유전자 발현 데이터의 퍼지 클러스터 평가를 위한 결정트리 기반의 베이지안 검증방법

유시호°, 조성배
연세대학교 컴퓨터과학과
bonanza@sclab.yonsei.ac.kr°, sbcho@cs.yonsei.ac.kr

A Bayesian Validation Method based on Decision Tree for Evaluating Fuzzy Clusters of Gene Expression Data

Si-Ho Yoo°, Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

퍼지 클러스터링 방법은 일반적인 클러스터링 방법과는 달리 하나의 샘플이 다수의 집단에 속할 수 있으며 그 속하는 정도를 표현하여 보다 유연한 클러스터 분할의 분석을 가능하게 한다. 유전자 발현 데이터는 노이즈가 많고 공통된 기능을 가진 유전자들의 집단이 존재하기 때문에 퍼지 클러스터링을 사용하면 더욱 효율적으로 분석할 수 있다. 이러한 퍼지 클러스터링 방법에 있어서 중요한 것은 얼마나 분할이 정확하게 이루어졌으며, 실제 데이터가 가지고 있는 분할과 결과가 얼마나 유사한가이다. 본 논문에서는 효과적인 유전자 클러스터의 평가를 위하여 베이지안 검증 방법을 제시하고, 결정트리로 생성된 규칙에 의하여 각 데이터의 특성에 따라 유연하게 검증하는 방법을 제안한다. 다양한 유전자 발현 데이터를 퍼지 c-means 알고리즘을 이용하여 클러스터링하고 제안하는 방법으로 검증한 결과, 그 유용성을 확인할 수 있었다.

1. 서론

다양한 클러스터링 방법들이 수 천개의 유전자 정보를 가진 유전자 발현 데이터의 분석에 사용되고 있다. 유전자들은 여러 가지 기능을 가지고 있으며 공통으로 중복된 기능을 가진 경우도 많기 때문에 일반적인 하드 클러스터링 방법보다는 퍼지 클러스터링 방법이 유전자 분석에 더 적합하다[1]. 하지만 이러한 퍼지 클러스터링의 결과는 검증이 어렵기 때문에 이를 그대로 분석에 적용하기에는 한계가 있다. 따라서 클러스터 결과를 체계적으로 검증하는 절차는 유전자 분석에 있어 매우 중요하다.

본 논문에서는 효과적인 유전자 클러스터링의 검증을 위하여 결정트리(Decision Tree:DT)[2]를 이용한 적응적인 베이지안 검증 방법을 제안한다. 기존의 퍼지 클러스터 검증 방법인 베이지안 검증은 주어진 데이터에 대한 클러스터 분할의 사후확률을 계산하여 최적의 클러스터 분할을 구하는 방법인데, 데이터의 특성을 무시하고 동일한 기준으로 데이터를 분할하였다. 유전자 발현 데이터의 경우, 실험환경과 데이터의 특성에 따라 다양한 패턴을 보이기 때문에 동일한 방식으로 분석하는 것보다는 데이터마다 각각 자신에 맞는 최적의 검증 조건하에서 분석하는 것이 더 효율적이다.

제안하는 방법은 각 데이터에서 필요한 정보를 추출하여 DT를 훈련시킨 후, 그 결과로 얻어진 규칙에 의해 데이터에 따른 최적의 검증 조건을 구하여 클러스터 분할을 평가하기 때문에 더 정확한 평가가 가능하다. 제안한 방법의 유용성을 검증하기 위해 널리 사용되고 있는 5가지 유전자 발현 데이터(Yeast cell-cycle, Serum, Leukemia, Lymphoma, SRBCT)에 대하여 실험한다.

2. 관련 연구

표 1과 같이 다양한 클러스터링 알고리즘이 유전자 발현 데이터 분석에 사용되었다. 하드 클러스터링 알고리즘인 SOM과 k-means 알고리즘을 사용하여 Leukemia와 Lymphoma 데이터를 분석한 Bolshakova와 Azuaje의 연구가 있으며[3], Yeast cell-cycle 데이터를 k-means 알고리즘과 single-linkage 알고리즘으로 분석한 Yeung의 연구도 있다[4]. 또한 퍼지 클러스터링 알고리즘인 퍼지 c-means 알고리즘을 이용하여 Serum과 Yeast cell-cycle 데이터를 분석한 Dembele의 연구도 있다[1]. 이러한 연구들에 사용된 검증방법들은 모두 데이터의 특성을 고려하지 않고 단지 클러스터 중심과 샘플들의 거리에 기반한 평가를 하였기 때문에 실제 형성된 클러스터 분할을 잘 반영하지 못하는 한계를 가진다.

