퍼지 클러스터링의 베이지안 검증 방법을 이용한 발아효모 세포주기 발현 데이타의 분석

(Analysis of Saccharomyces Cell Cycle Expression Data using Bayesian Validation of Fuzzy Clustering)

유 시 호 † 원 홍 희 †† 조 성 배 †††
(Si-Ho Yoo) (Hong-Hee Won) (Sung-Bae Cho)

요 약 유전자를 분석하는 방법 중 하나인 클러스터링은 비슷한 기능을 가진 유전자들을 집단화시켜서 유전자 집단의 기능을 분석하는데 이용되고 있다. 유전자들은 다양한 functional family에 속할 수 있기 때문에 각 유전자의 클러스터를 하나로 결정짓는 기존의 클러스터링 방법보다 퍼지 클러스터링 방법이 유전자 클러스터링에 더 적합하다. 본 논문에서는 퍼지 클러스터 결과를 효과적으로 검증할 수 있는 베이지안 검증 방법을 제안한다. 베이지안 검증 방법은 확률기반의 방법으로 주어진 데이타에 대해 가장 큰 사후확률을 가진 클러스터 분할을 선택한다. 먼저 본 논문에서 제안하는 베이지안 검증 방법과 기존의 대표적인 4가지 퍼지 클러스터 검증 방법들을 4가지 데이타에 대해 퍼지 c-means알고리즘을 대상으로 비교 평가한다. 그리고 발아효모 세포주기 발현 데이타를 클러스터링한 후, 제안하는 방법으로 그 결과를 검증하여 분석한다.

키워드 : 퍼지 클러스터링, 퍼지 c-means알고리즘, 베이지안 검증 방법, 발아효모 세포주기 발현 데이타

Abstract Clustering, a technique 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 or 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 a group. In this paper, a Bayesian validation method is proposed to evaluate the fuzzy partitions effectively. Bayesian validation method is a probability-based approach, selecting a fuzzy partition with the largest posterior probability given the dataset. At first, the proposed Bayesian validation method is compared to the 4 representative conventional fuzzy cluster validity measures in 4 well-known datasets where fuzzy c-means algorithm is used. Then, we have analyzed the results of Saccharomyces cell cycle expression data evaluated by the proposed method.

Key words : fuzzy clustering, fuzzy c-means algorithm, Bayesian validation method, Saccharomyces cell cycle expression data

1. 서 론

클러스터링 분석은 방대한 유전자 정보를 비슷한 속성의 군집으로 나누어 분석할 수 있도록 해주기 때문에 유전자 발현 데이타 분석에 많은 도움을 준다[1]. 이 방법은 비슷한 기능을 가진 유전자들을 집단화시켜서 집단내의 유전자들의 기능을 밝히거나, 미지의 유전자의 기능을 분석하는데 이용되고 있다. 기존의 연구들에서 많이 사용되고 있는 클러스터링 방법인 하드 클러스터링 방법은 각 데이타를 하나의 클러스터에 소속시키는 하드 분할(hard partition)방식을 사용한다. 하지만 일반적으로 유전자 발현 암 데이타와 같은 실세계의 데이타는 쉽게 나뉘어지기 힘들거나 클러스터 간의 경계가 명확하지 않기 때문에 하드 클러스터링 기법은 주어진 데이타의 성질을 손실할 수 있다. 또한 유전자는 외부 조건에 따라 다양한 functional family에 속할 수 있기 때문에 하드 클러스터링 기법을 적용하여 그러한 유전자

· 본 논문은 보건복지부의 보건의료기술 진흥사업의 지원에 의하여 이루어진 것임
† 학생회원 : 연세대학교 컴퓨터과학과
bonanza@sclab.yonsei.ac.kr
†† 비 회 원 : 연세대학교 컴퓨터과학과
cool@sclab.yonsei.ac.kr
††† 종신회원 : 연세대학교 컴퓨터과학과 교수
sbcho@sclab.yonsei.ac.kr
논문접수 : 2004년 2월 10일
심사완료 : 2004년 9월 16일

의 특성을 분석하는데 한계가 있다[2]. 이에 반해 퍼지 클러스터링 기법은 각 데이타가 소속 정도에 따라 여러 개의 클러스터에 속할 수 있도록 분할한다[3]. 따라서 퍼지 클러스터링 기법은 하드 클러스터링 기법에 비하여 노이즈에 강하며, 유전자 발현 데이타를 분석하는 데 적합하다.

이때, 중요한 것은 클러스터링 알고리즘을 통해 얻어진 클러스터 분할이 실제로 얼마나 잘 형성되었으며 실제 데이타와 얼마나 잘 부합하는가이다. 이러한 클러스터 분할을 분석하는 검증 방법이 필요한데, 기존의 연구들로는 Bezdek이 제안한 Partition Coefficient (PC)[4], Partition Entropy (PE)[5], 그리고 최근의 연구에서 많이 사용되고 있는 Xie-Beni's Index[6] 등이 있다. 하지만, 기존의 검증 방법들은 클러스터들의 중심간의 거리에 초점을 맞추었기 때문에, 실제 데이타의 구조에 대한 해석에 한계를 가질 수 밖에 없었다[7].

본 논문에서는 퍼지 클러스터 결과를 체계적으로 평가하기 위하여 각 클러스터의 멤버쉽 값을 이용하는 베이지안 검증 방법을 제안한다. 베이지안 검증 방법은 거리 기반의 기존 검증 방법들과는 다른 확률 기반의 방법이다. 기존 방법들의 평가 기준이 클러스터 분할들의 separateness나 compactness에 있는 것에 반해, 제안하는 방법은 클러스터내의 데이타들에 대한 클러스터 분할이 형성될 사후확률이 얼마나 큰가를 기준으로 평가하는 방법이다. 먼저, 알려진 4가지 데이타들에 대해 본 논문에서 제안하는 베이지안 검증 방법과 기존 검증 방법들의 성능을 비교 평가하고, 퍼지 c-means알고리즘과 베이지안 검증 방법을 이용하여 발아효모 세포주기 발현 데이타를 분석한다.

본 논문의 나머지 부분은 다음과 같이 구성된다. 2장은 연구의 배경이 되는 퍼지 c-means 알고리즘과 기존 검증 방법들에 대하여 소개하고, 관련연구에 대해 언급한다. 3장은 제안하는 베이지안 검증 방법의 이론과 구체적인 방법에 대하여 기술한다. 4장은 실험 환경과 결과를 설명하고, 5장은 논문의 결론과 향후 연구에 대해 언급한다.

