Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 39 (2006) 2405–2414
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

# Fuzzy Bayesian validation for cluster analysis of yeast cell-cycle data

Sung-Bae Cho*, Si-Ho Yoo

*Department of Computer Science, Yonsei University, 134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, South Korea*

Received 6 July 2005; received in revised form 9 December 2005; accepted 15 December 2005

**Abstract**

Clustering for the analysis of the genes organizes the patterns 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. Since the genes usually belong to multiple functional families, fuzzy clustering methods are more appropriate than the conventional hard clustering methods which assign a sample to only one group. In this paper, a Bayesian-like validation method selecting a fuzzy partition is proposed to evaluate the fuzzy partitions effectively. The theoretical interpretation of the obtained memberships is beyond the scope of this paper, and an empirical evaluation of the proposed method is conducted by comparing to the four representative conventional fuzzy cluster validity measures in four well-known datasets. Analysis of yeast cell-cycle data follows to evaluate the proposed method.


*Keywords:* Fuzzy clustering; Fuzzy *c*-means algorithm; Fuzzy Bayesian validation method; Yeast cell-cycle data

## 1. Introduction

Clustering groups thousands of genes by their similarity of expression levels and helps to analyze gene expression profiles [1]. 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 dataset, like gene expression profiles, does not have clear boundaries and they cannot be easily partitioned by hard clustering. Since some genes also have multiple functional families, analyzing the genes by hard clustering method has some limits. Fuzzy clustering, unlike the hard clustering, assigns a sample to multiple groups by their grade of membership values [2]. Fuzzy clustering method is more robust on noise and more appropriate in analyzing gene expression profiles than hard clustering method [3].

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. Many investigations about these matters have been conducted. Partition coefficient and partition entropy were first proposed by Bezdeck [4,5]. These two cluster validity indices produce the optimal partition at maximum validity measures. Xie–Beni's index [6] and Fukuyama–Sugeno index [7] are popular in the field of fuzzy clustering. The Xie–Beni index is a ratio of the fuzziness within cluster sum of squared distances to the product of the number of elements and the minimum between cluster separations, and the Fukuyama–Sugeno measures the compactness and separation of the resulting fuzzy partition after a dataset has been separated into several clusters. However, since the conventional validity indices are based on the distance between the clusters, they cannot fully represent the structure of the dataset [8].

In this paper, we propose a Bayesian-like validation method which evaluates the result of clustering by memberships of the fuzzy partition of a given dataset. We therefore call it fuzzy Bayesian validation method. Unlike the conventional validity indices, the fuzzy Bayesian validation method never uses the distance between the clusters. It selects the partition with the largest membership in a given

* Corresponding author. Tel.: +82 2 2123 2720.
*E-mail address:* sbcho@cs.yonsei.ac.kr (S.-B. Cho).

0031-3203/$30.00 
doi:10.1016/j.patcog.2005.12.007

dataset. We take the probabilistic Bayesian equation and formally replace probabilities by memberships. The effectiveness of the proposed method is demonstrated in four known datasets, where the fuzzy $c$-means algorithm is used as the clustering algorithm [9,10]. After the preliminary experiments, yeast cell-cycle data is analyzed by the proposed method.

The rest of the paper is organized as follows. In Section 2, the conventional cluster validation methods are described. Section 3 introduces the fuzzy Bayesian validation method and its mathematical backgrounds. Experimental environments and results are presented in Section 4. Conclusions and future works are presented in Section 5.

## 2. Conventional cluster validity measures

This section describes the cluster validity measures which evaluate the cluster results. Four conventional measures—partition coefficient, classification entropy, Fukuyama–Sugeno and Xie–Beni index—are explained.

- Partition coefficient (PC): Bezdeck proposed a validity index PC for fuzzy clustering as follows:

$$PC(U;c)=\frac{\sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^2}{n}, \tag{1}$$

where $n$ is the number of samples and $c$ represents the number of clusters. Optimal partition is obtained by maximizing the value of PC with respect to certain value of $c$. However, this index does not perform well in multi-class dataset, because the value of PC decreases monotonically when $c$ gets larger [5].

- Classification entropy (CE): CE is one of the most widely used cluster validity indices.

$$CE(U;c)=\frac{-\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}\log_a u_{ij}}{n}. \tag{2}$$

It looks similar to PC, but it takes log ratio of membership $u_{ij}$. Optimal partition is obtained by minimizing the value of CE with respect to certain value of $c$ [5]. CE also has monotonic decreasing tendency as $c$ gets larger, like PC.

- Fukuyama–Sugeno (FS): Fukuyama and Sugeno tried to model the cluster validation by exploiting the compactness and separateness [7].

$$FS(U,V;X)=\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m(\|X_j-V_i\|^2 - \|V_i-\bar{V}\|^2),\quad \bar{V}=\frac{1}{n}\sum_{i}^{n}x_i. \tag{3}$$

Here, $\|X_j-V_i\|^2$ is a compactness measure, and $\|V_i-\bar{V}\|^2$ is a degree of separation between each cluster and the mean of cluster centroids. Smaller value of FS indicates the better partition result in a given dataset. But like the other indices, FS has the problem of monotonic decreasing tendency when $c$ gets larger.

- Xie and Beni index (XB): Xie and Beni also proposed a validity index XB that focused on two properties: compactness and separateness [6].

$$XB(U,V;X)=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^2\|V_i-X_j\|^2}{nd_{\min}^2},\quad d_{\min}=\min_{i,j}\|V_i-V_j\|. \tag{4}$$

The numerator part indicates the compactness of fuzzy partition, while the denominator part does the strength of separation between clusters. The most desirable partition is obtained by minimizing XB over certain value of $c$. $d_{\min}$ indicates the minimum distance between the clusters.

These conventional methods have common problem that their index values decrease when $c$ gets larger, and there have been some researches for novel cluster validity indices such as PBM-index [11] and PCAES [12] to solve this problem.