표 1. 유전자 발현 데이터를 분석한 연구들

저자	알고리즘	검증방법	데이터
Yeung et al. (2001)	K-means Single-linkage	FOM	Yeast cell-cycle
Bolshakova and Azuaje(2002)	SOM K-means	Dunn's based-index	Leukemia Lymphoma
Dembele and Kastner(2003)	Fuzzy-c-means	Silhouette-index	Serum Yeast cell-cycle Human cancer

3. 방법

3.1 베이지안 검증 방법

베이지안 검증 방법은 데이터가 주어졌을 때, 해당 데이터에 대한 클러스터 분할의 사후확률을 구하여 클러스터 결과를 검증하는 방법이다. 식(1)과 같이 주어진 데이터에 대해 각 클러스터의 사후확률이 최대가 되는 것을 최적의 클러스터 분할로 평가한다.

$$\max P(Cluster \mid Dataset) \quad (1)$$

Bayes Theorem을 적용하면 다음과 같이 사전확률을 이용하여 사후확률값을 구할 수 있다[5].

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

이러한 과정을 이용하여 그림 1과 같이 모든 클러스터에 대한 $P(Cluster|Dataset)$들의 합을 구하여 이를 베이지안 스코어(BS)라고 정의한다. 이 베이지안 스코어는 그 값이 클수록 각 클러스터의 사후확률이 커지므로 좋은 클러스터 분할을 나타낸다고 볼 수 있다.

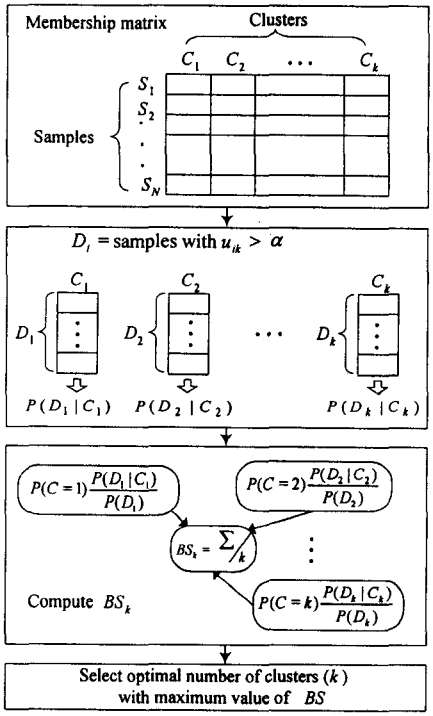

그림 1. 베이지안 검증 방법

그림 1을 보면, 먼저 퍼지 c-means 알고리즘의 결과인 소속행렬(membership matrix)에서 일정한 임계값보다 큰($u_{ik}>\alpha$) 소속행렬값(membership values)을 가진 샘플들만을 클러스터별로 선택한다. u_{ik}는 퍼지 c-means 알고리즘의 결과로 i번째 샘플이 k번째 클러스터에 속하는 정도를 나타내며 0과 1사이의 값을 가진다. 1에 가까운 값을 가질수록 해당 클러스터에 속하는 정도가 크다. 샘플들을 선택한 후, 각 클러스터별로 $P(D_k|C_k)$를 계산하여 최종적인 평가값(BS)을 계산한다.

3.2 결정트리 기반의 베이지안 검증방법

베이지안 검증 방법에서의 α값의 설정은 평가값(BS) 자체에 영향을 준다. 다음은 퍼지 이론에 기초한 α-cut의 정의이다[6].

$$A_\alpha = \{x \in X \mid u(x) \geq \alpha\}, \quad 0 \leq \alpha \leq 1 \quad (3)$$

A_α는 α보다 큰 소속정도를 가진 원소들의 집합이며, x는 각 원소들을 의미한다. α는 0과 1사이의 범위를 가지는데, 어떤 값을 가지느냐에 따라 A_α는 다양한 집합을 형성한다. 보통 소속함수가 선형이면 α=0 또는α=1과 같이 단순한 α의 설정만으로도 충분하지만, 선형이 아닌 경우는 다양한 α의 설정이 필요하다[6]. 그림 1의 D_k는 A_α와 같이 α의 값에 따라 매우 다양한 집합을 형성하기 때문에 데이터에 따른 적절한 α-cut을 설정하는 것은 매우 중요하다.

