진화연산과 적응적 α-cut 기반 평가를 이용한 유전자 발현 데이타의 퍼지 클러스터 분석

(Fuzzy Cluster Analysis of Gene Expression Profiles Using Evolutionary Computation and Adaptive α-cut based Evaluation)

박 한 샘 † 조 성 배 ††
(Han-Saem Park) (Sung-Bae Cho)

요 약 유전자 데이타의 클러스터링은 방대한 유전자 정보를 발현 정도에 따라 비슷한 그룹으로 나누어 분석하는 방법으로 유전자의 기능을 분석하는데 사용되어 왔다. 클러스터링의 한 종류인 퍼지 클러스터링은 하나의 샘플이 소속정도에 따라 여러 그룹에 동시에 소속되도록 나누는 방법으로, 하나의 유전자 데이타는 여러가지 유전 정보를 가질 수 있기 때문에 유전자 발현 데이타의 분석에 보다 적절한 방법이다. 그러나 보통 클러스터링 방법은 초기 값에 민감하고, 지역해에 빠질 수 있는 단점을 갖는다. 이런 단점을 해결하기 위해 본 논문에서는 진화 연산을 이용한 퍼지 클러스터링 방법을 제안한다. 이때, 적합도 평가를 위해서 모든 데이타에 대해 동일한 기준을 적용하는 베이지안 검증방법의 단점을 개선하여, 데이타의 특성을 고려하여 결정된 적응적 α-cut 기반 평가방법을 사용한다. SRBCT 데이타와 효모 세포주기 데이타를 이용해 실험을 하고 결과를 분석하여 제안하는 방법의 유용성을 확인하였다.

키워드 : 진화적 퍼지 클러스터링, 적응적 α-cut 기반 평가, 유전자 발현 데이타

Abstract Clustering is one of widely used methods for grouping thousands of genes by their similarities of expression levels, so that it helps to analyze gene expression profiles. This method has been used for identifying the functions of genes. Fuzzy clustering method, which is one category of clustering, assigns one sample to multiple groups according to their degrees of membership. This method is more appropriate for analyzing gene expression profiles because single gene might involve multiple genetic functions. Clustering methods, however, have the problems that they are sensitive to initialization and can be trapped into local optima. To solve these problems, this paper proposes an evolutionary fuzzy clustering method, where adaptive α-cut based evaluation is used for the fitness evaluation to apply different criteria considering the characteristics of datasets to overcome the limitation of Bayesian validation method that applies the same criterion to all datasets. We have conducted experiments with SRBCT and yeast cell-cycle datasets and analyzed the results to confirm the usefulness of the proposed method.

Key words : evolutionary fuzzy clustering, adaptive α-cut based evaluation, gene expression profiles

1. 서 론

클러스터링은 방대한 유전자 정보를 비슷한 속성의 그룹으로 나누어 분석할 수 있도록 해 줌으로써 유전자 발현 데이타를 분석하는데 유용하다. 이 방법은 비슷한 기능을 가진 유전자들의 집단을 형성하여, 집단 내 유전자들의 기능을 밝히거나, 미지의 유전자를 분석하는데 이용되고 있다[1]. 그러나 일반적으로 실세계의 데이타로부터 명확한 경계를 가진 집단을 구성하기는 어렵다[2]. 하나의 유전자가 여러 가지 유전 정보를 동시에 가질 수 있는 유전자 데이타는 그 대표적인 예라고 할 수 있으며, 퍼지 클러스터링은 이러한 유전자 발현 데이타

· 본 연구는 생체인식연구센터(BERC)를 통해 한국과학재단(KOSEF)에서 지원받았음
† 학생회원 : 연세대학교 컴퓨터과학과
sammy@sclab.yonsei.ac.kr
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
sbcho@cs.yonsei.ac.kr
논문접수 : 2005년 7월 22일
심사완료 : 2006년 6월 14일

를 분석하는 데 유용한 방법이다[3].

일반적인 클러스터링 알고리즘은 초기값에 매우 민감하며 목적 함수를 최소화시키는 방향으로 알고리즘이 진행되기 때문에 지역해에 빠지기 쉬운 문제점이 있다[4,5]. 또한 클러스터의 수를 고정시키고 실험을 하기 때문에 데이타에 대한 사전 지식이 없으면 올바른 분석을 하기가 어렵다. 예를 들어 클러스터 수를 모르는 데이타를 클러스터링하려면 여러 번 반복해서 실험해야 하기 때문에 시간비용이 커지게 된다. 이 외에 클러스터 결과의 검증에 대한 문제도 존재한다. 서로 다른 환경에서 수집된 데이타는 다른 특성을 가지기 때문에 모든 데이타를 동일한 기준에 의해 평가하는 것은 적절하지 못하다.

본 논문에서는 이와 같은 문제점을 해결하기 위해 진화연산 기법을 이용한 퍼지 클러스터링 방법과 데이타의 특성을 고려한 적응적 α-cut 기반 평가 방법을 제안한다. 최적화 문제를 해결하는데 뛰어난 성능을 보이는 유전자 알고리즘(genetic algorithm)[6]을 사용하여, 초기값에 덜 민감하고 보다 최적해에 근접한 클러스터링을 할 수 있다[7]. 클러스터링에 진화 연산을 적용한 연구는 많이 진행되어 왔는데, 클러스터 중심과 클러스터 내 개체들의 거리를 최소화시키기 위해 유전자 알고리즘을 사용한 Maulik의 연구[5], 하드 c-means 알고리즘과 퍼지 c-means 알고리즘의 목적함수를 최소화시키는데 유전자 알고리즘을 사용한 Hall의 연구 등이 대표적이다[4]. 하지만 이러한 연구들은 클러스터의 수를 고정시키고 주로 클러스터링 알고리즘의 목적함수 최소화를 위해 유전자 알고리즘을 이용하였기 때문에 다양한 클러스터 집단에 대한 평가를 동시에 할 수 없는 단점이 있다. 제안하는 방법은 하나의 클러스터 분할을 염색체로 표현하여 다양한 클러스터 분할을 형성하도록 하였고, 이 집단을 유전자 알고리즘을 이용해 진화시켜 최적의 클러스터 분할을 찾아낸다. 각 세대마다 퍼지 c-means 알고리즘을 사용하여 집단내 개체들을 클러스터링하고 각 개체의 적합도 평가에는 퍼지 클러스터 평가 척도인 베이지안 검증방법을 개선한 적응적 α-cut 기반 평가방법을 사용한다. 결정트리(decision tree)의 규칙을 이용하여 각 데이타에 따라 클러스터 결과를 검증하는 기준을 달리하여 적절한 평가가 이루어 지도록 한다. 제안하는 방법을 공개된 유전자 발현 데이타인 SRBCT 데이타와 효모 세포주기 데이타에 적용하여 제안하는 방법의 유용성을 확인하였고, 마지막으로 제안하는 방법이 찾아낸 최적의 클러스터 분할을 분석하였다.