## 3. Fuzzy Bayesian validation method

All the previous indices including PC, CE, FS and XB focused on only the compactness and the variation within cluster. However, those indices are limited in their ability 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.

$$\lim_{c->n}\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m\|x_j-v_i\|^2=0. \tag{5}$$

As shown in Eq. (5), 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 indices lose their ability to validate fuzzy partition for large values of $c$ [13,14]. The fuzzy Bayesian validation method is inspired by the classic Bayesian concept of probability theory, selecting a fuzzy partition with the largest membership given the dataset. It chooses a partition which has maximum membership, given the dataset as an optimal cluster partition [15].

$$\max P(Cluster|Dataset). \tag{6}$$

By formally transferring the principles of the classic Bayes' theorem to memberships we obtain

$$P(Cluster|Dataset)=\frac{P(Cluster)P(Dataset|Cluster)}{P(Dataset)}. \tag{7}$$

By applying multiplication and independence rules we get

$$\begin{aligned}P(Cluster|Dataset)&=P(Cluster|d_1,d_2,\ldots,d_N)\\&=P(Cluster|d_1)\times P(Cluster|d_2)\\&\quad\times\cdots\times P(Cluster|d_N).\end{aligned} \tag{8}$$

The sum of $P(Cluster|Dataset)$ for all $c$ is calculated using Eqs. (7) and (8) and this value is defined as Bayesian score (BS). We propose that this score indicates how well the fuzzy partition represents the dataset. Experimental evidence will be given in Section 4, but a theoretical foundation is up to future work. The larger the BS the better the cluster partition is:

$$BS = \frac{\sum_{i=1}^{c} P(C_i|D_i)}{C} = \frac{\sum_{i=1}^{c} P(C_i|d_{i1}, d_{i2}, \ldots, d_{iN})}{C}$$

$$= \frac{\sum_{i=1}^{c} P(C_i|d_{i1})P(C_i|d_{i2}) \ldots P(C_i|d_{iN})}{C}$$

$$= \frac{\sum_{i=1}^{c} \prod_{j=1}^{N_i} P(C_i)P(d_{ij}|C_i)/P(d_{ij})}{C},$$

$$D_i = \{d_{ij}|\mu_{ij} > \alpha, 1 \leqslant j \leqslant n\}, \quad N_i = n(D_i). \tag{9}$$

In Eq. (9), $n(D_i)$ is the number of $D_i$'s and we select only a sample which has larger membership value ($u_{ij}$) than certain threshold $\alpha$ for calculation. There are two reasons for doing this: to exclude the samples with zero membership value ($u_{ij} = 0$) for multiplication and since the main purpose of fuzzy clustering is 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 $\alpha$-cut. $P(C_i)$ and $P(d_{ij})$ are calculated as follows:

$$P(C_i) = \frac{\sum_{j=1, u_{ij}>\alpha}^{n} u_{ij}}{\sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}} \tag{10}$$

$$P(d_{ij}) = \sum_{i=1}^{c} P(C_i)P(d_{ij}/C_i) = \sum_{i=1}^{c} P(C_i)u_{ij}. \tag{11}$$

Since each membership value $u_{ij}$ represents the belongness of a data $x_i$ to certain cluster $c$, $u_{ij}$ can be substituted for $P(d|C)$. Fig. 1 shows the outline of the proposed method. $D_1$ includes the samples in cluster $c_1$ whose membership value is larger than $\alpha$. Finally, BS is obtained and used to select the optimal fuzzy partition.

The algorithm of fuzzy Bayesian validation method can be summarized as follows:

- *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 BS using the calculated values of Step 2.
- *Step* 5: Evaluate the fuzzy partition with the maximum value of BS as optimal one.

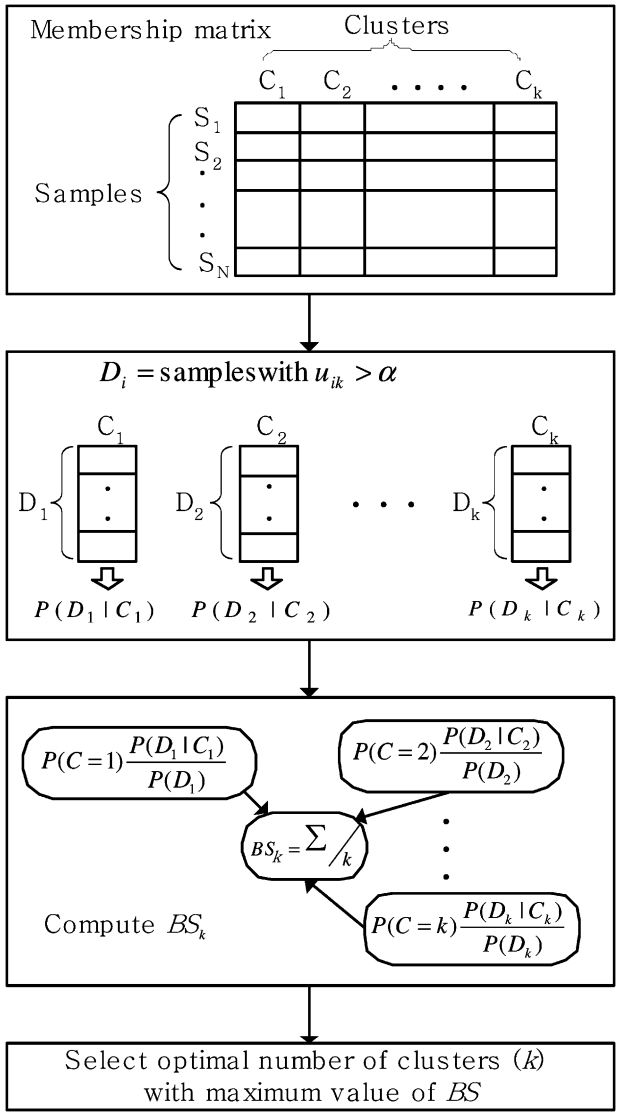


Fig. 1. Fuzzy Bayesian validation method.