기존의 베이지안 검증 방법은 모든 데이터에 대해 동일한 α-cut을 설정하고 클러스터 분할을 평가하였는데 데이터마다 샘플들의 분포가 다르기 때문에 정확한 검증이 불가능하였다. 본 논문에서는 이를 해결하기 위하여 데이터에 적응적인 α-cut 결정방법으로 결정트리 기반의 베이지안 검증 방법을 제안한다. 제안하는 방법은 그림 2와 같이 퍼지 클러스터링 과정과 규칙생성 과정으로 나뉘어진다.

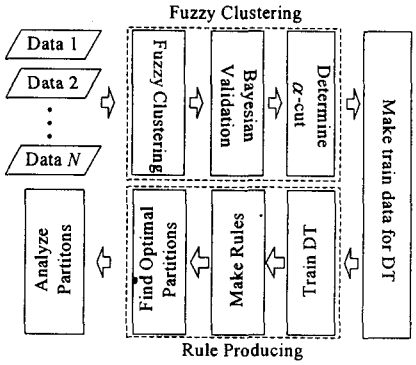

그림 2. 제안하는 방법

퍼지 클러스터링 과정에서는 먼저 클래스 수가 알려진 N개의 유전자 발현 데이터를 퍼지 c-means 알고리즘을 사용하여 클러스터링한다. 그 절차는 다음과 같다.

1) 클러스터 수(c)와 퍼지 계수(m)값을 설정한다.
2) 다음의 조건을 만족하도록 u_{ij}를 초기화 한다.

$$\sum_{i=1}^{c} u_{ij} = 1, \; 1 \leq j \leq n \quad (4)$$

3) 각 클러스터의 중심 v_i를 계산한다. x_j는 j번째 샘플이다.

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m} \quad (5)$$

4) 소속 행렬 U를 계산한다.

$$u_{ij} = \frac{\left(\frac{1}{d^2(x_j, v_i)}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{c}\left(\frac{1}{d^2(x_j, v_k)}\right)^{\frac{1}{m-1}}} \quad (6)$$

5) 목적함수 J_m을 계산한다. $d^2(,)$은 유클리디안거리 제곱이다.

$$J_m(X,U,V) = \sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^m d^2(x_j, v_i) \quad (7)$$

6) 다음의 종료조건이 만족할 때까지 3)~5)를 반복한다.

$$|\{J_m^{(l)} - J_m^{(l-1)}\}| \leq \varepsilon \quad (8)$$

위의 과정을 거친 클러스터 결과는 베이지안 검증 방법에 의해 검증된 후, 각 데이터 별 최적의 α-cut을 결정하여 DT의 훈련 데이터의 레이블로 설정한다. 규칙생성과정에서는 DT를 훈련시키고, 이를 바탕으로 규칙들을 생성한다. 이렇게 만들어진 규칙을 가지고 각 데이터에 대한 최적의 클러스터 분할을 찾는다. 최종적으로 검증이 끝난 클러스터 분할을 가지고 유전

자 발현 데이터를 분석한다.

DT의 훈련 데이터의 속성은 각 데이터의 퍼지 클러스터링의 결과인 소속행렬을 이용하여 생성한다. 소속행렬값을 0.0부터 1.0까지 0.1씩 증가시켜가며 총 10개의 구간으로 나누고 각 구간 내 범위의 소속행렬값을 가진 샘플들의 빈도를 계산하여 데이터의 총 샘플수로 나눈 값을 속성으로 정의하여 사용한다. 이렇게 정한 각 속성을 $A_1 \sim A_{10}$이라고 정의한다.

4. 실험 및 결과

4.1 실험 데이터

총 6개의 유전 발현 데이터를 사용하여 클러스터링하고 그 결과를 이용하여 DT의 훈련데이터를 생성하였다. 사용한 데이터는 표 2와 같다.

표 2. 데이터 설명

인덱스	데이터	샘플 수	속성 수	클래스 수
CC	Yeast cell-cycle	421	17	20
SE	Serum	517	19	10
LE	Leukemia	38	50	2
LY	Lymphoma	45	44	2
SR	SRBCT	63	96	4

4.2 실험 결과

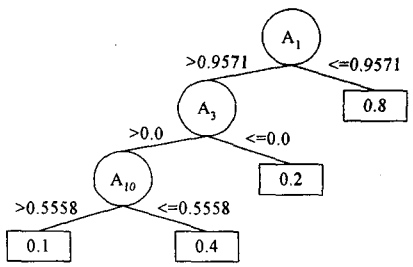

그림 3. DT에 의해 생성된 규칙

훈련 데이터로부터 만들어진 규칙은 그림 3과 같고 위의 규칙에 의해 결정된 각 데이터의 α-cut은 표 3과 같다.