2. 배 경

2.1 DNA 마이크로어레이

마이크로어레이 기술의 등장으로 한번의 실험으로 대량의 유전자 정보를 습득할 수 있게 되었다. 마이크로어레이는 고밀도의 cDNA나 oligonucleotide를 슬라이드 위에 배열해 놓은 것으로 DNA 칩이라고도 한다. 본 논문에서는 두 가지의 cDNA 마이크로어레이 데이타인 SRBCT데이타와 효모 세포주기 데이타가 실험을 위해 사용되었다.

cDNA 마이크로어레이는 실험을 거친 수천 개 이상의 유전자를 일정한 간격으로 배열한 후 유전자들의 발현 패턴에 따른 색 변화로부터 유전자 발현 정보를 얻을 수 있도록 만들어진 바이오칩이다. 어레이 상의 각 셀은 두 개의 다른 환경에서 채집된 유전물질에 녹색의 Cy3와 빨간색의 Cy5라는 각기 다른 형광물질을 합성하여 동일한 양으로 보합한 것이다. 이것을 레이저 형광 스캐너로 읽어 들이면 녹색부터 빨간색에 이르는 발현 정도를 얻을 수 있는데, 식 (1)과 같이 Cy5/Cy3의 비율에 밑이 2인 로그를 취한 값을 그 셀의 발현 정보 값으로 얻게 된다[8,9]. 식 (1)에서 Int는 강도(intensity)를 의미한다.

$$gene\_expression = \log_2 \frac{\text{Int(Cy5)}}{\text{Int(Cy3)}} \quad (1)$$

2.2 퍼지 c-means 알고리즘

퍼지 c-means 알고리즘은 Bezdek에 의해 제안된 알고리즘으로, 가장 널리 이용되는 퍼지 클러스터링 방법이다. 주어진 데이타 집합이 $X = \{x_1, x_2, ..., x_n\}$이고 퍼지 클러스터링의 중심 벡터가 $V = \{v_1, v_2, ..., v_c\}$일 때, 목적함수 J_m은 각 데이타 x_j와 각 클러스터 중심 v_i와의 거리와 클러스터에 대한 소속 정도(degree of membership) 값으로 정의된다.

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

여기서, μ_{ij}는 x_j의 i번째 클러스터에 대한 소속정도를 의미하며, 소속행렬 $U = [\mu_{ij}]$의 원소이다. $d^2(,)$는 유클리드 거리의 제곱이고, 매개변수 m은 각 데이타의 소속 정도에 대한 퍼지 파라미터를 나타내며, 보통 1보다 큰 수를 사용한다[10,11]. m이 1이 되면 퍼지 c-means 알고리즘은 하드 c-means 알고리즘과 동일해진다.

퍼지 c-means 알고리즘의 수행절차는 다음과 같다.

- 단계 1: 클러스터 수 c와 퍼지 파라미터 m을 결정한다.
- 단계 2: 식 (3)의 조건을 만족하도록 μ_{ij}를 초기화한다.

$$\sum_{i=1}^{c}\mu_{ij} = 1,\ 1 \le j \le n \quad (3)$$

• 단계 3: 각 클러스터의 중심 v_i를 계산한다(i=1, 2, ..., c).

$$v_i = \frac{\sum_{j=1}^{n} \mu_{ij}^{m} x_j}{\sum_{j=1}^{n} \mu_{ij}^{m}} \tag{4}$$

• 단계 4: 소속 행렬 U를 계산한다.

$$\mu_{ij} = \frac{\left(\frac{1}{d^2(x_j, v_i)}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{c}\left(\frac{1}{d^2(x_j, v_k)}\right)^{\frac{1}{m-1}}} \tag{5}$$

• 단계 5: 아래의 종료조건이 만족될 때까지 단계 3과 단계 4를 반복한다. 여기에서 $J_m^{(l)}$은 l번째 루프에서의 목적함수 값을 의미한다.

$$|\{J_m^{(l)} - J_m^{(l-1)}\}| \leq \varepsilon \tag{6}$$

3. 제안하는 방법

본 논문에서는 최적의 클러스터 분할을 찾기 위해 그림 1과 같이 진화연산을 이용한 클러스터링 방법을 제안한다. 제안하는 방법은 크게 두 부분으로 나누어, 유전자 알고리즘을 이용한 퍼지 클러스터링 알고리즘을 통하여 최적의 클러스터 분할을 검색하는 부분과, 데이타로부터 필요한 정보를 추출하여 미리 결정트리에 의해 생성된 규칙으로 베이지안 검증에 필요한 최적의 α-cut을 결정하는 적응적 α-cut 기반 평가부분으로 구성된다.

3.1 적응적 α-cut 기반 평가

여기에서는 베이지안 검증방법과 이를 개선하기 위해 사용될 데이타의 특성을 고려한 α-cut을 결정트리의 규칙을 이용해 결정하는 과정을 설명한다.

3.1.1 베이지안 검증방법

베이지안 검증 방법은 확률기반의 검증방법으로, 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.

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

베이즈 이론을 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 식 (8)은 다음의 식 (9)와 같이 표현될 수 있다.

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

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.