2. 배 경

클러스터링은 주어진 전체 데이타 집합을 유사한 성질을 갖는 몇 개의 클러스터로 분할하는 것이며, 대량의 데이타를 분석하는 데 용이하다. 클러스터링 알고리즘은 클러스터로 분할시키는 정도에 따라 하드 클러스터링 기법과 퍼지 클러스터링 기법으로 나눌 수 있다. 전자는 각 데이타를 하나의 클러스터에 소속시키는 하드 분할 방식을 사용하는데, 일반적으로 실세계의 데이타는 쉽게 나뉘어지기 힘들거나 클러스터 간의 경계가 분명하지 않기 때문에 주어진 데이타의 성질을 손실할 수 있다. 후자는 각 데이타가 소속 정도에 따라 여러 개의 클러스터에 속할 수 있도록 분할하기 때문에, 노이즈에 강하며 실세계의 데이타를 분석하는 데 적합하다. 퍼지 클러스터링 알고리즘으로 퍼지 c-means 알고리즘, Gustafson-Kessel 알고리즘, Gath-Geva 알고리즘 등이 있다[3,8]. 본 논문에서는 대표적인 퍼지 클러스터링 방법인 퍼지 c-means (FCM) 알고리즘을 사용하였다.

2.1 퍼지 c-means 알고리즘

퍼지 c-means 알고리즘은 Bezdek에 의해 제안된 가장 널리 이용되는 퍼지 클러스터링 방법이다[3,9]. 즉, 퍼지 이론을 적용한 목적 함수의 반복 최적화에 기반을 둔 방식으로 각 데이타가 특정 클러스터에 속하는 소속 정도를 줌으로써 데이타에 대한 보다 정확한 정보를 제공한다[10].

그림 1은 퍼지 c-means 알고리즘의 순서도이다. 먼

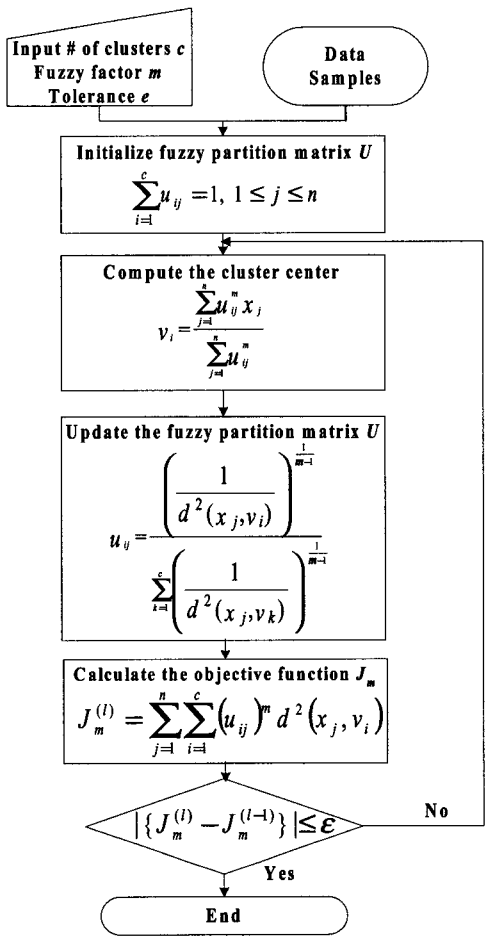

그림 1 퍼지 c-means 알고리즘

저 클러스터 수(c)와 퍼지 계수(m)의 값을 결정하고 x_j의 소속 정도인 u_{ij}를 합이 1이 되게 초기화 한다. x_j는 j번째 샘플을 의미하며, u_{ij}는 i번째 클러스터에 j번째 샘플이 속하는 소속정도를 의미한다. 이렇게 계산된 u_{ij}를 가지고 클러스터의 중심 v_i를 구한다. 마지막으로 클러스터의 중심과 x_j의 거리를 계산하여 소속 행렬 u_{ij}를 구한다. 전 단계의 목적함수와 현 단계의 목적함수의 차를 계산하여 그 차가 특정 임계값(ε)보다 작아질 때까지 위의 과정을 반복한다.

목적 함수는 주어진 데이타 집합이 $X = \{x_1, x_2,..., x_n\}$이고 퍼지 클러스터링의 중심 벡터가 $V = \{v_1, v_2,..., v_c\}$ 일 때, 각 데이타 x_j의 각 클러스터 중심 v_i와의 거리와 클러스터 소속 정도(membership) 값으로 정의된다.

$$J_m(X,U,V) = \sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^m d^2(x_j, v_i) \quad (1)$$

여기서 u_{ij}는 x_j의 i번째 클러스터에 대한 소속 정도를 나타내며 ($c \times n$)의 소속 행렬 $U = [u_{ij}]$의 원소이다. $d^2(,)$는 유클리디안 거리(Euclidean distance)의 제곱이고, 매개 변수 m은 각 데이타가 각 클러스터에 소속되는 정도를 의미하며 1 보다 큰 값을 사용한다.

2.2 클러스터 검증 방법

(1) Partition Coefficient (PC)

Partition Coefficient는 가장 많이 사용되고 있는 퍼지 클러스터링 평가 척도 중의 하나로 다음과 같은 간단한 계산을 통해 클러스터 결과를 검증한다[5].

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

식 (2)에서 u_{ij}는 멤버쉽 함수값이고 n은 샘플의 수, 그리고 c는 클러스터의 수이다. 이 방법은 각 경우에 대한 총합을 PC로 두고 그 값이 1에 가까워질수록 클러스터가 잘 형성된 것으로 본다. 하지만 c가 증가할수록 PC의 값이 단조 감소하는 현상을 보이는 한계를 가지고 있기 때문에 실제 다수의 클래스를 가진 경우가 많은 유전 데이타에 적용하기에는 부적합하다.

(2) Classification Entropy (CE)

Partition Coefficient와 더불어 가장 많이 사용되고 있는 퍼지 클러스터링 평가 척도 중의 한 방법이다[5].

$$CE(U;c) = \frac{-\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}\log_a u_{ij}}{n} \quad (3)$$

식 (3)을 보면 PC와 거의 비슷하지만, u_{ij}에 로그값을 취했다. 이 방법은 최종 결과값인 CE값이 작을수록 클러스터가 잘 형성된 것으로 본다. 이 방법 역시 PC와 마찬가지로 c값이 증가함에 따라 CE값이 단조 증가하는 현상을 보인다.