## 4. Experimental results

### 4.1. Experiments with benchmark datasets

To show the usefulness of the proposed method, comparisons with four fuzzy cluster validity indices (PC, CE, FS, XB) for the fuzzy partitions obtained from FCM are conducted on four datasets: Iris, Wine and Image datasets downloaded from UCI machine learning repository (http://www.ics.uci.edu/~mlearn/MLSummary.html) and SRBCT i.e. gene expression dataset. Iris dataset contains 120 samples in four dimensional measurement spaces. Iris consists of two or three clusters because of the substantial overlap of two of the clusters. Wine dataset includes 178 samples in 13 dimensional measurement spaces and has three clusters. Image dataset contains 210 samples in 19 dimensional measurement spaces where seven clusters are

Table 1
Cluster validity values on the Wine dataset for $c = 2, \ldots, 7$

| $c$ | PC | CE | FS | XB | BS |
|---|---|---|---|---|---|
| 2 | **0.9358** | **0.0476** | 29.3216 | 0.4516 | 0.1583 |
| 3 | 0.9258 | 0.0564 | −1.5084 | **0.4312** | 0.2707 |
| 4 | 0.8814 | 0.0932 | −8.9773 | 0.8636 | **0.2773** |
| 5 | 0.8308 | 0.1344 | −12.2939 | 1.3137 | 0.2556 |
| 6 | 0.8180 | 0.1494 | −15.3106 | 1.5032 | 0.2477 |
| 7 | 0.7964 | 0.1673 | **−18.4707** | 1.3638 | 0.2160 |

Table 2
Cluster validity values on the Image dataset for $c = 2, \ldots, 8$

| $c$ | PC | CE | FS | XB | BS |
|---|---|---|---|---|---|
| 2 | 0.9468 | **0.0379** | 35.2037 | **0.2617** | 0.2190 |
| 3 | 0.9270 | 0.0550 | −22.3275 | 0.5519 | 0.3846 |
| 4 | **0.9539** | 0.0384 | −65.1495 | 0.3935 | 0.4120 |
| 5 | 0.9448 | 0.0464 | −74.4622 | 0.3895 | 0.3918 |
| 6 | 0.9292 | 0.0599 | −85.9394 | 0.7165 | 0.4672 |
| 7 | 0.8980 | 0.0866 | −91.0603 | 0.8023 | **0.5490** |
| 8 | 0.9224 | 0.0657 | **−104.061** | 0.5720 | 0.5077 |

Table 3
Cluster validity values on the Iris dataset for $c = 2, \ldots, 7$

| $c$ | PC | CE | FS | XB | BS |
|---|---|---|---|---|---|
| 2 | **0.9916** | **0.0060** | −311.725 | **0.0619** | **0.8077** |
| 3 | 0.9781 | 0.0156 | −426.900 | 0.1539 | 0.7161 |
| 4 | 0.9704 | 0.0226 | −459.027 | 0.2189 | 0.6568 |
| 5 | 0.9569 | 0.0331 | −459.622 | 0.5045 | 0.5876 |
| 6 | 0.9560 | 0.0333 | −462.487 | 0.9038 | 0.6653 |
| 7 | 0.9510 | 0.0366 | **−481.403** | 0.6820 | 0.6825 |

known as an optimal partition. The SRBCT consists of four types of cancer (RMS, NB, BL, EWS) and has 63 samples in 96 dimensional measurement spaces [16].

All the experiments are repeated six times on each dataset by increasing the α-cut value from 0.1 to 0.6 by 0.1 and the average score is used as BS value. We have used $m = 1.2$ for the fuzziness parameter. Table 1 shows the results of Wine dataset. PC and CE produce the optimal fuzzy partition at $c = 2$, FS at $c = 7$, and XB at $c = 3$, whereas BS yields $c = 4$ as the optimal fuzzy partition. XB is the only measure producing the correct clusters ($c = 3$). Since the difference between the BS values at $c = 3$ and $c = 4$ is very small (0.0066) compared to other margins, it can be said that the proposed method makes the optimal fuzzy partition at $c = 3$ or $c = 4$.

Table 2 shows the results of Image dataset. All the other methods except the proposed method make a wrong decision on the optimal number of clusters ($c = 7$): PC at $c = 4$, CE at $c = 2$, FS at $c = 8$ and XB at $c = 2$, respectively. Tables 3 and 4 show the results of Iris dataset and SRBCT dataset, respectively. In Table 3, all the methods select $c = 2$ as an optimal fuzzy partition except FS that has the optimal value at $c = 7$. In the case of SRBCT, 4 clusters are known as the optimal number of clusters, and PC, CE, XB, and BS find out the optimal fuzzy partition at $c = 4$, whereas FS finds it at $c = 7$.

Table 4
Cluster validity values on the SRBCT dataset for $c = 2, \ldots, 7$

| $c$ | PC | CE | FS | XB | BS |
|---|---|---|---|---|---|
| 2 | 0.8758 | 0.0969 | 164.4075 | 1.1294 | 0.1918 |
| 3 | 0.9205 | 0.0709 | 86.6529 | 0.8127 | 0.5612 |
| 4 | **0.9393** | **0.0616** | 27.3224 | **0.5657** | **0.7073** |
| 5 | 0.9100 | 0.0850 | −0.9891 | 0.8487 | 0.6731 |
| 6 | 0.8922 | 0.0977 | −22.6041 | 0.7798 | 0.6411 |
| 7 | 0.8989 | 0.0979 | **−34.773** | 0.8670 | 0.6852 |

FS is found as the most unreliable indices since it cannot yield the correct number of clusters for all four datasets. PC, CE and XB tend to monotonically decrease as $c$ increases on Wine and Image datasets. BS makes the correct number of clusters except Wine dataset and does not show monotonic decreasing tendency as $c$ increases.