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

여기서 $n(D_i)$는 D_i의 개수이며, 일정한 확률보다 큰 멤버쉽값(u_{ij} >α)을 가진 샘플들만을 선택한다. 그 이유

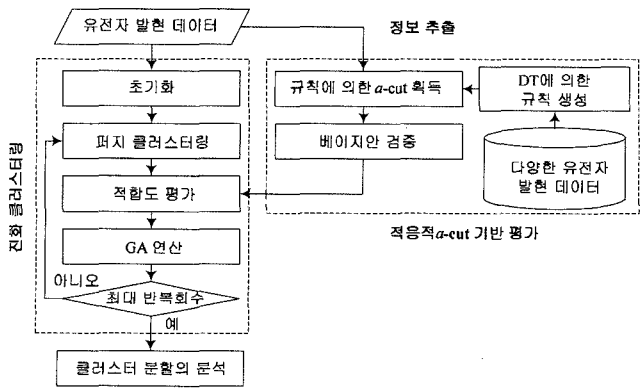


그림 1 제안하는 방법의 개요

는 BS의 계산과정 중에는 곱셈 계산이 있기 때문에 u_{ij}=0인 샘플들의 경우 올바른 값이 나올 수 없게 되고, 또한 퍼지 클러스터링을 하는 궁극적인 이유는 명확하지 않은 개체들의 소속정도를 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 하며 α-cut의 결정은 베이지안 검증방법에서 매우 중요한 역할을 한다. 각 확률은 다음 식과 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. 베이지안 검증방법은 최종적으로 Bayesian score(BS)를 계산하여 클러스터를 검증하는데, 다음과 같은 과정을 거친다. 베이지안 검증방법의 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: 단계2에서 선택한 D_j에 대해 $P(D_j|C_j)$, $P(D_j)$, $P(C_j)$ 값을 계산한다.
- 단계 4: 단계3에서 계산한 값을 이용하여 BS를 계산한다.
- 단계 5: BS의 값을 최대로 하는 클러스터 분할을 최적의 분할로 평가한다.

3.1.2 적응적 α-cut기반 평가

앞에서 설명했듯이 데이타마다 샘플의 분포 등 특성이 다르기 때문에 그에 맞추어 적응적으로 α-cut을 설정해 주는 것이 필요하다. 본 논문에서는 데이타의 도메인을 고려하여 α-cut을 결정하기 위해 결정트리의 규칙에 의해 각 데이타에 맞는 α-cut을 자동으로 구한다.

먼저 실험에 사용될 N개의 유전자 발현 데이타를 퍼지 c-means 알고리즘을 통해 클러스터링하고, 그 결과를 베이지안 검증방법으로 평가한 후, 각 데이타별로 적절한 α-cut을 결정해 결정트리의 학습데이타을 구성한다. 규칙생성과정에서는 결정트리를 학습하고, 이를 바탕으로 규칙을 생성한다.

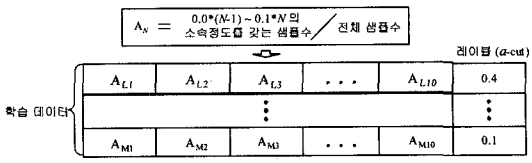


그림 2 결정트리의 학습 데이타 생성 과정

결정트리 학습 데이타의 속성은 그림 2와 같이 퍼지 클러스터링을 거쳐 나온 각 데이타의 소속행렬을 이용해 생성한다. 소속행렬의 소속정도를 0.0부터 1.0까지 0.1단위로 증가시키며 총 10구간으로 나누고, 각 구간의 소속정도를 가진 샘플의 빈도를 계산해 데이타의 총 샘플수로 나눈 값을 속성으로 정의해 사용한다. 이렇게 정한 각 속성을 A_1~A_{10}으로 정의한다. 이렇게 구성된 학습데이타를 가지고 데이타가 입력되면 결정트리에 의해 만들어진 규칙에 의해 α-cut을 결정한다. 이 과정은 실험결과 4.2.1에서 설명된다. 이후 이렇게 결정된 α-cut을 이용해 베이지안 검증방법의 BS를 구하고 이를 각 개체의 적합도 평가에 이용한다.

3.2 진화 연산을 이용한 퍼지 클러스터링

여기에서는 최적의 클러스터 분할을 찾기 위한 방법인 유전자 알고리즘을 이용한 퍼지 클러스터링 알고리즘에 대해 설명하고, 유전자 알고리즘의 단계별로 제안하는 방법에 대해 기술한다.

3.2.1 개체의 표현

본 논문에서는 일반적인 이진표현을 사용하지 않고, 실수표현법을 사용한다. 이를 통해 하나의 개체가 여러 클러스터 중심에 대한 정보를 포함하는 하나의 클러스터 분할을 표현할 수 있게 된다. 하나의 클러스터 분할은 K개의 클러스터를 포함하고 각 중심값이 N차원으로 표현된다고 할 때, 개체는 NxK 공간으로 표현된다.

본 논문은 다양한 크기의 클러스터 수를 가진 집단을 평가하는 것이 목적이기 때문에 가변길이로 개체를 표현한다. 그림 3과 같이 크기가 다른 개체들이 존재하여, 각 개체가 서로 다른 개수와 값을 갖는 클러스터 중심으로 표현된다.

3.2.2 집단 초기화와 적합도 평가

하나의 집단은 집단내의 클러스터 수만큼의 샘플을 임의로 선택하여 그 샘플들을 클러스터 중심으로 초기화하는데, 개체 수만큼 위의 과정을 반복하여 전체 집단을 초기화한다. 본 논문은 가변길이 개체를 사용하는데, 데이타 샘플수의 제곱근을 최대 길이로 제한하였다[13]. 최저 길이는 2로 설정하여 최소 두 개 이상의 클러스터

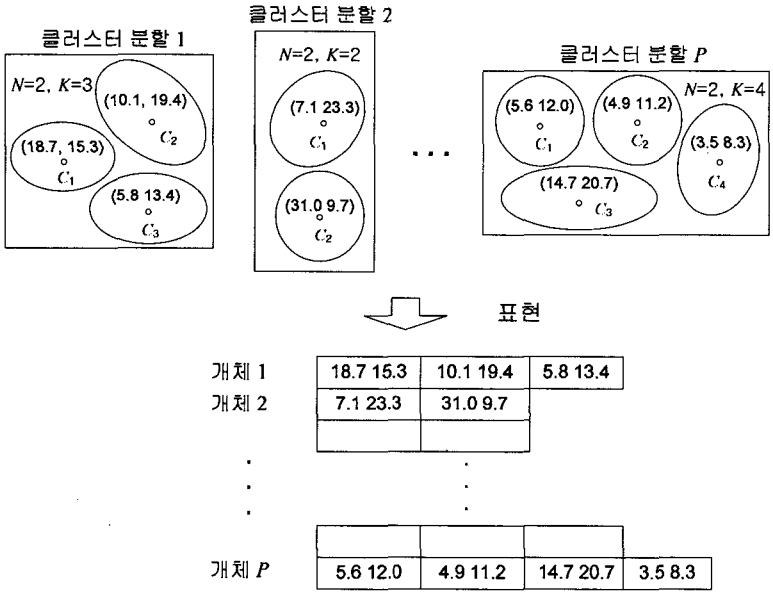


그림 3 가변길이 개체의 표현

를 가진 개체로 집단을 설정하였다.

