최적의 유전자 클러스터 분석을 위한 퍼지 c-Means 알고리즘 기반의 베이지안 검증 방법

유시호° 원홍희 조성배
연세대학교 컴퓨터과학과
bonanza@candy.yonsei.ac.kr° cool@candy.yonsei.ac.kr sbcho@cs.yonsei.ac.kr

Bayesian Validation Method based on Fuzzy c-Means Algorithm for Analysis of Optimal Gene Clustering

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

요 약

수천 개의 유전자 발현 정보를 가지고 있는 DNA 마이크로어레이 기술의 발달로 대량의 생물정보를 빠른 시간 내에 분석하는 것이 가능하게 되었다. 유전자를 분석하는 방법 중 하나인 클러스터링 방법은 비슷한 기능을 가진 유전자들을 집단화시켜서 집단내의 유전자들의 기능을 밝히거나, 미지의 유전자를 분석하는데 이용되고 있다. 본 논문에서는 유전자 데이터를 분석하기 위한 퍼지 클러스터링 방법과 이를 효과적으로 검증할 수 있는 베이지안 검증 방법을 제안한다. 퍼지 c-means 알고리즘을 사용하여 클러스터를 생성하고, 클러스터 결과를 기존의 퍼지 클러스터 검증 방법들과 본 논문에서 제안하는 베이지안 검증 방법을 사용하여 비교 평가한다. 베이지안 검증 방법은 각 유전자의 클러스터 멤버쉽을 확률로 이용하여 각 클러스터에 속할 확률을 계산하고, 이 값을 가장 크게 해주는 클러스터 집단을 선택한다. 이 방법은 기존의 퍼지 클러스터 검증 방법들과는 달리 클러스터 수에 무관한 평가가 가능한 장점을 가지고 있다. Serum과 Yeast 데이터에 대한 실험 결과, 베이지안 검증 방법의 유용성을 확인할 수 있었다.

1. 서 론

지속적으로 증가하는 생물정보는 사람의 처리능력을 넘어섰으며 데이터 마이닝과 같은 인공지능 기법이 필수적으로 요구된다. 한번에 수천 개의 유전자 발현 정보를 획득할 수 있는 DNA 마이크로어레이 기술은 대량의 생물정보를 가진 대표적인 신기술로 질병의 진단 및 예측에 있어 새로운 분석방법들과 연계하여 많은 연구가 진행중이다. 이러한 새로운 기술들을 이용하여 유전자의 메커니즘을 규명하는 것은 질병의 치료 및 신약의 개발에 많은 도움을 줄 것으로 기대된다.

유전자를 분석하는 방법 중 하나인 클러스터링 방법은 비슷한 기능을 가진 유전자들을 집단화시켜서 집단내의 유전자들의 기능을 밝히거나, 미지의 유전자 기능을 분석하는데 이용되고 있다. 이 중에서도 퍼지 클러스터링은 일반적인 클러스터링과는 달리 하나의 유전자가 하나의 클러스터에만 속하지 않고 다수의 클러스터에 중복되게 속할 수 있으며, 그 속한 정도를 나타낼 수 있다. 생명체의 경우 하나의 유전자가 다수의 기능을 가질 수 있기 때문에 유전 데이터의 분석에는 퍼지 클러스터링이 일반적인 클러스터링 방법보다 더 효율적이라고 할 수 있다. 하지만 이러한 퍼지 클러스터링의 결과는 검증된 것이 아니기 때문에 이를 그대로 분석에 적용하기에는 한계가 있다. 그렇기 때문에 클러스터 결과를 체계적으로 검증하는 절차가 꼭 필요하며 실제로 이와 관련된 연구들이 많다.

일반적으로 Partition coefficient, Classification entropy, Proportion Exponent, Xeni-Beni index 등이 많이 사용되고 있으며[1], 일반적인 클러스터 검증 방법에 사용되는 Silhouette index 방법으로 임계값을 설정하여 퍼지 클러스터링에 사용한 연구도 있고[2], 최종 클러스터들의 응집성(compactness)과 분리성(separation)에 따라 평가하는 방법을 제안하는 논문도 있다[3].