After comparing the proposed method to four conventional validity indices, experiments with three synthetic datasets are conducted. Three datasets are called synthetic-5-2, synthetic-10-2 and synthetic-5-3. The names imply the number of clusters and dimensions. For example, there are five clusters in two-dimensional space for synthetic-5-2. The distributions of three datasets are illustrated in Fig. 2.

Fig. 3 shows the BS values for the synthetic-5-2, synthetic-10-2 and synthetic-5-3 according to the number of clusters. BS has found out the actual number of clusters of the three synthetic datasets correctly.

### 4.2. Experiments with yeast cell-cycle data

In this section, 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 (160 min). 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. Four hundred and twenty-one genes are extracted and used for experiments because they are known as informative genes in clustering [17]. The fuzzy $c$-means algorithm of well-known fuzzy clustering method is used for clustering [9,10].

Fig. 4 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 $c = 30$, and XB at $c = 6$, respectively. Unlike the other methods, BS leads to the optimal value at $c = 29$. PC shows monotonic

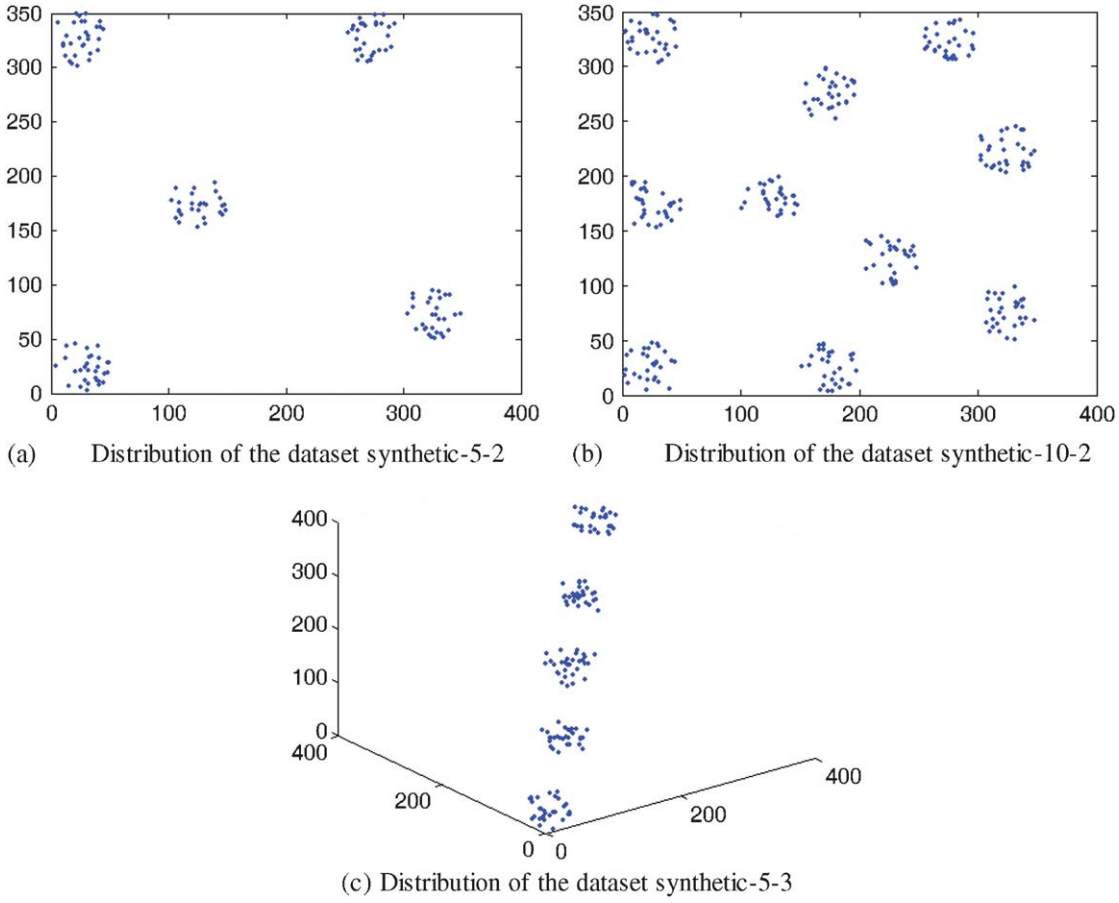


Fig. 2. Distributions of the synthetic datasets. (a) Distribution of the dataset synthetic-5-2; (b) distribution of the dataset synthetic-10-2 and (c) distribution of the dataset synthetic-5-3.

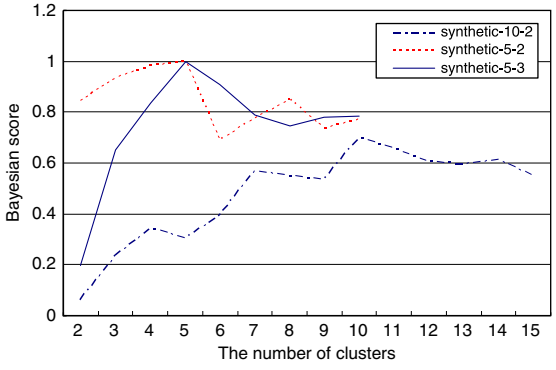


Fig. 3. BS of the synthetic datasets.

decreasing tendency, whereas CE shows monotonic increasing tendency.

We have compared the result of BS which produces the optimal fuzzy partition at $c = 29$ with biological knowledge of yeast cell-cycle data [17]. Yeast cell-cycle data represents expression levels of the genes in each of the five cell cycles (Early G1–Late G1–S–G2–M). Each cell cycle includes the genes that show higher expression levels at that cycle time than others.

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 5 shows the assigned clusters and the cycles that 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 6 arranges the genes whose biological functions are known and their cluster number in parentheses. 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. Tables 7–10 describe the genes and their membership values in certain cluster. Parentheses in the category of the 2nd membership indicate the corresponding gene's cluster number.