(3) Fukuyama-Sugeno (FS)

Fukuyama와 Sugeno가 개발한 클러스터 검증 방법으로 얼마나 해당 클러스터들이 compact한가를 측정한다[11].

$$FS(U,V;X) = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m\left(\left\|X_j - V_i\right\|^2 - \left\|V_i - \bar{V}\right\|^2\right),$$

$$\bar{V} = \frac{1}{n}\sum_{i}^{n}x_i \quad (4)$$

식 (4)에서 m은 퍼지한 정도를 나타내는 변수이고 X_j는 j번째 샘플, V_i는 i번째 클러스터의 중심을 나타낸다. 본 논문에서는 유전 데이타에 대하여 m=1.2 일 때 가장 적절한 클러스터 결과를 가져온다는 참고 논문[12]을 바탕으로 m=1.2로 설정하였다. 최종 결과값인 FS가 작은 값일수록 클러스터 결과가 좋은 것인데, 이 방법 역시 c가 증가할수록 FS값은 단조 감소하는 단점이 있다.

(4) Xie-Beni Index (XB)

Xie-Beni Index 방법은 퍼지 분할 결과가 얼마나 compact하고 separate한가를 측정하는 평가 척도이다[6]. compactness는 데이타의 총 변화의 가중치와 관측 회수의 비율에 의해서 결정되며, separation은 클러스터 간의 거리에 의해서 측정된다.

$$XB(U,V;X) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^2\left\|V_i - X_j\right\|^2}{nd_{\min}^2},$$

$$d_{\min} = \min_{i,j}\left\|V_i - V_j\right\| \quad (5)$$

식 (5)에서 d_{min}은 가장 가까운 클러스터간의 거리를 나타낸다. 이 방법은 최종값인 XB값이 작을수록 compact하고 separate한 정도가 큰 좋은 클러스터라고 할 수 있다.

2.3 DNA 마이크로어레이 클러스터링 관련연구

DNA 마이크로어레이 기반의 유전자 클러스터링 분야에는 표 1과 같이 다양한 연구가 진행되고 있다. Yeast 데이타를K-means 알고리즘과 Single-linkage알고리즘으로 분석한 Yeung의 연구[13]와 하드 클러스터링 알고리즘인 SOM과 하드 K-means알고리즘을 사용하여 Leukemia와 Lymphoma 데이타를 분석한 Bolshakova와 Azuaje의 연구[14]도 있다. Eisen [15]은 퍼지 k-means 알고리즘으로 같은 Yeast 데이타를 분석하였고, Dembele와 Kastner [12]는 퍼지 클러스터링 알고리즘을 통해 Serum과 Yeast 데이타를 분석하였다. 표 1에 나와있는 관련 연구들에서 사용한 검증 방법들은 모두가 거리 기반의 검증 방법들로 클러스터의 중심

표 1 DNA 마이크로어레이 클러스터링 관련 연구들

저자	사용한 알고리즘	검증 방법	데이타
Yeung et al. (2001)	K-means, Single-linkage	Figure of Merits	Yeast data
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 data
Dembele and Kastner (2003)	Fuzzy c-means	Silhouette index	Serum Yeast data Human cancer

간 거리계산이나, 클러스터 내 개체들간의 거리계산에 초점을 맞춘 방법들이다. Dunn's Index는 클러스터 내 샘플들의 거리(intracluster)와 클러스터간의 거리(inter-cluster)를 계산하여 클러스터 분할을 평가하는 방법이고 Silhouette index 역시 클러스터 내 모든 샘플들의 평균거리와 다른 클러스터 내 샘플들과의 평균거리를 이용한 방법으로 2.2에서 소개한 방법들과 유사한 수식을 사용한다.

3. 제안하는 검증 방법

퍼지 클러스터링의 결과를 검증하기 위해 제안된 기존의 검증 방법들은 클러스터가 얼마나 떨어져 있는가를 측정하기 위해서 클러스터의 중심간 거리나 클러스터내의 개체와 중심간의 거리를 계산하였다. 하지만, 이러한 거리 계산만으로는 실제 형성된 클러스터 분할을 잘 반영하지 못한다는 한계가 있다.

$$\lim_{c \to n} \left\| x_j - v_i \right\|^2 = 0 \tag{6}$$

식 (6)과 같이 클러스터 수(c)가 전체 샘플 수(n)에 가까워짐에 따라 클러스터의 중심과 어떤 샘플간의 거리는 0에 가까워진다. 위에서 x_j는 j번째 샘플, v_i는 i번째 클러스터의 중심을 의미한다. 따라서, 식 (6)과 같은 거리계산을 이용하는 기존의 검증 방법들은 클러스터 수(c)가 많아짐에 따라 단조 감소하는 경향을 보인다 [16,17].

베이지안 검증 방법은 확률기반의 검증 방법으로, 데이타가 주어졌을 때 해당 데이타에 대한 클러스터 분할의 사후확률을 구하여 클러스터 결과를 검증하는 방법이다[18]. 베이지안 검증 방법은 이처럼 주어진 데이타에 대해 각 클러스터의 사후확률이 최대가 되는 것을 최적의 클러스터 분할로 한다.

$$\max P(Cluster \mid Dataset) \tag{7}$$

Bayes' Theorem을 적용하면 다음과 같이 사전확률을 이용하여 사후확률 값을 구할 수 있다.

$$P(Cluster \mid Dataset) = \frac{P(Cluster)P(Dataset \mid Cluster)}{P(Dataset)} \tag{8}$$

각 데이타가 서로 독립이라 가정하면 multiplication rule과 independence rule에 의해 식 (8)은 다음의 식 (9)와 같이 표현될 수 있다.

$$\begin{aligned} P(Cluster \mid Dataset) &= P(Cluster \mid d_1, d_2, ..., d_N) \\ &= P(Cluster \mid d_1) \times P(Cluster \mid d_2) \times ... \times P(Cluster \mid d_N) \end{aligned} \tag{9}$$