적합도 평가를 위해서는 적응적 α-cut기반 평가를 사용하였는데, 먼저 개체내의 클러스터 중심값을 기반으로 퍼지 c-means 알고리즘을 이용해 데이타의 모든 샘플을 클러스터링한다. 그리고 각 샘플과 클러스터 중심간의 거리를 계산하여 식 (13)과 같이 소속행렬값을 구하고, 식 (14)을 이용해 클러스터 중심을 갱신한다. 염색체내의 클러스터 중심정보는 이렇게 새로 계산된 클러스터 중심으로 바뀐다.

$$u_{ij} = \left(\frac{1}{d^2(x_j, v_i)}\right)^{\frac{1}{m-1}} \Bigg/ \sum_{k=1}^{c}\left(\frac{1}{d^2(x_j, v_i)}\right)^{\frac{1}{m-1}} \tag{13}$$

$$v_i = \sum_{j=1}^{n} u_{ij}^m x_j \Bigg/ \sum_{j=1}^{n} u_{ij}^m \tag{14}$$

그 이후에는 바뀐 개체들을 평가하며 이 과정에서 데이타에 대한 사전 지식을 반영한 적응적 α-cut기반 평가방법을 사용한다. 개체들이 서로 다른 클러스터 분할을 표현하고 있기 때문에 다양한 평가값이 계산되며 이 결과를 바탕으로 선택이 이루어진다. 적합도에 비례한 확률을 주어 적합도가 높은 개체의 생존확률을 높이는 룰렛휠(Roulet wheel)방법이 선택을 위해 사용되었다[14].

3.2.3 교차와 돌연변이 연산

본 논문은 가변길이 개체를 사용하기 때문에, 일반적인 교차나 돌연변이 연산을 적용할 수 없다. 교차 연산은 개체의 크기를 고려하여 교차점을 선택한 후 앞부분의 클러스터 중심값을 바꾸는 방식으로 이루어진다. 염색체의 길이가 l인 경우에, [1, l-1] 사이의 범위에서 교차점이 선택되는데, 그림 4는 교차 연산의 예를 보여준다.

실수표현법을 사용하기 때문에 돌연변이 연산은 다음과 같은 과정을 통해 이루어진다. δ가 [0, 1] 사이의 값을 가진 균일한 분포의 변수라고 하고, ν가 돌연변이가 일어나는 지점의 값이라고 할 때, 돌연변이에 의해 바뀌

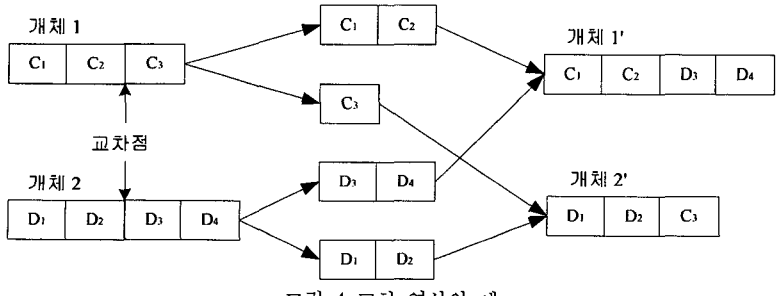


그림 4 교차 연산의 예

는 ν가 식 (15)와 (16)에 의해 계산된다[5]. ν의 값이 0이 아닐 때는 식 (15)를 사용하고, ν가 0일 때는 식 (16)을 사용한다. 이 때 '+'와 '-' 부호가 사용될 확률은 동일하다.

$$\nu \pm 2\times\delta\times\nu,\ \nu \neq 0 \quad (15)$$

$$\nu \pm 2\times\delta,\ \nu = 0 \quad (16)$$

4. 실험 결과

4.1 실험 환경

4.1.1 실험 데이타

실험을 위해서 SRBCT와 효모 세포주기 데이타가 사용되었으며, 결정트리를 이용한 적응적 α-cut 결정 실험을 위해 추가로 림프종(lymphoma)[15], 백혈병(leukemia)[16], 혈청(serum)[17] 데이타가 사용되었다.

- SRBCT(Small Round Blue Cell Tumors) 데이타: SRBCT 데이타는 총 63개의 샘플로 구성되며 6567개의 유전자를 갖는다. NB(neuroblastoma), RMS (rhabdomhosarcoma), NHL(non-Hodgkin lymphoma), EWS (Ewing family of tumors)의 4 클래스로 나뉘며 네 가지 모두 암의 일종이다. 본 논문에서는 Kahn의 연구[18]를 참고해 클러스터링 과정에 중요하다고 알려진 96개의 유전자를 사용하여 63개의 샘플을 클러스터링하였다.
- 효모 세포주기(Yeast cell-cycle) 데이타: 효모 세포주기 유전자 발현 데이타는 두 번의 세포주기동안 약 6000개 유전자들의 발현정도를 측정한 데이타이다[19]. 10분 간격으로 두번의 효모 세포주기에 해당되는 160분에 걸쳐 유전자들의 발현정도를 17개의 시점에서 측정했다. 이 데이타는 생물학적 기능에 따라 분류된 유전자들이 주기별로 지정되어 있어서 분석을 위해 자주 사용된다. 본 논문에서는 실험과정에서 의미있는 발현 변화를 보이는 421개의 유전자를 선택하여 클러스터링 하였다[19].