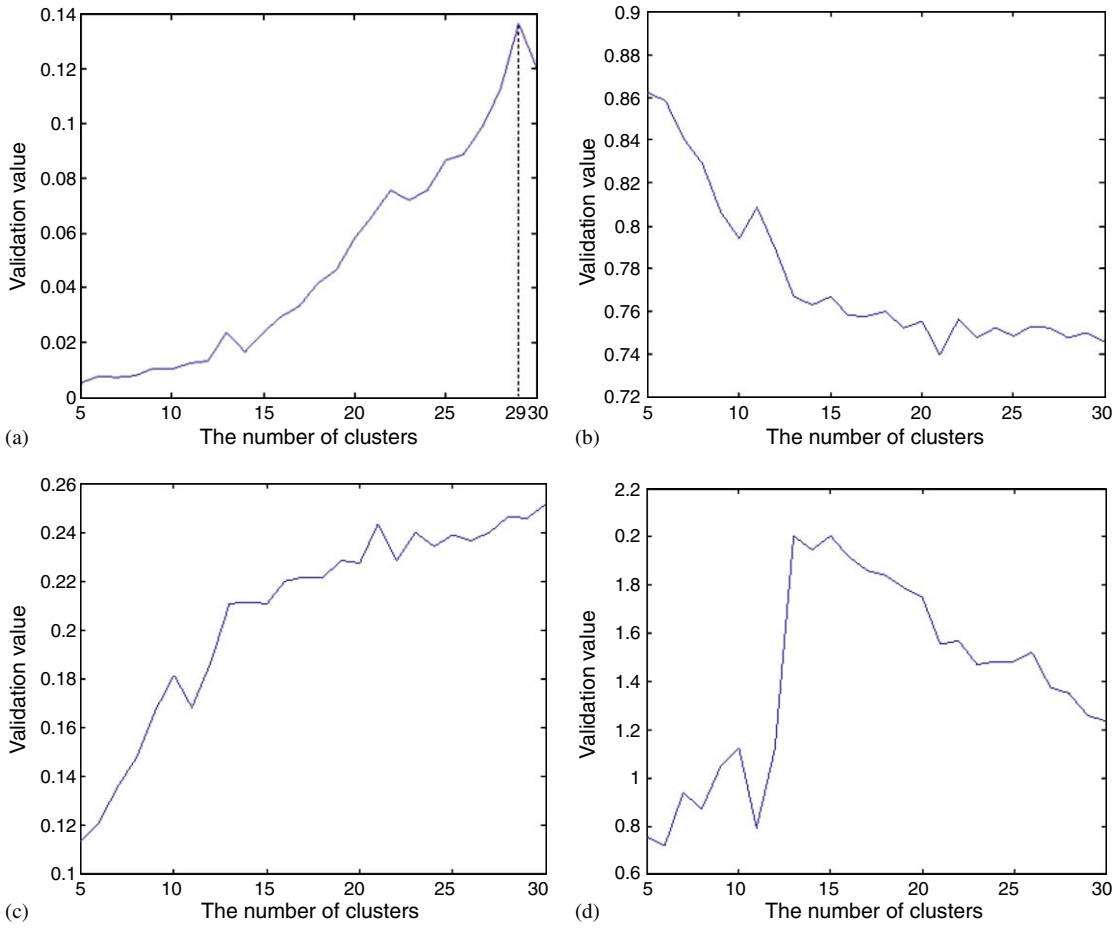


Fig. 4. Values of $c$ for yeast cell-cycle data by each cluster validity measure. (a) BS, (b) PC, (c) CE, (d) XB.

Table 5
Analysis of cell cycle and clusters

| Time (×10 min) | Cell cycle | Cluster showing peak expression levels on the corresponding cycle |
|---|---|---|
| | Intercourse | Cluster19, Cluster20 |
| 1–3 | $G_1$ phase | Cluster4, Cluster5, Cluster6, 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 | Cluster3, Cluster10, Cluster19, Cluster20, Cluster21 |
| 9–11 | $G_1$ phase | Cluster4, Cluster5, Cluster6, 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 |

Table 7 shows the information of cluster 20 in Early G1. Table 8 shows the information of cluster 26 in Late G1, which is composed of the genes that have chromosome segregation function. Table 9 shows the information of cluster 11 in intercourse between S and G2 cycles. Table 10 shows the information of cluster 19 in M cycle.

Table 6
Analysis of cell cycle and genes

| 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) |

Table 7
Cluster 20

| Gene name | 1st membership | 2nd membership | Descriptions |
|---|---|---|---|
| YBR052c | 0.959326 | 0.024133(19) | Strong similarity to S. pombe brefeldin A resistance protein obr1 |
| YBR053c | 0.999252 | 0.000641(21) | Similarity to rat regucalcin |
| YCL040w/GLK | 0.995427 | 0.003458(19) | Glucokinase |
| YCR005c/CIT | 0.627321 | 0.113450(19) | Citrate (si)-synthase, peroxisomal |
| YDL119c | 0.964506 | 0.018641(25) | Similarity to bovine Graves disease carrier protein |
| YDR368w/YPR | 0.993018 | 0.003227(21) | Similar to aldo/keto reductases |
| YDR493W | 0.904762 | 0.042710(21) | Hypothetical protein |
| YDR511W | 0.956695 | 0.031362(21) | Hypothetical protein |
| YHR022C | 0.996261 | 0.002902(21) | Weak similarity to ras-related protein |
| YHR038W | 0.665447 | 0.302911(21) | Killed in mutagen |
| YHR113W | 0.534192 | 0.234211(21) | Similarity to vacuolar aminopeptidase Ape1p |
| YKL163W/PIR | 0.999100 | 0.000490(21) | Protein with internal repeats |
| YLR258W/GSY | 0.866731 | 0.105164(25) | UDP-glucose–starch glucosyltransferase, isoform 2 |
| YNL073W/MSK | 0.987382 | 0.006315(19) | Lysyl-tRNA synthetase, mitochondrial |
| YNL173C | 0.999746 | 0.000097(19) | GTP-binding protein of the pheromone-response pathway |
| YNR001C/CIT | 0.834271 | 0.151766(19) | Citrate (si)-synthase, mitochondrial |
| YOR317W/FAA | 0.770655 | 0.220442(19) | Long-chain-fatty-acid CoA ligase |

Genes in this cycle tend to function as a transcriptional factor.