표 3. DT의 규칙에 의해 결정된 최적의 α-cut

	CC	LE	LY	SR	SE
α-cut	0.4	0.8	0.8	0.2	0.1

첫 번째 속성(A_1)의 값이 0.9571보다 작거나 같으면 α-cut은 0.8로 분류가 되고, 그렇지 않으면 A_3의 값에 의해 α-cut이 0.2로 결정된다. 즉 LE와 LY는 0.0~0.1사이의 소속정도를 가진 샘플들이 다른 데이터에 비해 많이 존재함을 알 수 있다. A_3이 0보다 크면 다시 A_{10}에 의해 α-cut이 0.1과 0.4로 분류되기 때문에 SR은 0.2~0.3사이의 소속정도를 가진 샘플들이 존재하지 않음을 알 수 있으며 CC와 SE는 0.8~0.9사이의 소속정도를 가진 샘플들이 다른 데이터에 비해 상대적으로 적은 분포를 보임을 알 수 있다. 이와 같이 각 데이터의 특성에 따라 α-cut을 결정할 수 있다.

표 3의 결과가 얼마나 정확하게 각 데이터의 클러스터 분할을 평가하는지 알아보기 위해 그림 3의 규칙에 의해 판명된 α-cut을 베이지안 검증방법에 적용하여 각 데이터를 분석하였다. 그림 4는 CC와 SE의 결과이고 그림5는 LE, LY, SR의 결과이다. 클러스터 수(c)가 증가되면서 변화하는 BS값을 보면, CC는 c=20 근처에서 가장 높은 BS값을 기록하며 SE는 c=8~10에서 가장 높은 BS값을 보인다. 그림 5에서는 LE는 c=2, LY는 c=7, 그리고 SR은 c=4에서 각각 가장 높은 BS값을 보인다. 제안하는 방법이 LY를 제외한 나머지 모든 데이터에 대해 정확한 클러스터 분할의 평가를 하고 있음을 알 수 있다.

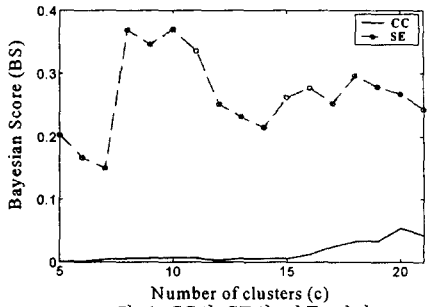

그림 4. CC와 SE의 검증 결과

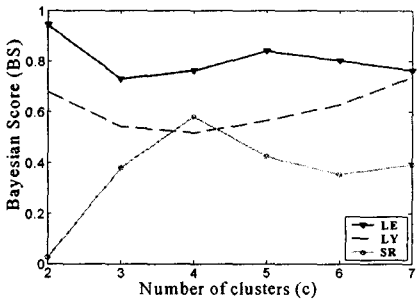

그림 5. LE, LY, SR의 검증 결과

5. 결론

본 논문에서는 데이터에 적응적인 유전자 클러스터링의 결과를 검증하기위해 결정트리 기반의 베이지안 검증 방법을 제안하였다. 제안하는 방법은 기존의 방법과는 다르게 데이터로부터 정보를 얻어 데이터에 적합한 최적의 α-cut을 자동으로 찾아내어 클러스터 분할을 검증한다. 실험 결과, 제안하는 방법의 우수한 성능을 확인할 수 있었다.

감사의 글

본 연구는 보건복지부 보건의료기술 진흥사업의 지원에 의하여 이루어진 것임.

참고 문헌

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

[2] S. Ruggieri, "Efficient C4.5," *IEEE Trans. on Knowledge and Data Engineering*, vol. 14, no. 2, March/April, 2002.

[3] K. Y. Yeung, *et al.*, "Validating clustering for gene expression data," *Bioinformatics*, vol. 17, no. 4, pp. 309-318, 2001.

[4] N. Bolshakova and F. Azuaje, "Cluster validation techniques for genome expression data," *SIGPRO*, vol. 21, no. 82, pp. 1-9. 2002.

[5] 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.

[6] P. Baranyi, *et al.*, "A new method for avoiding abnormal conclusion for α-cut based rule interpolation," *8th IEEE Int. Conf. on Fuzzy Systems*, Seoul, Korea, 1999, pp. 383-388, 1999.