본 논문에서는 퍼지 c-means 알고리즘을 사용하여 클러스터를 생성하고, 클러스터 결과를 베이지안 검증 방법을 사용하여 평가한다. 또한 기존의 퍼지 클러스터 검증 방법(Partition Coefficient, Classification Entropy, Xeni-Beni index)들과 본 논문에서 제안하는 베이지안 검증 방법을 비교 평가한다. 기존의 검증 방법들은 클러스터의 수가 증가함에 따라 단조 감소 또는 단조 증가하는 한계를 가지고 있으나, 베이지안 검증 방법은 클러스터수와 무관하게 클러스터링의 결과를 검증할 수 있다. 각 유전자의 클러스터 멤버쉽을 확률로 이용하여 각 클러스터에 속할 확률을 가장 크게 해주는 클러스터 집단을 선택한다. Serum과 Yeast 유전 발현 데이터를 통한 실험 결과, 베이지안 검증 방법의 유용성을 확인할 수 있었다.

2. 퍼지 클러스터 검증 방법

본 논문에서는 가장 널리 사용되며, 그 방법 자체에 대한 검증이 끝난 대표적인 3가지 퍼지 클러스터 검증 방법을 사용하여 실험하였다[1].

2.1 Partition Coefficient(PC)

가장 많이 사용되고 있는 검증 방법 중의 하나로 간단

한 계산을 통해 퍼지 결과를 검증한다. 식(1)에서 u_{ij}는 멤버쉽 값이고, n은 샘플 수, 그리고 c는 클러스터의 수이다. 이 방법은 각 경우에 대한 총합을 PC로 두고 그 값이 1에 가까워질수록 클러스터가 잘 형성된 것으로 본다. 하지만 c의 값이 증가함에 따라 단조 감소하는 한계를 가지고 있기 때문에 실제 다수의 클래스로 구성된 유전 데이터에 적용하기에는 부적합하다.

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

2.2 Classification Entropy(CE)

Partition coefficient와 마찬가지로 가장 많이 사용되고 있는 퍼지 클러스터링 평가 척도 중의 한 방법이다. 식(2)를 보면 이 방법은 PC와 거의 비슷한데, 다만 u_{ij}에 로그값을 취했다. 이 방법은 최종 결과값인 CE값이 작을수록 클러스터가 잘 형성된 것으로 본다. 하지만 이 방법 역시 geometrical한 속성이 부족하며, PC와 마찬가지로 c값이 증가함에 따라 CE값이 단조감소현상을 보인다.

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

2.3 Xie-Beni index

Xie-Beni index 방법은 퍼지 파티션 결과가 얼마나 compact하고 separate한가를 측정하는 평가 척도이다. 응집성(compactness)은 데이터의 총 변화의 가중치와 관측회수의 비율에 의해서 결정되며, 분리성(separation)은 클러스터간의 거리에 의해서 측정된다. 식(3)에서 d_{min}은 가장 가까운 클러스터간의 거리를 나타낸다. 이 방법은 최종값인 XB값이 작을수록 compact하고 separate한 정도가 큰 좋은 클러스터라고 말할 수 있다.

$$XB(U,V;X) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^2\|V_i - X_j\|^2}{nd_{\min}^2} \quad (3)$$

$$d_{\min} = \min_{i,j}\|V_i - V_j\|$$

3. 베이지안 검증 방법

기존의 퍼지 클러스터링 검증 방법들은 클러스터의 수(c)에 따라서 감소하는 현상을 보인다. 즉 클러스터 수가 증가함에 따라 단조로운 감소를 보이는 경우가 많다. 특히 PC나 CE의 경우 대부분 클러스터수가 적을 때(c=2) 가장 적합한 결과를 보이기 때문에 다수의 클래스를 가진 데이터에 적용하는 경우 틀린 결과를 보이기 쉽다.

본 논문에서 제안하는 클러스터 검증 방법인 베이지안 검증 방법은 클러스터의 수(c)와 무관하게 클러스터의 평가가 가능하기 때문에 기존의 검증 방법들이 갖는 한계점을 보완한다. 베이지안 검증 방법은 식(4)와 같이 주어진 데이터에 대하여 퍼지 클러스터의 결과인 멤버쉽 함수 값을 이용하여 클러스터 집단이 형성될 확률을 계산하고, 계산된 결과들 중에서 해당 클러스터에 속할 확률들의 합이 가장 큰 클러스터 집단을 최적의 클러스터 집단으로 평가한다[4].

또한 퍼지한 결과를 그대로 검증 방법에 도입하기 위

$$\max_i P(Cluster_i \mid Dataset_i) \quad (4)$$

해서 α-cut을 도입하여 특정 임계값 이상의 멤버쉽 함수 값을 가진 샘플들만을 선택하여 계산하였다. 그림 1에서 멤버쉽 행렬은 퍼지 클러스터 결과로 각 유전자가 각 클러스터에 속한 정도를 매트릭스로 표현하였다. 그리고 D_1은 c_1에 속하면서 $u_{ik}>\alpha$인 조건을 만족하는 샘플들의 집합을 나타낸다. 그리고 최종적으로 Bayesian score(BS)를 계산하여 클러스터를 검증한다.