We have chosen special genes whose the 1st membership value lies between 0.35 and 0.7, and the 2nd membership value is larger than 0.3. These fuzzy genes are belonged to multiple clusters and they provide useful information in gene analysis. Table 11 shows the fuzzy genes. YNL078W belongs to cluster 30 (0.431654) and cluster 19 (0.431654) simultaneously. Actually cluster 3 is related to mating pathway and cluster 19 is related to glycolysis respiration in the same Early G1 cycle. YNL078W plays multiple roles in Early G1 cycle. YPR019W, YHR038W and YHR113W are also fuzzy genes that have multiple functions in cell's life.

Fig. 5 shows how the fuzzy genes are overlapped in their functional groups. YBR160W in Late G2 cycle belongs to cluster 12 (0.398234) and cluster 6 (0.34645) simultaneously. Cluster 12 is related to cell-cycle regulation; cluster 6 is related to chromosome segregation. Other fuzzy genes in G2 cycle are YIL050W, YCR086, YDR464W, YKL052C, and YPR111W. These genes belong to multiple clusters as shown in Fig. 5. Fuzzy genes in M cycle are YHR023W and YOR315W. YHR023W belongs to cluster 18 (0.665914) which is related to chromosome segregation and cluster 7 (0.32451) which is related to cell-cycle regulation.

### 4.3. Related works

Studies about cluster validation of the DNA microarray data are shown in Table 12. Bolshakova and Azuaje used

Table 8
Cluster 26

| Gene name | 1st membership | 2nd membership | Descriptions |
|---|---|---|---|
| YBR073w/RDH | 0.394689 | 0.272282(6) | Required for meiosis, putative helicase |
| YBR088c/POL | 0.949515 | 0.033347(5) | Proliferating cell nuclear antigen (PCNA) |
| YBR089w | 0.996386 | 0.001917(28) | Questionable ORF |
| YCL024w | 0.846978 | 0.099224(12) | Similarity to ser/thr protein kinase Gin4p |
| YDL003W/RHC | 0.997543 | 0.002296(12) | Mitotic Chromosome Determinant |
| YDL010w | 0.580032 | 0.398084(28) | Similarity to hypothetical protein YBR014c and glutaredoxins |
| YDL227C/HO | 0.459841 | 0.352540(12) | Homothallic switching endonuclease |
| YDR097C/MSH | 0.999828 | 0.000115(12) | DNA mismatch repair protein |
| YFL008W/SMC | 0.896596 | 0.088707(12) | Chromosome segregation protein |
| YGR152C/RSR | 0.845395 | 0.109358(12) | GTP-binding protein, affects budding site |
| YHR153c/SPO | 0.970821 | 0.018095(12) | Early meiotic protein |
| YHR154W | 0.936678 | 0.022641(12) | Putative DNA damage responsive cell-cycle checkpoint protein |
| YJL074C/SMC | 0.981196 | 0.008007(28) | Required for structural maintenance of chromosomes |
| YJL187C/SWE | 0.873642 | 0.078148(6) | Phosporylates Cdc28, SPB separation, nucleardivision |
| YKL042W/SPC | 0.898970 | 0.058697(5) | Spindle pole body component |
| YKL113C/RAD | 0.899896 | 0.071100(6) | ssDNA endonuclease and 5′-3′exonuclease |
| YLR183C | 0.985893 | 0.008883(6) | Similarity to YDR501w |
| YML021C/UNG | 0.698835 | 0.265839(28) | Uracil N-glycosylase |
| YMR076C/PDS | 0.869097 | 0.105145(6) | Precocious dissociation of sister chromatids for chromosome |
| YMR078C/CHL | 0.990038 | 0.003983(12) | transmission in mitosis and normal telomere |
| YNL102W/CDC | 0.967338 | 0.027452(6) | DNA-directed DNA polymerase alpha, 180 KD subunit |
| YNL312W/RFA | 0.997584 | 0.001559(6) | DNA replication factor A, 36 kDa subunit |
| YOL090W/MSH | 0.739115 | 0.143342(5) | DNA mismatch repair protein |
| YOR074C/CDC | 0.999579 | 0.000317(28) | Thymidylate synthase |

Table 9
Cluster 11

| Gene name | 1st membership | 2nd membership | Descriptions |
|---|---|---|---|
| YBL032w | 0.441659 | 0.312490(17) | Weak similarity to hnRNP complex protein homolog YBR233w |
| YCR073c/SSK | 0.560968 | 0.178256(15) | MAPKKK, high osmolarity sensitive with SSK2 |
| YCR085w | 0.983793 | 0.009209(9) | Hypothetical protein |
| YHR031C | 0.503745 | 0.220493(13) | Similarity to Pif1p |
| YHR086w/NAM | 0.995097 | 0.001189(9) | Meiotic recombination protein |
| YIL131C/FKH | 0.994513 | 0.003082(14) | Similarity to Drosophila fork head protein |
| YJL092W/HPR | 0.741732 | 0.137793(14) | ATP-dependent DNA helicase |
| YJL099W/CHS | 0.713494 | 0.076135(15) | Chitin biosynthesis |
| YJR076C/CDC | 0.739040 | 0.134265(9) | Involved in proper bud growth |
| YJR110W | 0.970474 | 0.019734(13) | Similarity to Caenorhabditis elegans hypothetical protein T24A11.1 |
| YMR003W | 0.922438 | 0.064507(9) | Hypothetical protein |
| YMR190C/SGS | 0.868787 | 0.108565(9) | DNA helicase |
| YMR215W | 0.968721 | 0.009645(0) | Similarity to GAS1 protein |
| YPL116W/HOS | 0.883831 | 0.050475(14) | Protein with similarity to Hda1p, Hos2p, Hos1p |

