.

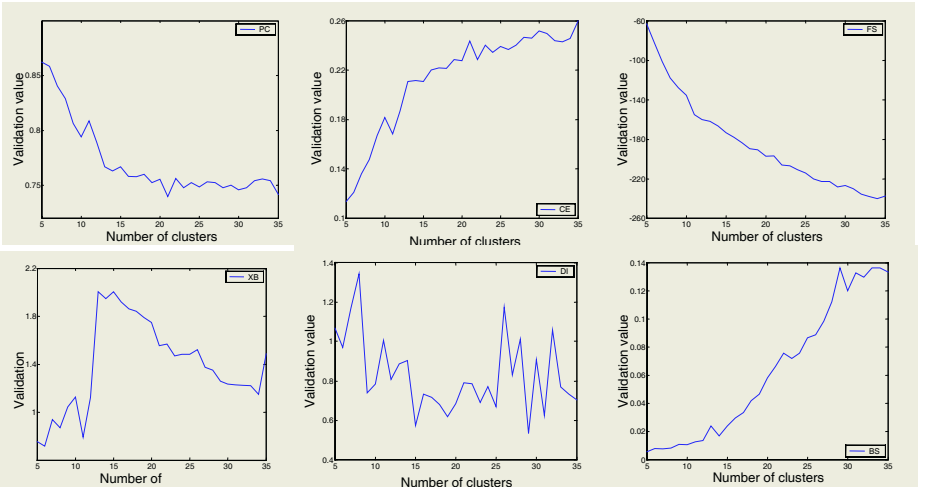


Fig. 2. Preferable values of c for yeast cell-cycle data by each cluster validity measure

Table 2. Analysis of cell cycle and clusters

Time (×10 min)	Cell-cycle	Cluster showing peak expression levels on corresponding cycle
0-3	G_1 phase	Cluster5, Cluster6, Cluster4, Cluster24
	intercourse	Cluster2, Cluster12, Cluster26, Cluster28
3-5	S phase	Cluster8, Cluster13, Cluster14, Cluster16
	intercourse	Cluster11
5-7	G_2 phase	Cluster13
	intercourse	Cluster18
7-9	M phase	Cluster7, Cluster17
	intercourse	Cluster10, Cluster21, Cluster3, Cluster20, Cluster19
9-11	G_1 phase	Cluster5, Cluster6, Cluster4, Cluster24
	intercourse	Cluster2, Cluster12, Cluster26, Cluster28
11-13	S phase	Cluster8, Cluster13
	intercourse	Cluster11
13-15	G_2 phase	Cluster0, Cluster13
	intercourse	Cluster18
15-17	M phase	Cluster7, Cluster17

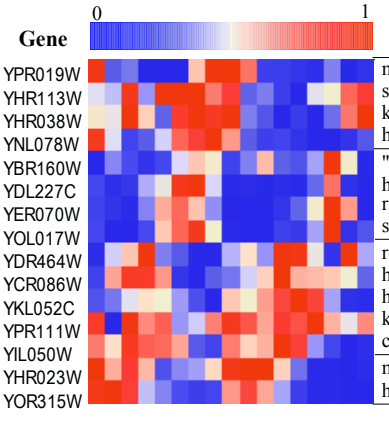
Other fuzzy genes in second category (cluster 12, cluster 24, and cluster 26) are related to Late G_1 phase. Gene like YBR160W, belongs to cluster 12 (0.3982) and cluster 6 (0.3464) simultaneously. Cluster 12 is related to cell cycle regulation and cluster 6 is related to chromosome segregation. Cluster 9, cluster 11, and cluster 13 are related to G_2 phase and cluster 7 and cluster 18 are related to M phase in cell cycle rotation as shown in Figure 3.

We have plotted the fuzzy genes which are analyzed in Figure 3 and their relations are shown in Figure 4. We have used PCA (Principal Component Analysis) to reduce the dimensions of the genes to three and displayed all genes in 3-dimensional space). Fuzzy genes are represented as black cross (X) and rests of genes are represented as

different shapes (diamonds, rectangle, triangle, and circle) according to their belonged clusters. As shown in Figure 4, it is clear to see that YHR113W and YHR038W are located between cluster 20 and cluster 21 which are related to Early G_1 phase. Also YHR023 and YOR315W which belong to cluster 7 and cluster 18, are located between these two clusters. These two clusters are related to M phase in cell cycle rotation. Between the other clusters related to Late G_1 phase and G_2 phase, there exist fuzzy genes, providing useful information for further research about unknown genes. Fuzzy genes which have multiple functional families do not have clear boundaries and belong to multiple clusters simultaneously.