4.1.2 파라미터 설정

모든 실험은 100세대까지 진화시켰고, SRBCT 데이타는 집단의 크기를 100으로, 효모 세포주기 데이타는 200으로 설정하였다. SRBCT 데이타의 집단의 크기를 더 작게 설정한 것은 샘플수가 효모 세포주기 데이타보다 적기 때문이다. 최대 클러스터 수는 SRBCT의 경우 8로, 효모 세포주기 데이타는 20으로 설정하였다[16]. 교차율은 0.8, 돌연변이율은 0.01이 사용되었고, 퍼지 c-means 알고리즘의 퍼지 파라미터는 모든 실험에 동일하게 1.2로 설정되었다.

4.2 실험 결과

4.2.1 적응적 α-cut 결정

앞에서 설명한 5가지의 유전자 발현 데이타를 사용하여 결정트리의 학습 데이타를 생성하였는데, 생성된 규칙은 그림 5와 같다. 규칙에 의해 학습 데이타의 첫 속성(A_1)이 0.9571 이상이면 α-cut이 0.8로 결정되며, 이와 동시에 세 번째 속성(A_3)도 0.0 이하이면 α-cut이 0.2로 결정된다. 마지막으로 열 번째 속성(A_{10})은 α-cut 0.1과 0.4를 결정하는 기준이 된다.

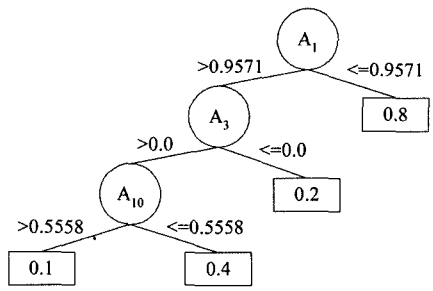


그림 5 결정트리에 의해 생성된 규칙

결정트리에 의해 결정된 SRBCT와 효모 세포주기 데이타의 α-cut은 0.2와 0.4이다. 두 데이타 모두 A_1의 값은 0.9571보다 큰 값을 갖지만, A_3에서 다른 방향으로 갈라지게 된다. 두 데이타로부터 추출한 결정트리의 학습 데이타는 표 1과 같다. SRBCT 데이타와 효모 세포주기 데이타를 비교해 보면, SRBCT 데이타는 A_2~A_9 속성이 효모 세포주기 데이타보다 작은 수치를 보이고, A_{10} 속성은 더 큰 수치를 보인다. 이는 SRBCT 데이타의 샘플들이 효모 세포주기 데이타의 샘플들보다 클러스터 간의 경계가 명확하여 0.9 이상의 큰 소속정도를 갖는 샘플들이 많기 때문이다. 표 1은 이 두 데이타의 특성이 다르다는 것을 명확히 보여주며, 따라서 이 점을 고려하여 검증을 해야 클러스터 분할을 올바르게 평가할 수 있음을 뒷받침한다.

4.2.2 최적의 클러스터 분할 탐색

그림 6은 세대에 따라 진화하는 SRBCT 데이타의 평균 적합도 추이를 보여준다. 처음에 빠르게 진화하다가

표 1 결정트리의 학습 데이타

데이타	학습 데이타의 속성									
	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
효모 세포주기	1.000	0.231	0.114	0.086	0.071	0.064	0.059	0.069	0.162	0.546
SRBCT	1.000	0.016	0.000	0.016	0.000	0.016	0.000	0.000	0.048	0.937

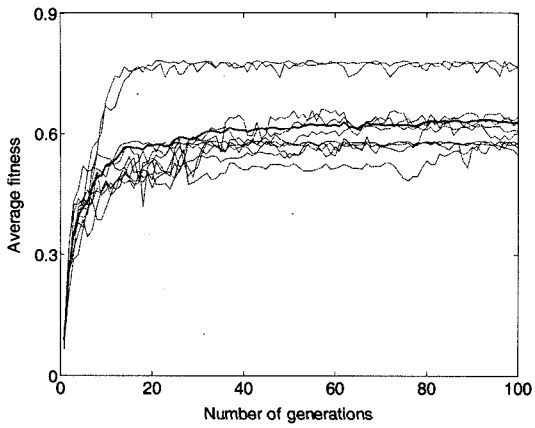


그림 6 SRBCT 데이타의 평균 적합도 추이 변화 (P=100)

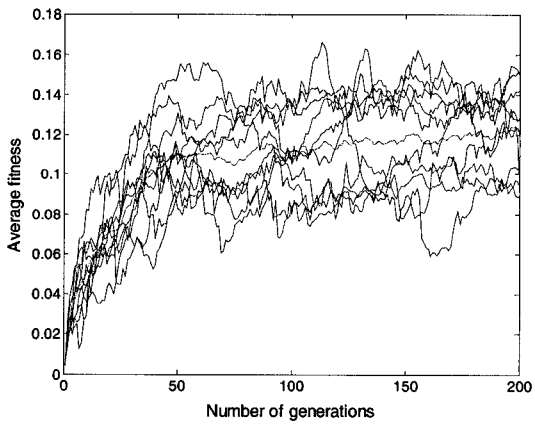


그림 7 효모 세포 주기 데이타의 평균 적합도 추이 (P=200)

20세대에 가까워지면 0.6 정도의 값으로 평균 적합도가 수렴함을 알 수 있다. 그림 7은 효모 세포주기 데이타의 평균 적합도 추이를 보여준다. SRBCT 데이타의 결과보다 느리게 수렴하지만 80세대 정도까지 평균 적합도가 꾸준히 증가하다가 그 이후 0.12 정도의 값에 수렴하는 결과를 보이고 있다. 두 실험 모두 10회 반복실험을 하였다. 효모 세포주기 데이타의 평균 적합도 변화가 SRBCT 데이타의 결과에 비해 심한 변화를 보이는데 이는 두 데이타가 가진 다른 특성이 반영된 결과이다. 이 차이는 표 1을 보면 알 수 있는데 SRBCT 데이타는 소속 정도가 0.9보다 크거나 0.1보다 작은 데이타가 많은 반면 효모 세포주기 데이타는 다양한 분포를 보였다.