Table 10
Cluster 18

| Gene name | 1st membership | 2nd membership | Descriptions |
|---|---|---|---|
| YBR043c | 0.872449 | 0.051899(0) | Similarity to benomyl/methotrexate resistance protein |
| YDR146c/SWI | 0.812059 | 0.088460(0) | Transcription factor |
| YGL021W/ALK | 0.999422 | 0.000238(7) | DNA damage-responsive protein |
| YGL255W/ZRT | 0.894156 | 0.038686(15) | High-affinity zinc transport protein |
| YGR108W/CLB | 0.999238 | 0.000564(7) | Cyclin, G2/M-specific |
| YHR023w/MYO | 0.665914 | 0.324511(7) | Myosin-1 isoform heavy chain |
| YIL158W | 0.980776 | 0.009280(13) | Similarity to hypothetical protein YKR100c |
| YLR131c/ACE | 0.958667 | 0.012878(13) | Metallothionein expression activator |
| YLR353W/BUD | 0.994215 | 0.003963(7) | Budding protein |
| YML033W | 0.982098 | 0.015135(7) | Similarity to YDR458c |
| YMR001C/CDC | 0.983288 | 0.016009(7) | Involved in regulation of DNA replication |
| YOL014W | 0.759511 | 0.141879(7) | Hypothetical protein |
| YOL069W/NUF | 0.902811 | 0.045171(17) | Spindle pole body protein |
| YOR315W | 0.617416 | 0.318186(7) | Hypothetical protein |

Table 11
Fuzzy genes

| Fuzzy gene | 1st cluster | 2nd cluster |
|---|---|---|
| YHR023w/MYO1 | 0.665914 | 0.32451 |
| YHR038W/ | 0.665447 | 0.30291 |
| YOR315W/ | 0.617416 | 0.31818 |
| YIL050W/ | 0.611315 | 0.37182 |
| YBR158w/ | 0.598468 | 0.39674 |
| YDL010w/ | 0.580032 | 0.39808 |
| YPR019W/CDC54 | 0.538395 | 0.32615 |
| YDR464w/SPP41 | 0.530323 | 0.320204 |
| YLL002w/ | 0.524684 | 0.310751 |
| YOL017W/ | 0.513527 | 0.462174 |
| YPL256C/CLN2 | 0.491816 | 0.345759 |
| YKR012C/ | 0.462177 | 0.41311 |
| YDL227C/HO | 0.459841 | 0.35254 |
| YGR189C/ | 0.445827 | 0.38395 |
| YBL032w/ | 0.441659 | 0.31249 |
| YNL078W/ | 0.431654 | 0.313888 |
| YBR160w/CDC28 | 0.398234 | 0.34645 |
| YLR236C/ | 0.389045 | 0.377567 |
| YJR092W/BUD4 | 0.379821 | 0.368409 |
| YDL138W/RGT2 | 0.378443 | 0.318087 |
| YEL017w/ | 0.37469 | 0.302481 |
| YCL038c/ | 0.371376 | 0.364128 |
| YER016w/BIM1 | 0.367631 | 0.349896 |
| YER190w/_f | 0.36546 | 0.317462 |

SOM and hard $k$-means algorithm for clustering and Silhouette index for cluster validation [18]. Demblele and Kastner used fuzzy $c$-means algorithm to analyze serum and yeast cell-cycle data [19]. Also, Eisen and Yeung analyzed yeast cell-cycle data by fuzzy $k$-means algorithm and $k$-means algorithm, respectively [20,21]. Validity indices used in these studies are all based on the distance between the clusters or between the samples in a cluster.

## 5. Conclusion

In this paper, a new cluster validation method for the fuzzy partition has been proposed. Fuzzy Bayesian validation method evaluates the fuzzy partition by the membership for the dataset at hand. The best fuzzy partition is obtained by finding the maximum BS value with respect to $c$. We set $\alpha$-cut as threshold in computing the value of BS to evaluate various kinds of cluster partitions. The performance of the proposed method was tested on four well-known datasets to demonstrate the usefulness with fuzzy $c$-means algorithm. Unlike the conventional methods, the proposed method evaluated the fuzzy partition independent of $c$ and more reliable than other methods. Also, we have analyzed the yeast cell-cycle data with the proposed method. To confirm the superiority of the proposed method, the results were verified with biological knowledge.

While yeast genes may not have large numbers of alternative mRNA transcripts, the method proposed here will allow human transcriptome researchers to identify how a specific gene may trigger alternative splice variants (and perhaps different functions) over time. It would be interesting to explore human cancer gene expression datasets using this