이러한 과정을 이용하여 식 (10)과 같이 모든 클러스터에 대한 $P(Cluster|Dataset)$들의 합을 구하여 이를 베이지안 스코어(Bayesian score)라고 정의한다. 이 베이지안 스코어(BS)는 그 값이 클수록 각 클러스터의 사후확률이 커지므로 좋은 클러스터 분할을 나타낸다고 볼 수 있다.

$$\begin{aligned} BS &= \frac{\sum_{i=1}^{c} P(C_i \mid D_i)}{C} = \frac{\sum_{i=1}^{c} P(C_i \mid d_{i1}, d_{i2}, ..., d_{iN})}{C} \\ &= \frac{\sum_{i=1}^{c} P(C_i \mid d_{i1}) P(C_i \mid d_{i2}) ... P(C_i \mid d_{iN})}{C} \\ &= \frac{\sum_{i=1}^{c} \prod_{j=1}^{N_i} P(C_i) P(d_{ij} \mid C_i) / P(d_{ij})}{C}, \\ D_i &= \{ d_{ij} \mid \mu_{ij} > \alpha, 1 \le j \le n \}, \; N_i = \mathrm{n}(D_i) \end{aligned} \tag{10}$$

여기서 $\mathrm{n}(D_i)$는 D_i의 개수이며, 일정한 확률보다 큰 멤버쉽값($u_{ij} > \alpha$)을 가진 샘플들만을 선택한다. 그 이유는 BS의 계산과정 중에는 곱셈 계산이 있기 때문에 u_{ij}=0인 샘플들의 경우 올바른 값이 나올 수가 없다. 또한 퍼지 클러스터링을 하는 궁극적인 이유는 명확하지 않은 개체들의 소속 정도를 분석하기 위함인데, 모든 개체에 대해 검증을 하는 것보다는 특정한 임계값 이상의 소속 정도를 가진 개체들로 클러스터 분할을 평가하는 것이 더 정확하기 때문이다. 이러한 임계값을 α-cut이라 하고, 각 확률은 다음과 같이 계산될 수 있다.

$$P(C_i) = \frac{\sum_{j=1, u_{ij} > \alpha}^{n} u_{ij}}{\sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}} \tag{11}$$

$$P(d_{ij}) = \sum_{i=1}^{c} P(C_i) P(d_{ij}) = \sum_{i=1}^{c} P(C_i) u_{ij} \tag{12}$$

퍼지 클러스터링 결과로 멤버쉽 행렬이 주어질 때, 이의 멤버쉽 값은 각 샘플이 각 클러스터에 속할 확률이라고 볼 수 있다. 그러므로 각 샘플의 멤버쉽값 u_{ij}를 $P(d_{ij}|C_i)$로 정의할 수 있다. 그림 2는 베이지안 검증 방법의 전체적인 개요인데, 여기서 D_1은 c_1에 속하면서 $u_{ij}>\alpha$인 조건을 만족하는 샘플들의 집합을 나타내고 최종적으로 Bayesian score (BS)를 계산하여 클러스터를 검증한다.

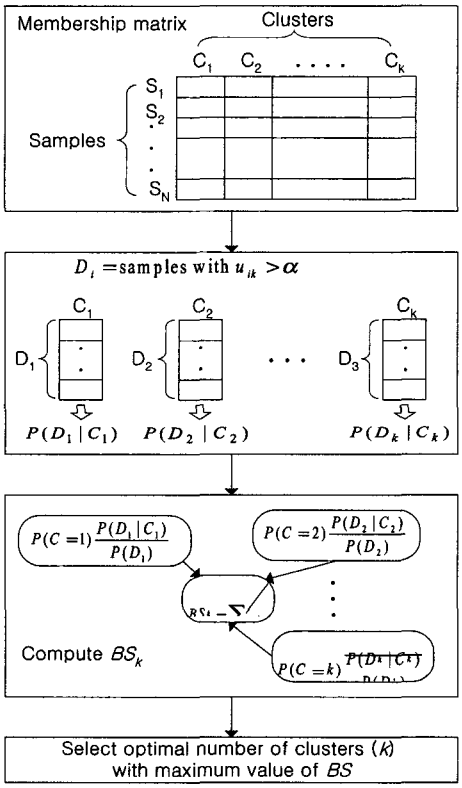

그림 2 베이지안 검증 방법의 개요

베이지안 검증 방법의 알고리즘은 다음과 같다. 이때, 베이지안 검증 방법의 최종 결과값인 BS는 0과 1사이의 확률값을 가지며, 가장 큰 값을 가진 분할을 최적의 클러스터 분할로 평가한다.

- 단계 1: 퍼지 클러스터링의 결과인 멤버쉽 행렬 U_{ij}를 구한다.
- 단계 2: U_{ij}에서 $u_{ij}>\alpha$를 만족하는 샘플들을 각 클러스터 별로 선택한 샘플들의 집합인 D_j를 구한다.
- 단계 3: 단계2)에서 선택한 D_j에 대해 $P(D_j|C_j)$, $P(D_j)$, $P(C_j)$ 값을 계산한다.
- 단계 4: 단계3)에서 계산한 값을 이용하여 BS를 계산한다.
- 단계 5: BS의 값을 최대로 하는 클러스터 분할을 최적의 분할로 평가한다.

4. 결 과

4.1 실험 데이타

기존 퍼지 클러스터링 검증 방법들과 본 논문에서 제안하는 베이지안 검증 방법의 성능을 비교하기 위해서 우선 알려진 4가지 데이타에 대해서 실험하였다. 사용한 데이타는 UCI Machine Learning Repository(http://www.ics.uci.edu/~mlearn/MLSummary.html)에 있는 Iris, Wine, Image데이타와 유전 발현 데이타인SRBCT 데이타이다.

Iris는 식물의 줄기와 관련된 데이터로 3개의 클래스를 가지고 있으며, 150개의 샘플과 4개의 속성을 가지고 있다. 3개의 클래스 중에서 2개의 클래스는 상당 부분 겹치는 클래스로 보통 2~3개의 클래스로 보기도 한다. Wine데이타는 총 3가지 클래스를 가지고 있으며, 178개의 샘플과 13개의 속성을 가지고 있다. Image 데이타는 7개의 클래스를 가지고 있고, 210개의 샘플과 19개의 속성으로 이루어져 있다. 마지막 SRBCT 데이타는 총 4개의 클래스를 가지고 있으며 모두가 암의 종류와 관련된 샘플들이다. 이 데이타는 63개의 샘플과 96개의 속성을 가지고 있다[19].