Table 3. Analysis of cell cycle and functional groups

Cell-cycle	Functional groups	Genes
Early G_1 phase	DNA replication	YBL023C(10) YEL032W(10) YPR019W(10)
	Mating pathway	YJL157C(3) YKL185W(3)
	Glycolysis, Respiration	YCR005C(20) YCL040W(20) YLR258W(20)
	Biosynthesis	YIL009W(21) YLL040C(21)
Late G_1 phase	Cell cycle regulation	YBR160W(12) YDL127W(12) YGR109C(12) YPR120C(12)
	Chromosome segregation	YDL003W(26) YFL008W(26) YJL074C(26) YKL042W(26) YMR076C(26) YMR078C(26)
	DNA replication	YBR278W(24) YKL045W(24) YLR103C(24) YPR018W(24)
S phase	Chromosome segregation	YDR113C(16) YGR140W(16) YHR172W(16)
	DNA replication	YBL002W(8) YBL003C(8)
	Miscellaneous	YCR035C(14) YER016W(14) YJR137C(14)
G_2 phase	Directional growth	YJL099W(11) YJR076C(11)
	DNA replication	YDR224C(27) YDR225W(27)
M phase	Cell cycle regulation	YGL116W(7) YPR119W(7)
	Transcriptional factor	YDR146C(18) YLR131C(18)
	Directional growth	YCL037C(17)



Gene description	Clusters
member of the Cdc46p/Mcm2p/Mcm3p family	10, 3
similarity to vacuolar aminopeptidase Ape1p	20, 21
killed in mutagen	20, 21
hypothetical protein	3, 19, 25
"g1,g2" CDC28 cyclin-dependent kinase	6, 12, 24
homothallic switching endonuclease	5, 12, 26
ribonucleoside-diphosphate reductase, large subunit	12, 24, 26
similarity to YFR013w	12, 26
regulates spliceosome components	9 ,11
hypothetical protein	9 ,11
hypothetical protein	9 ,11
kinase involved in late nuclear division	9 ,13
cyclin like protein interacting with Pho85p	11, 13
myosin-1 isoform heavy chain	7, 18
hypothetical protein	7, 18

Fig. 3. Analysis of fuzzy genes (gene description and cluster number)

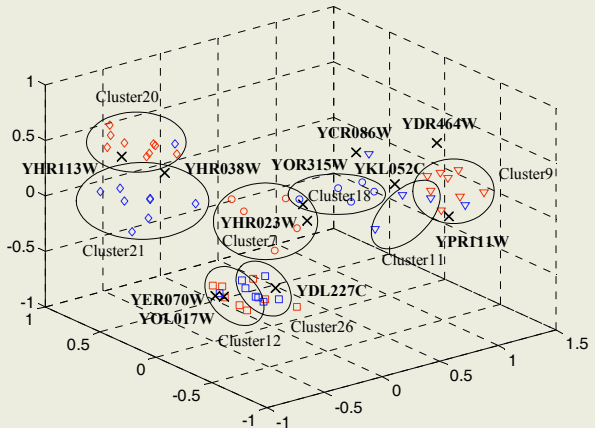


Fig. 4. 3D plot display of fuzzy genes

5 Concluding Remarks

In this paper, a new cluster validation method for the fuzzy partition has been proposed. Bayesian validation method evaluates the fuzzy partition by the posterior probability for the dataset at hand. The best fuzzy partition is obtained by finding the maximum BS value with respect to the number of clusters. We have established α-cut as threshold in computing the value of BS to evaluate various kinds of cluster partitions. We have analyzed the yeast cell-cycle data with the proposed method. To confirm the superiority of the proposed method, the results are verified with biological knowledge.

References

1. A. P. Gasch and M. B. Eisen, Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, vol. 3, no. 11, research 0059.1-0059.22. 2002.
2. J. C. Bezdeck, Cluster validity with fuzzy sets. *J. Cybernit.*, vol. 3, pp. 58-72, 1974.
3. X. L. Xie and G. Beni, A validity measure for fuzzy clustering. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841-846, 1991.
4. Y. Fukuyama and M. Sugeno, A new method of choosing the number of clusters for the fuzzy c-means method. *Proceedings of 5th Fuzzy Systems Symposium*, pp. 247-250, 1989.
5. D. W. Kim, K. H. Lee, D. and H. Lee, Fuzzy cluster validation index based on inter-cluster proximity. *Pattern Recognition Letters*, vol. 24, pp. 2561-2574, 2003.
6. K. Y. Yeung, et al., "Validating clustering for gene expression data," *Bioinformatics*, vol. 17, no. 4, pp. 309-318, 2001.
7. N. Bolshakova and F. Azuaje, "Cluster validation techniques for genome expression data," *SIGPRO*, vol. 21, no. 82, pp. 1-9. 2002.
8. A. P. Gasch and M. B. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome Biology*, vol. 3, no. 11, research 0059.1-0059.22, 2002.

9. D. Dembele and P. Kastner,"Fuzzy c-means method for clustering microarray data," *Bioinformatics*, vol. 19, no. 8, pp. 973-980, 2003.
10. J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673-679, 2001.
11. R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart and R. W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, pp. 65-73, 1998.