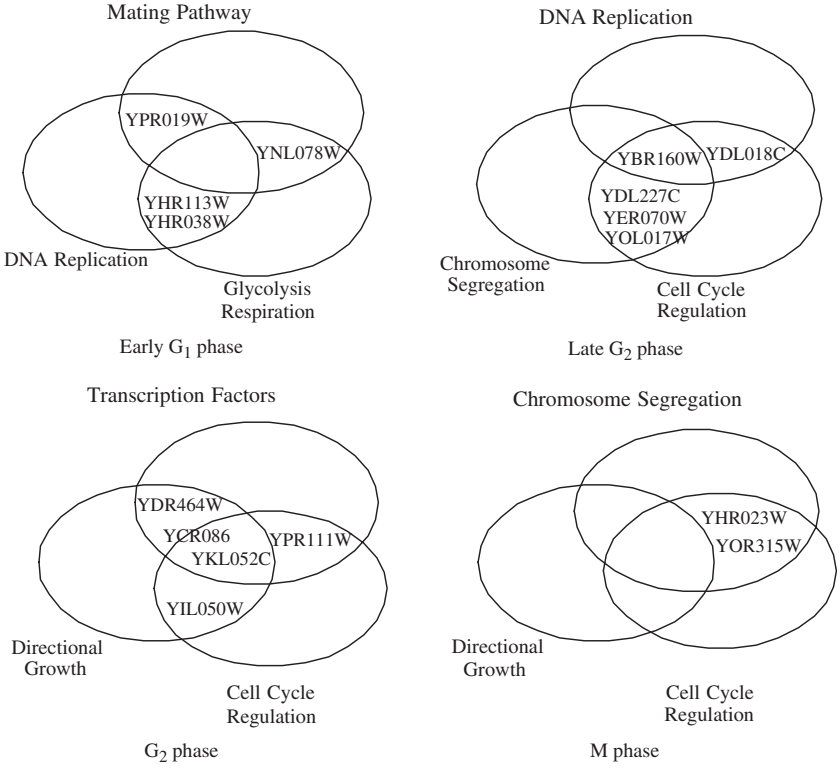


Fig. 5. Analysis of multiple functional genes.

Table 12
Related woks on DNA microarray data

| Author | Algorithm | Validity index | Data |
|---|---|---|---|
| Dembele and Kastner (2003) | Fuzzy *c*-means | Silhouette index | Serum Yeast cell-cycle human cancer |
| 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 |

approach to identify genes that perform different activities over time during the development of a cancer. In addition, deciding α-cut adaptively to the characteristics of the dataset can be considered as future work. The main challenge will be to theoretically interpret the fuzzy Bayesian validation method whether it is a probability or possibility indicator or even of some other nature.

Obviously, the evaluation of the cluster results is an important part for clustering, one of the analyzing methods for DNA microarrays, which can be used to identify the interactions of genes or other elements [1–3]. Systems biology, on the other hand, is a comprehensive quantitative analysis of the manner in which all the components of a biological system interact functionally over time [22]. The underlying principle is more crucial than the list of genes for systems biology, where interactions among them that could reveal the principles are important. We hope the BS approach to validate the clustering results can be applied to systems biology as a method to find underlying principles.

## Acknowledgements

This work was supported by Biometrics Engineering Research Center and a grant of Korea Health 21 R&D Project, Ministry of Health & Welfare, and Republic of Korea. The authors would like to thank Prof. Sushmita Mitra and the anonymous reviewers for their helpful comments to polish up the manuscripts.

## References

[1] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, P. Natl. Acad. Sci. USA 96 (12) (1999) 6745–6750.
[2] A.P. Gasch, M.B. Eisen, Exploring the conditional coregulation of yeast gene expression through fuzzy *k*-means clustering, Genome Biol. 3 (11) (2002) research 0059.1–0059.22.
[3] F. Höppner, F. Klawonn, R. Kruse, T. Runkler, Fuzzy Cluster Analysis, Wiley, New York, 2000.
[4] J.C. Bezdeck, Numerical taxonomy with fuzzy sets, J. Math. Biol. 1 (1974) 58–72.
[5] J.C. Bezdeck, Cluster validity with fuzzy sets, J. Cybernetics 3 (3) (1974) 58–72.
[6] X.L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE T. Pattern Anal. 3 (3) (1991) 841–846.
[7] Y. Fukuyama, M. Sugeno, A new method of choosing the number of clusters for the fuzzy *c*-means method, in: Proceedings of Fifth Fuzzy Systems Symposium, 1989, pp. 247–250.
[8] D.W. Kim, K.H. Lee, D.H. Lee, Fuzzy cluster validation index based on inter-cluster proximity, Pattern Recogn. Lett. 24 (2003) 2561–2574.
[9] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[10] S.L. Chiu, Fuzzy model identification based on cluster estimation, J. Intell. Fuzzy Syst. 2 (3) (1994) 267–278.
[11] M.K. Pakhira, S. Bandyopadhyay, U. Maulik, Validity index for crisp and fuzzy clusters, Pattern Recogn. 37 (2004) 487–501.
[12] K.-L. Wu, M.-S. Yang, A cluster validity index for fuzzy clustering, Pattern Recogn. Lett. 26 (2005) 1275–1291.
[13] N.R. Pal, J.C. Bezdeck, On cluster validity for the fuzzy *c*-means model, IEEE T. Fuzzy Syst. 3 (3) (1995) 370–379.
[14] M.R. Rezaee, B.P.F. Lelieveldt, J.H.C. Reiber, A new cluster validity index for the fuzzy *c*-means, Pattern Recogn. Lett. 19 (1998) 237–246.
[15] Y. Barash, N. Friedman, Context-specific Bayesian clustering for gene expression data, J. Comput. Mol. Cell Biol. 9 (2) (2001) 12–21.
[16] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, P.S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nat. Med. 7 (6) (2001) 673–679.
[17] 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, R.W. Davis, A genome-wide transcriptional analysis of the mitotic cell cycle, Mol. Cell 2 (1998) 65–73.
[18] N. Bolshakova, F. Azuaje, Cluster validation techniques for genome expression data, SIGPRO 21 (82) (2002) 1–9.
[19] D. Dembele, P. Kastner, Fuzzy *c*-means method for clustering microarray data, Bioinformatics 19 (8) (2003) 973–980.
[20] K.Y. Yeung, et al., Validating clustering for gene expression data, Bioinformatics 17 (4) (2001) 309–318.
[21] A.P. Gasch, M.B. Eisen, Exploring the conditional coregulation of yeast gene expression through fuzzy *k*-means clustering, Genome Biol. 3 (11) (2002) research 0059.1–0059.22.
[22] A. Aderem, Systems biology: its practice and challenges, Cell 121 (2005) 511–513.