위의 4가지 데이타를 가지고 제안하는 방법과 기존 검증 방법을 비교하여, 제안하는 베이지안 검증 방법의 성능을 확인한 후, 발아효모 세포주기 발현 데이타에 대해 실험하였다. 효모의 세포주기 유전자 발현 데이타는 두 개의 세포 주기를 거치는 동안의 약 6000개 유전자들의 발현 정도를 나타내는 데이타이다. 본 논문에서는 세포 주기별로 발현 정도의 차이를 보이는 유전자 421개(http://yscdp.stanford.edu/yeast_cell_cycle/cellcycle.html)만을 사용하여 실험하였다. 즉, 유전자 세포 주기의 17개 시점에서 선택된 발현 데이타를 이용하여 클러스터링하였다[20].

대부분의 클러스터링 알고리즘은 각 샘플간의 유사도를 기반으로 클러스터링한다. 유사도는 각 샘플의 속성값에 따라 결정되기 때문에, 편차가 큰 속성이 존재할 경우 전체의 유사도를 그 속성이 결정하게 된다. 본 논문에서는 식 (13)을 사용하여 각 유전자의 발현 정도를 [0, 1] 범위로 정규화하였으며, $\min(e_i)$는 각 유전자의 발현 정도 e_i의 최소값이고, $\max(e_i)$는 e_i의 최대값을 의미한다.

$$e_i = \frac{e_i - \min(e_i)}{\max(e_i) - \min(e_i)} \tag{13}$$

4.2 실험 결과

(1) 성능 비교 실험

베이지안 검증 방법의 BS값을 계산할 때, α-cut을 0.1~0.6까지 설정하고 실험하여 그 평균값을 최종 BS값으로 평가하였다. 표 2는 Wine데이타의 실험 결과이다. Wine데이타는 3개의 클래스를 가진 데이타로, 표 2에서 보는 바와 같이 *PC*와 *CE*는 c=2를 최적의 클러스터 분할로 평가 하였고, *FS*는 c=8, *XB*는 c=3, 제안하는 베이지안 검증 방법(*BS*)은 c=4를 최적의 분할로 평가하였다. *XB*를 제외한 나머지 모든 검증 방법들은 실제 클래스인 c=3을 맞추지 못하였다. 제안하는 베이지안 검증 방법 역시, 최적의 클러스터 수를 c=4로 평가하였지만, c=3일 때와 c=4일 때의 *BS*값의 차(0.0066)가 매우 작기 때문에 c=3 또는 c=4를 최적의 클러스터 분할로 평가한 것으로 볼 수 있다. *DI*는 Dunn's index로 2.3의 관련 연구와 비교하기 위해서 실험하였다. *DI*는 *PC*, *CE*와 비슷한 양상을 보이면서 c=2를 최적의 클러스터 분할로 평가한다.

표 3은 Image데이타의 실험 결과이다. 최적의 클러스터 분할인 c=7은 제안하는 방법인 *BS*만이 맞추고 있고, 나머지 기존 검증 방법들은 전부 다른 클러스터 분할을 최적의 분할로 평가하였다. *PC*는 c=4, *CE*와 DI는 c=2, *FS*는 c=8, 그리고 *XB*는 c=2에서 최적의 값을 보여주고 있다.

표 4와 표 5는 각각 Iris데이타와 SRBCT데이타에 대한 결과를 나타내고 있다. Iris데이타의 경우, c=2 또는 c=3에서 최적의 분할을 형성한다. 표 4에서 *FS*를 제

표 2 Wine데이타 실험 결과(c=3)

c	PC	CE	FS	XB	*DI*	BS
2	**0.9358**	**0.0476**	29.3216	0.4516	**1.4520**	0.1583
3	0.9258	0.0564	-1.5084	**0.4312**	1.3314	0.2707
4	0.8814	0.0932	-8.9773	0.8636	0.945	**0.2773**
5	0.8308	0.1344	-12.2939	1.3137	0.7487	0.2556
6	0.8180	0.1494	-15.3106	1.5032	0.7254	0.2477
7	0.7964	0.1673	-18.4707	1.3638	0.6162	0.2160
8	0.7937	0.1748	**-20.8222**	1.2147	0.6696	0.2719

표 3 Image데이타 실험 결과(c=7)

c	PC	CE	FS	XB	DI	BS
2	0.9468	**0.0379**	35.2037	**0.2617**	**1.8185**	0.2190
3	0.9270	0.0550	-22.3275	0.5519	1.2417	0.3846
4	**0.9539**	0.0384	-65.1495	0.3935	1.4623	0.4120
5	0.9448	0.0464	-74.4622	0.3895	1.0081	0.3918
6	0.9292	0.0599	-85.9394	0.7165	0.5850	0.4672
7	0.8980	0.0866	-91.0603	0.8023	0.5734	**0.5490**
8	0.9224	0.0657	**-104.061**	0.5720	0.5828	0.5077

표 4 Iris데이타 실험 결과(c=2)

c	PC	CE	FS	XB	DI	BS
2	**0.9916**	**0.0060**	-311.725	**0.0619**	**3.9295**	**0.7512**
3	0.9781	0.0156	-426.900	0.1539	2.4267	0.5825
4	0.9704	0.0226	-459.027	0.2189	1.8279	0.4494
5	0.9569	0.0331	-459.622	0.5045	1.1272	0.3046
6	0.9560	0.0333	-462.487	0.9038	1.3143	0.4712
7	0.9510	0.0366	**-481.403**	0.6820	1.3141	0.4610

표 5 SRBCT데이타 실험 결과(c=4)

c	PC	CE	FS	XB	DI	BS
2	0.8758	0.0969	164.4075	1.1294	0.8338	0.1918
3	0.9205	0.0709	86.6529	0.8127	0.9454	0.5612
4	**0.9393**	**0.0616**	27.3224	**0.5657**	**1.1721**	**0.7073**
5	0.9100	0.0850	-0.9891	0.8487	0.7477	0.6731
6	0.8922	0.0977	-22.6041	0.7798	0.7405	0.6411
7	0.8989	0.0979	**-34.773**	0.8670	0.6908	0.6852

외한 나머지 방법들이 모두 c=2에서 최적의 값을 보여주고 있다. *PC, CE, XB, DI* 모두 비슷한 양상을 보인다. SRBCT데이타의 경우, c=4에서 최적의 분할을 형성하는데 *BS*를 포함한 *PC, CE, XB*등의 방법들이 c=4에서 최적치를 보이지만, *FS*의 경우는 c=7에서 최적치를 보인다.