4.2.3 단일 퍼지 c-means 알고리즘과의 성능 비교

여기에서는 단일 FCM과 제안하는 방법인 GA+FCM을 *BS*와 목적함수값을 이용해 비교한다. 표 2는 SRBCT 데이타에 대해서 두 방법의 결과를 비교한 표이다. 클러스터링 결과가 좋을 경우 *BS*값은 커지고 목적함수값은 작아진다. 큰 차이는 아니지만 제안하는 방법의 결과가 단일 FCM의 결과보다 좋은 값을 보인다.

표 2 FCM과의 성능 비교 (SRBCT 데이타)

	단일 FCM		GA+FCM	
실험	*BS*	목적함수값	*BS*	목적함수값
1	0.58028	156.9920	0.58042	156.9918
2	0.58034	156.9925	0.58036	156.9920
3	0.58031	156.9921	0.58041	156.9919
4	0.58029	156.9922	0.58036	156.9920
5	0.58041	156.9926	0.58036	156.9920
6	0.58034	156.9921	0.58041	156.9919
7	0.58031	156.9920	0.58042	156.9918
8	0.58028	156.9922	0.58042	156.9918
9	0.58033	156.9922	0.58041	156.9919
10	0.58036	156.9920	0.58042	156.9918
평균	0.58033	156.9922	0.58040	156.9919

표 3 FCM과의 성능 비교 (효모 세포주기 데이타)

	단일 FCM		GA+FCM	
실험	*BS*	목적함수값	*BS*	목적함수값
1	0.03354	164.472	0.13256	166.883
2	0.00875	163.670	0.11246	161.542
3	0.03238	165.057	0.12661	162.911
4	0.03825	162.653	0.08058	162.073
5	0.02165	163.758	0.10667	162.798
6	0.04096	164.086	0.09778	162.312
7	0.02806	163.052	0.11873	162.042
8	0.04473	164.877	0.13659	162.773
9	0.02478	162.452	0.12898	162.905
10	0.04645	169.216	0.11246	161.542
평균	0.03195	164.329	0.11534	162.778

표 3은 역시 10번 반복한 효모 세포주기 데이타의 실험결과이다. 효모 세포주기 데이타의 경우는 큰 차이로 제안하는 방법이 더 좋은 결과를 보였다.

제안하는 방법과 단일 FCM을 비교한 결과, 제안하는 방법이 최적해에 더 근접한 결과를 보이는 것을 확인하였다.

4.3 결과 분석

4.3.1 SRBCT 데이타의 실험 결과 분석

SRBCT 데이타에 제안하는 방법을 적용하여 얻은 실험 결과를 Khan의 연구[18]와 비교하였다. 그림 8은 제안하는 방법으로 찾은 4개의 클러스터와 실제 샘플이 속한 클래스를 비교한 것이다. 63개 샘플이 모두 자신이 속한 클래스에 올바로 소속되었다. 클래스 EWS와 RMS의 가운데 위치한 샘플 EWS-T13은 EWS 클래스와 RMS 클래스에 동시에 비슷한 소속정도로 속한다는 실험 결과를 얻었다.

이처럼 SRBCT의 실험 결과가 좋은 이유는 원래의

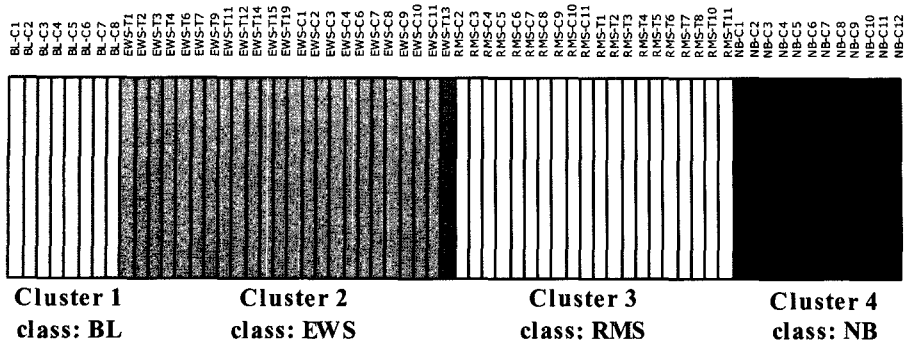


그림 8 SRBCT 데이타의 클러스터링 결과 (96개 유전자)

데이타에서 추출한 96개의 유전자가 이미 클래스별로 분포가 다른 발현정도를 보이기 때문이다. 또한 Khan의 연구[18]에서와 같이 1차로 최소한의 불필요한 정보만을 제거한 2308개의 전체 유전자를 이용한 실험결과, 표 4와 같이 8개의 클러스터가 발견되었다.

표 5는 제안하는 방법이 찾아낸 샘플로 소속정도가 0.3 이상이면서 다수의 클러스터에 동시에 속하는 퍼지 샘플들이다. 총 4개의 샘플을 찾았으며, 각각의 소속정도와 클러스터 번호가 표에 표시되어 있다. 마지막에 표시된 샘플 EWS-T13은 96개의 유의한 유전자를 사용했을 때 발견된 퍼지 샘플이다. Cluster 3과 Cluster 4에 동시에 속하는 것으로 나타났는데 이 두 클러스터가 EWS와 RMS에 속하는 클러스터라고 유추해 볼 수 있다.

표 5 SRBCT 데이타의 퍼지 샘플 목록

퍼지 샘플	첫 번째 클러스터	두 번째 클러스터
EWS-T19	0.549942(3)	0.379079(4)
NB-C5	0.564467(6)	0.313732(7)
RMS-C6	0.349232(7)	0.333511(6)
RMS-C8	0.456904(5)	0.339977(1)
EWS-T13	0.718476(3)	0.210988(4)

표 4의 결과를 토대로 각 클러스터간의 관계를 분석하여 그 결과를 그림 9에 정리하였다. Cluster 0과 Cluster 3은 EWS클래스에 완전히 소속되고 Cluster 1, Cluster 4, Cluster 5, Cluster 7은 소속 샘플의 일부가 EWS클래스에 속한다. 절반 정도의 샘플이 EWS와 RMS클래스에 동시에 속하는 Cluster 5는 EWS와 RMS클래스의 중간 성질을 가진 클러스터로 분석할 수 있다. Cluster 6은 NB클래스에, Cluster 2는 BL클래스에 속하며 Cluster 1은 EWS, BL, NB 세 개의 클래스에 동시에 속한다.