총 4가지 데이타에 대한 검증 방법들의 성능평가결과, *FS*의 성능이 가장 떨어졌으며, *PC, CE, XB, DI*등 기존 방법들은 Wine과 Image데이타에서 c의 값이 증가함에 따라 단조 감소 또는 증가하는 경향을 보였다. 제안하는 베이지안 검증 방법의 경우, Wine데이타의 경우를 제외하고는 모두 올바른 클러스터 분할을 찾아내었으며, c의 값에 따라 단조 감소 또는 증가하는 경향을 보이지 않았다.

(2) 발아효모 세포주기 발현 데이타의 분석

발아효모 세포주기 발현 데이타에 대한 클러스터 검증 실험 결과는 그림 3과 같다. 그림 3에서 x축은 클러스터 수를, y축은 검증 방법들의 평가값을 나타낸다. 각 검증 방법에 대하여 클러스터 수를 5부터 35까지 증가시켜가면서 평가하였다.

실험 결과, *PC*와 *CE* 방법은 c=5에서 최적으로 평가되었고, *FS*는 c=35에서, *XB*는 c=13 근처에서 최적으로 평가되었다. *DI*의 경우는 c=7 정도에서 최고치를 보이다가 감소-증가를 반복하는 양상을 보인다. 반면에 베이지안 검증 방법은 c=29근처에서 최적 평가값을 가졌다. 모든 방법마다 중간에 지역적인 해를 보이는데 이것은

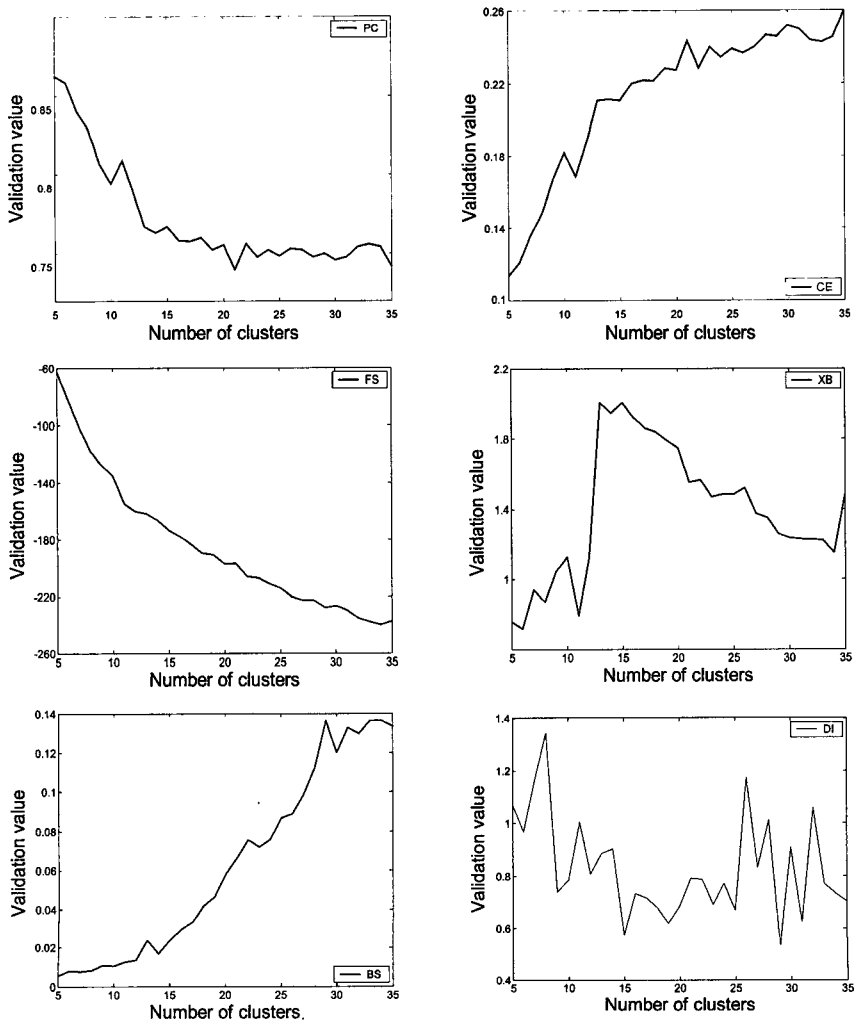

그림 3 클러스터 수에 따른 검증 결과(발아효모 세포주기 발현 데이타)

발아효모 세포주기 발현 데이타 자체의 특성상 매 주기마다 반복되는 유전자들의 유사성 때문인 것으로 해석된다. 본 논문에서는 가장 큰 값을 갖는 c=29일 때를 최적의 클러스터 수로 정하여 결과를 분석하였다.

발아효모 세포주기 발현 데이타는 세포의 5주기(Early G_1 - Late G_1 - S - G_2 - M)에 따른 각 유전자의 발현 정도를 나타내기 때문에 주기를 갖는 유전자 클러스터가 중요하다고 볼 수 있다. 세포의 5주기는 실제 세포가 성장하는 과정을 시간에 따라 분류한 주기로, 각 주기에 따라 발현되는 유전자들이 차이를 보이기 때문에 생성된 클러스터들은 특정한 주기에서 소속된 유전자들의 발현이 최대가 되는 형태의 주기성을 보인다.

각 주기에서 최고점을 갖는 클러스터들은 표 6과 같으며, 특정 주기에서 최고점을 갖는 클러스터는 다른 주기에서는 대부분 낮은 발현 정도를 보였다. 각 주기와 주기 사이에 최고점을 갖는 클러스터에 속한 유전자들은 다음 주기에 발현될 유전자들의 regulator 역할을 할 것으로 예상할 수 있다. 표 6에서 intercourse는 각 주기 사이에 중복되어 존재하는 주기를 뜻하며, 유전자들의 발현 정도가 각 주기의 사이에서 최대치를 보이는 클러스터들이 포함된다.

표 7은 생물학적으로 알려진 유전자들의 세포 주기에 따른 실제 소속을 보여준다[20]. 퍼지 클러스터링 실험 결과로 얻은 결과가 생물학적으로 의미가 있는지를 확

표 6 발아효모 세포주기 발현 데이타의 퍼지 클러스터링 실험 결과 정리(c=29)

실험 시간 (×10 min)	세부 주기	각 주기에서 최고점을 보이는 클러스터
	intercourse	Cluster20 Cluster19
0-3	G_1 기	Cluster5 Cluster6 Cluster4 Cluster24
	intercourse	Cluster2 Cluster12 Cluster26 Cluster28
3-5	S 기	Cluster8 Cluster13 Cluster14 Cluster16
	intercourse	Cluster11
5-7	G_2 기	Cluster13
	intercourse	Cluster18
7-9	M 기	Cluster7 Cluster17
	intercourse	Cluster10 Cluster21 Cluster3 Cluster20 Cluster19
9-11	G_1 기	Cluster5 Cluster6 Cluster4 Cluster24
	intercourse	Cluster2 Cluster12 Cluster26 Cluster28
11-13	S 기	Cluster8 Cluster13
	intercourse	Cluster11
13-15	G_2 기	Cluster0 Cluster13
	intercourse	Cluster18
15-17	M 기	Cluster7 Cluster17

표 7 생물학적으로 알려진 유전자들의 세포 주기에 따른 분류

세포 주기	상세 기능	소속된 유전자 및 클러스터 번호
Early G_1 기	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 기	Cell cycle regulation	YBR160W(12) YDL127W(12) YGR109C(12)YPR120C(12)
	Chromosomesegregation	YDL003W(26) YFL008W(26) YJL074C(26)YKL042W(26) YMR076C(26) YMR078C(26)
	DNA replication	YBR278W(24) YKL045W(24) YLR103C(24)YPR018W(24)
S 기	Chromosomesegregation	YDR113C(16) YGR140W(16) YHR172W(16)
	DNA replication	YBL002W(8) YBL003C(8)
	Miscellaneous	YCR035C(14) YER016W(14) YJR137C(14)
G_2 기	Directional growth	YJL099W(11) YJR076C(11)
	DNA replication	YDR224C(27) YDR225W(27)
M 기	Cell cycle regulation	YGL116W(7) YPR119W(7)
	Transcriptional factor	YDR146C(18) YLR131C(18)
	Directional growth	YCL037C(17)

인하기 위하여 각 클러스터에 속한 유전자와 생물학적으로 알려진 기능을 분석해보았다. 모든 비교 분석은 이미 생물학적 의미가 밝혀진 유전자들의 결과[20]와 비교하여 분석하였다. 하지만 아직 그 기능군이 알려지지 않은 유전자들을 포함한 클러스터들의 경우는 표 7의 분석에서 제외하였다.

29개의 클러스터로 발아효모 세포주기 발현 데이타를 실험하고 그 중에서 특별히 가장 큰 멤버쉽 값이 0.35이상 0.7미만이고 두 번째로 큰 멤버쉽 값이 0.3 이상인 유전자들만을 선택하였다. 이런 유전자들은 동시에 다수의 클러스터에 속할 수 있는 퍼지 유전자로, 유전자 분석에 있어서 의미 있는 정보들을 제공한다. 표 8은 이러한 퍼지 유전자들의 소속정도를 보여주고 있으며 그림 4는 이 중에서도 생물학적인 기능이 밝혀진 유전자들을 분석한 결과이다. 그림 4에는 각 유전자들의 설명이 나와있고, 유전자들이 소속된 클러스터 번호가 나타나 있다. 클러스터 번호가 유사한 유전자들끼리 총 4개 그룹으로 나누어서 분석을 하였다. 3, 10, 20, 21등의 클러스터에 속하는 유전자들은 Early G_1 기에 관련되어있고, 12, 24, 26등의 클러스터에 속하는 유전자들의 경우는 Late G_2 기에 관련되어 있다. 9, 11, 13번째 클러스터는 G_2 기에 그리고 7, 18등의 클러스터는 M 기에 관련된 유전자들을 포함하고 있다.

분석 결과, cluster3은 early G_1 기의 mating pathway에 관련된 유전자들의 집단이고, cluster19는 역시 같은 early G_1 기의 glycolysis respiration에 관련된 유전자들의 집합으로 판명되었다. 따라서 YNL078W는 퍼지 유전자로 다수의 기능을 가지고 있음을 확인 할 수 있다. 또한 YPR019W, YHR038W, YHR113W와 같은 유전자들도 early G_1 기에 속하면서 다수의 세부 클러스터에 소속되는 퍼지 유전자들이다. Late G_2 기의 YBR160W같은 유전자는 cluster12(0.398234)와 cluster6(0.34645)에 동시에 속하며 cluser12는 cell cycle regulation에, cluster6은 chromosome segregation에 관련된 클러스터이다. G_2 기의 퍼지 유전자로는 YIL050W, YCR086, YDR464W, YKL052C, 그리고 YPR111W등이 있다. 이러한 퍼지 유전자들은 그림 4에 나와있는 것처럼 다수의 클러스터에 속한다. M 기의 퍼지 유전자로는 YHR023W와 YOR315W등이 있다. YHR023W는 cluster18(0.665914)과 cluster7(0.32451)에 동시에 속하는 퍼지 유전자로 M 기의 chromosome segregation과 cell cycle regulation집단에 동시에 속한다.

표 8 퍼지 유전자 테이블

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

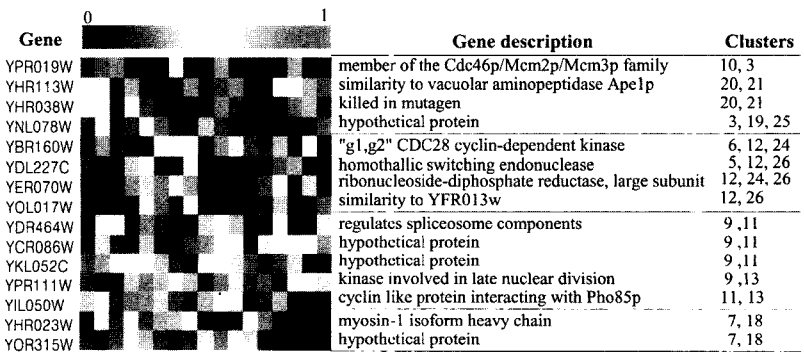

그림 4 퍼지 유전자 분석(유전자 설명, 소속 클러스터 번호)

그림 5에서는 그림 4에서 분석한 각 퍼지 유전자들이 속하는 클러스터가 어떤 분포를 보이는가를 소속 클러스터별로 나타내 보았다. 각 퍼지 유전자는 검은 크로스 모양으로 표시하였다. Early G_1 기의 Cluster20과 Cluster21사이에 YHR113W와 YHR038W가 분포하는 것을 확인할 수 있다. 그리고 Late G_2 기와 M 기, G_2 기에 있는 퍼지 유전자들도 각 주기 내에 존재하는 클러스터의 경계면에 존재하는 분포를 보인다. 다수의 기능군에 속하는 퍼지 유전자들의 경우, 그림 5와 같이 명확한 경계를 형성하지 않으며 다수의 클러스터에 중복되게 소속되는 것을 확인할 수 있다.