4.3.2 효모 세포주기 데이타의 실험 결과 분석

효모 세포주기 데이타의 실험 결과는 Cho의 연구[19]와 비교하여 알려진 유전자들의 기능을 참고해 분석을 시도하였다. 0.3이상의 소속정도를 갖고 둘 이상의 클러스터에 소속되는 퍼지 유전자의 분석에 초점을 맞추었다.

그림 10은 실험결과 얻어진 퍼지 유전자들의 설명과 발현정도를 나타낸다. 총 25개의 유전자들을 소속된 클러스터 번호에 따라 크게 네 개의 그룹으로 나누어 분석하였다. YBL032w, YHR031C, YCL063w는 Cluster 4, Cluster 7, Cluster 11에 속하며 이 세 개의 클러스

표 4 SRBCT 데이타의 클러스터링 결과(2308개 유전자)

클러스터 번호	클러스터에 속하는 샘플들
Cluster 0	EWS-C6 EWS-C8 EWS-C9 EWS-C10 EWS-C11
Cluster 1	EWS-C1 EWS-C2 EWS-C3 BL-C1 BL-C2 BL-C3 BL-C4 NB-C1
Cluster 2	BL-C5 BL-C6 BL-C7 BL-C8
Cluster 3	EWS-T6 EWS-T7 EWS-T9 EWS-T11 EWS-T12 EWS-T14 EWS-T15 **EWS-T19**
Cluster 4	EWS-T13 RMS-C4 RMS-T5 RMS-T6 RMS-T7 **RMS-T8** RMS-T10 RMS-T11
Cluster 5	EWS-C4 EWS-T1 EWS-T2 EWS-T3 EWS-T4 RMS-C8 RMS-C11 RMS-T1 RMS-T2 RMS-T3 RMS-T4
Cluster 6	NB-C2 NB-C3 NB-C4 **NB-C5** NB-C6 NB-C7 NB-C8 NB-C9 NB-C10 NB-C11 NB-C12
Cluster 7	EWS-C7 RMS-C2 RMS-C3 RMS-C5 **RMS-C6** RMS-C7 RMS-C9 RMS-C10

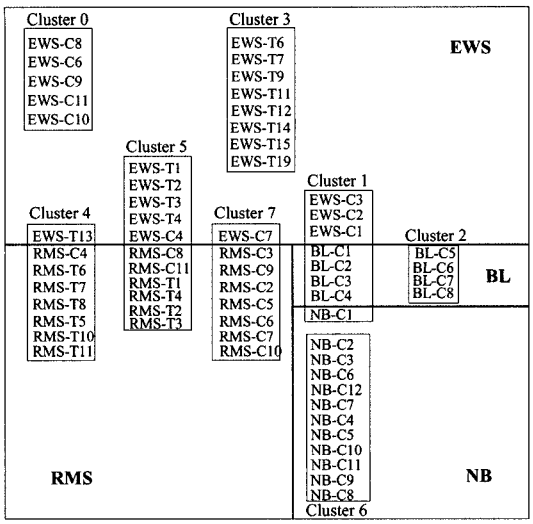


그림 9 SRBCT 데이타의 클러스터별 분류 결과(2308개 유전자)

터를 하나로 묶었다. 마찬가지로 Cluster 5, Cluster 12, Cluster 13, Cluster 15를 하나의 그룹으로 묶었다. 이 경우에는 Cluster 5를 중심으로 Cluster 12, Cluster 13, Cluster 15가 관계를 갖는다. 나머지 두 그룹은 Cluster 0, Cluster 1, Cluster 2를 하나, Cluster 10, Cluster 12를 다른 하나의 그룹으로 묶었다. 좌측의 발현정도를 나타낸 그림을 보면 각 그룹별로 다른 패턴을 보임을 확인할 수 있다.

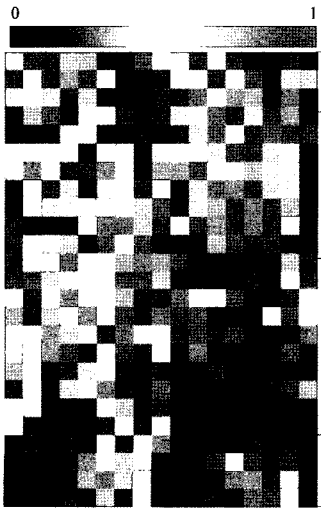
그림 11은 네 개의 그룹에서 각각 하나의 클러스터를 선택하여, 각각의 발현정도의 변화를 그래프로 나타낸 것이다. Cluster 7은 26개의 유전자로 구성되는데 효모 세포주기 중 G_2기에 가장 높은 발현 정도를 보이는 유전자들로 구성된 것으로 미루어 G_2기와 관련된 그룹임을 알 수 있다. Cluster 5는 S기에서 높은 발현정도를 보여 S기와 관련된 그룹임을 알 수 있으며, 마지막 그룹의 Cluster 12는 G_1기와 S기 사이에서 최대 발현 정도를 보인다. Cluster 12는 그림 9에서 Cluster 5가 속한 두 번째 그룹에도 속하는 것으로 보아 실제로 두 클러스터에 모두 연관되어 있다고 할 수 있다.

5. 결론 및 향후 연구

본 논문은 최적의 클러스터 분할을 효율적으로 탐색하기 위해 진화 연산을 이용한 클러스터링 방법을 제안하였고, 클러스터링 결과의 적절한 평가를 위해 베이지안 검증 방법의 단점을 개선하여 데이타에 따라 평가기준을 다르게 한 적응적 α-cut기반 평가방법을 제안하였다. 제안하는 방법을 SRBCT 데이타와 효모 세포주기 데이타에 적용해 실제로 진화가 잘 이루어지는 것을 확인하였으며, 제안하는 방법을 통해 찾아낸 최적의 클러스터 분할을 기존 연구와 비교하여 분석하였다. 추후 연구로 좀 더 다양한 데이타에 제안하는 방법을 적용할 것이다.