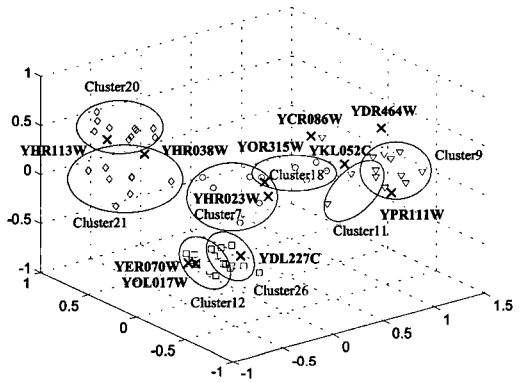

그림 5 퍼지 유전자 분석(PCA를 사용하여 유전자들의 속성을 3차원으로 줄여서 나타냄)

5. 결 론

본 논문에서는 새로운 퍼지 클러스터 검증 방법인 베이지안 검증 방법을 제안하여 발아효모 세포주기 발현 데이타를 분석하였다. 특히 퍼지 클러스터링의 결과로 얻어지는 멤버쉽 행렬로부터 각 유전자의 클러스터 소속을 결정할 때, α-cut값을 임계값으로 두어 그 이상의 값을 갖는 유전자들만을 선택하여 클러스터링 함으로써 유전자들의 다양한 클러스터 형성을 가능하게 하고, 보다 정확한 검증결과를 얻을 수 있었다.

기존의 클러스터링 검증 방법 4가지(PC, CE, FS, XB)와 본 논문에서 제안한 베이지안 검증 방법을 비교 평가하였는데, 베이지안 검증 방법은 기존 방법들 보다 우수한 클러스터 평가 성능을 보였고, 그 신뢰도 역시 기존 방법에 비해 떨어지지 않는 것을 확인하였다. 본 논문에서는 발아효모 세포주기 발현 데이타에 대해 제안하는 방법을 적용하여 다양한 functional family에 소속된 유전자들을 분석하고 이를 생물학적으로 검증하여 제안한 방법의 유용성을 확인하였다.

향후 연구 과제로는 보다 적응적인 α-cut을 결정하는 방법이나 데이타 자체로부터 α-cut을 결정하는 데 필요한 정보를 추출하는 방법을 들 수 있다. 이렇게 향상된 방법을 레이블이 알려지지 않은 데이타나 기능이 알려지지 않은 유전자 분석에 적용하면 보다 유용한 결과를 얻을 수 있을 것으로 기대된다. 또한 생물학적으로 알려진 유전자들의 기능군과 실험 결과로 나온 클러스터내의 유전자들이 어느 정도 매칭이 되는가를 통계적으로 분석하는 방법이 정확한 결과 분석에 필요하다.

참 고 문 헌

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc Natl Acad Sci*, USA, vol. 96, no.12, pp. 6745-6750, 1999.

[2] 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.

[3] F. Höppner, F. Klawonn, R. Kruse and T. Runkler, *Fuzzy Cluster Analysis*, Wiley, 2000.

[4] J. C. Bezdeck, "Numerical taxonomy with fuzzy sets," J. *Math. Biology*, vol. 1, pp. 58-72, 1974.

[5] J. C. Bezdeck, "Cluster validity with fuzzy sets," *J. Cybernit*, vol. 3, pp. 58-72, 1974.

[6] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Trnas. Pattern Analysis and Machine Intelligence*, vol. 3, no. 3, pp. 841-846, 1991.

[7] D. W. Kim, K. H. Lee and D. H. Lee, "Fuzzy cluster validation index based on inter-cluster proximity," *Pattern Recognition Letters*, vol. 24. pp. 2561-2574, 2003.

[8] I. Gath and A.B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, pp. 773-781, 1989.

[9] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, 1989.

[10] S. L. Chiu, "Fuzzy model identification based on cluster estimation," *J. Intelligent and Fuzzy Systems*, vol. 2, no. 3, pp. 267-278, 1994.

[11] 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.

[12] D. Dembele and P. Kastner, "Fuzzy c-means method for clustering microarray data," *Bioinformatics*, vol. 19, no. 8, pp. 973-980, 2003.

[13] K. Y. Yeung, et al., "Validating clustering for gene expression data," *Bioinformatics*, vol. 17, no.

4, pp. 309-318, 2001.
[14] N. Bolshakova and F. Azuaje, "Cluster validation techniques for genome expression data," *Signal Processing*, vol. 21, no. 82, pp. 1-9. 2002.
[15] 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.
[16] N. R. Pal and J. C. Bezdeck, "On cluster validity for the fuzzy c-means model," *IEEE Trans. Fuzzy Systems*, vol. 3, no. 3, pp. 370-379, 1995.
[17] M. R. Rezaee, B. P. F. Lelieveldt and J. H. C. Reiber, "A new cluster validity index for the fuzzy c-means," *Pattern Recognition Letters*, vol. 19, pp. 237-246, 1998.
[18] Y. Barash and N. Friedman, "Context-specific Bayesian clustering for gene expression data," *Journal of Computational Molecular Cell Biology*, vol. 9, no. 2, pp. 12-21, 2001.
[19] 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.
[20] 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.

유 시 호
정보과학회논문지 : 소프트웨어 및 응용
제 31 권 제 9 호 참조

원 홍 희
2002년 2월 연세대학교 컴퓨터과학과(학사). 2004년 2월 연세대학교 컴퓨터과학과(석사). 관심분야는 바이오정보기술, 패턴인식

조 성 배
정보과학회논문지 : 소프트웨어 및 응용
제 31 권 제 1 호 참조