참 고 문 헌

[1] U. Alon, *et al.*, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide



Gene description	Clusters
weak similarity to hnRNP complex protein homolog YBR233w	4, 7
similarity to Pif1p	4, 7
weak similarity to yeast translation regulator Gcd6p	7, 11
hypothetical protein	5, 15
hypothetical protein	5, 15
suppressor of rvs161 and rvs167 mutations	5, 13
involved in the HOG1 high-osmolarity signal transduction	5, 13
hypothetical protein	5, 13
forms schmoos poorly	5, 13
Rap1p interacting factor. telomere silencing	5, 12
DNA replication factor A, 13 kDa subunit	5, 12
similarity to rat regucalcin	0, 1
protein with internal repeats	0, 1
involved in regulating membrane traffic	0, 1
ubiquinone biosynthesis, methyltransferase	0, 1
similarity to bovine Graves disease carrier protein	0, 2
hypothetical protein	1, 2
cyclin like protein interacting with Pho85p	1, 2
acyl CoA synthase	1, 2
hypothetical protein	1, 2
hypothetical protein	1, 2
cyclin, G1/S-specific	10, 12
phosporylates Cdc28, SPB separation, nuclear division	10, 12
for chromosome transmission in mitosis and normal telomere	10, 12
required for normal transcription at a number of loci	10, 12

그림 10 제안하는 방법이 찾은 효모 세포주기 데이타의 퍼지 유전자들의 정보
(발현정도, 유전자 설명, 소속 클러스터 번호)

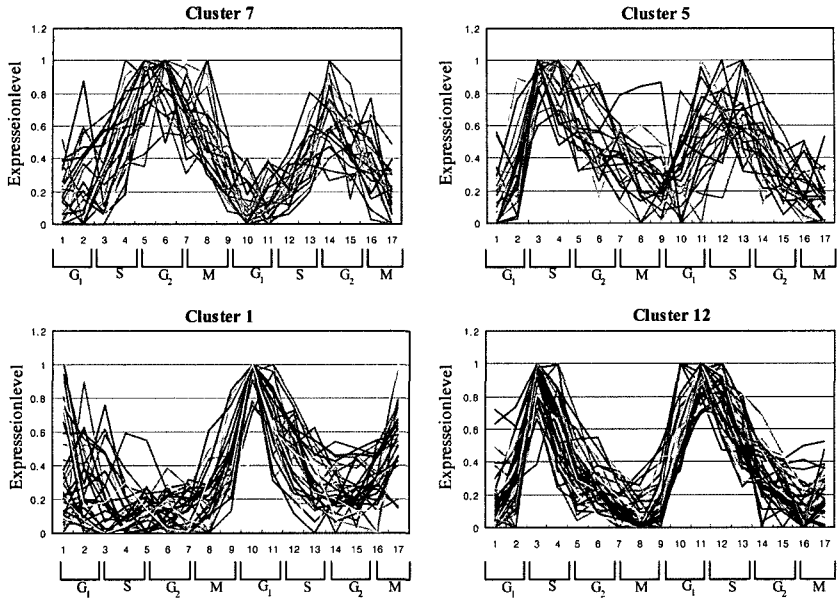


그림 11 클러스터 별 유전자들의 발현 정도(효모 세포주기 데이타)

arrays," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 6745-6750, June 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] N. Bolshakova and F. Azuaje, "Cluster validation techniques for genome expression data," *SIGPRO*, vol. 21, no. 82, pp. 1-9, 2002.

[4] L. O. Hall, *et al.*, "Clustering with a genetically optimized approach," *IEEE Trans. on Evolutionary Computation*, vol. 3, no. 2, pp. 103-112, 1999.

[5] U. Maulik and S. Bandyopadhyay, "Genetic algorithm-based clustering technique," *Pattern Recognition*, vol. 33, pp. 1455-1465, 2000.

[6] L. Chamber, *Practical Handbook of Genetic Algorithm*, CRC Press, 1995.

[7] J. N. Bhuyan, *et al.*, "Genetic algorithm for clustering with an ordered representation," *in Proc. 4th Int. Conf. Genetic Algorithms*, pp. 408-415, 1991.

[8] M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA*, 95, pp. 14863-14868, 1998.

[9] M. E. Futschik, A. Reeve and N. Kasabov, "Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue," *Artificial Intelligence in Medicine*, 28, pp. 165-189, 2003.

[10] R. E. Hammah and J. H. Curran, "Validity measures for the fuzzy cluster analysis of orientations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, 2000.

[11] F. Hoppner, *et al.*, *Fuzzy Cluster Analysis*, Wiley, pp. 43-39, 1999.

[12] S.-H. Yoo, H.-H. Won and S.-B. Cho, "Analysis of Saccharomyces cell cycle expression data using Bayesian validation of fuzzy clustering," *Journal of Korea Information Science Society*, vol. 31, no. 12, pp. 1591-1601, 2004.

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

[14] K. Krishna and M. N. Murty, "Genetic k-means algorithm," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 20, no. 3, pp. 433-439, 1999.

[15] A. A. Alizadeh, *et al.*, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503-511, February 2000.

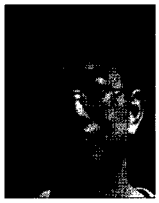
[16] T. R. Golub, *et al.*, "Molecular classification of cancer class discovery and class prediction by gene-expression monitoring," *Science*, vol. 286, no. 15, pp. 531-537, October 1999.

[17] V. R. Iyer, *et al.*, "The transcriptional program in the response of human fibroblast to serum," *Science*, vol. 283, pp. 83-87, 1999.

[18] J. Khan, *et al.*, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature*, vol. 7, no. 6, pp. 673-679, June 2001.

[19] R. J. Cho, *et al.*, "A genome-wide transcriptional

analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, pp. 65-73, 1998.



박 한 샘

2004년 2월 연세대학교 컴퓨터과학과(학사). 2006년 2월 연세대학교 컴퓨터과학과(석사). 2006년 3월~현재 연세대학교 컴퓨터과학과 박사과정. 관심분야는 패턴인식, 생물정보학, 지능형로봇, HCI

조 성 배

정보과학회논문지 : 소프트웨어 및 응용
제 33 권 제 3 호 참조