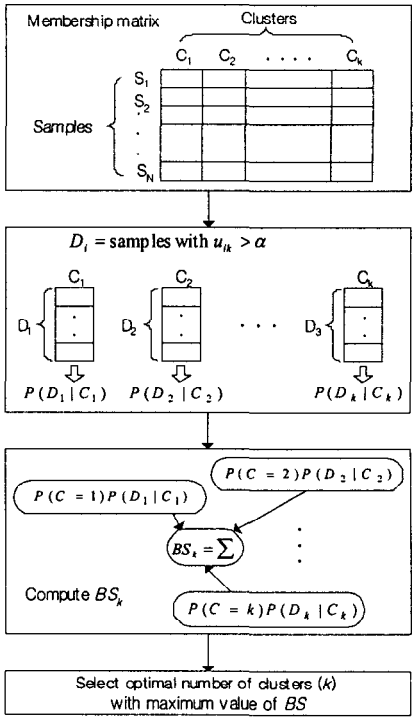

그림 1. 베이지안 검증 방법

그림 1에서 $P(D_k|C_k)$와 BS_k는 각각 식(5)와 식(6)을 이용하여 계산한다.

$$P(D_k \mid C_k) = \prod_{i \in D_k} P(x_i \mid C_k) \quad (5)$$

$$BS_k = \sum_k P(C=k)P(D_k \mid C_k) \quad (6)$$

베이지안 검증 방법의 최종 결과값인 BS는 큰 값을 가질수록 최적의 클러스터 분할로 평가한다.

4. 실험 및 결과

4.1 실험 환경

Serum 데이터는 알려진 517개의 유전자를 사용하였고 19개의 샘플이 있다. Yeast 데이터는 6153개의 유전자를 사용하였고, 93개의 샘플이 있다. 두 데이터 모두 temporal 데이터이다. 실험은 클러스터의 수를 변화해가

면서 각 경우에 대한 검증을 해보았다. 베이지안 검증 방법의 경우, α-cut의 값은 0.1부터 0.4까지 0.1단위로 증가시키면서 실험을 반복하였다. 0.5이상은 하나의 클러스터에 대해 절반 이상의 확률을 가지므로 실험에서 제외시켰다. 일반적인 클러스터링의 경우처럼 하나의 유전자가 하나의 클러스터에 속하는 경우가 많기 때문에 굳이 퍼지 클러스터링 결과에 적용시키는 의미가 없어진다.

4.2 실험 결과

그림 2는 Serum 데이터에 대하여 α-cut의 α값을 0.1부터 0.1씩 증가시키면서 관찰한 BS값의 변화를 보여준다.

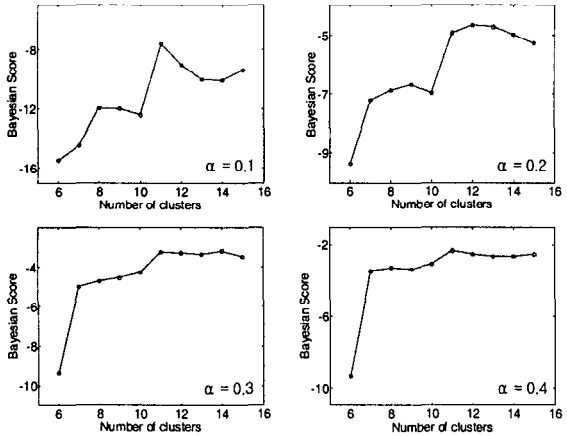

그림 2. α-cut 실험 결과(Serum 데이터)

Serum 데이터는 지금까지 10개 정도의 클러스터를 가지고 있는 것으로 밝혀져 있는데[5], 베이지안 검증 방법의 경우 α값에 따라 조금씩 차이는 보이지만, 보통 10~12개의 클러스터를 형성할 때 BS값은 가장 큰 값을 보인다.

표 1. α-cut 실험 결과(Yeast 데이터)

	α=0.1	α=0.2	α=0.3	α=0.4
10	-980.37	-461.30	-245.99	-135.88
15	-868.68	-333.28	-148.75	-69.930
20	-742.92	-235.02	-78.415	-32.594

표 1은 α값에 따른 Yeast 데이터의 결과로 클러스터의 수가 20개일 때 모든 경우가 가장 큰 BS값을 보인다. Yeast 데이터는 보통 20개 또는 그 이상의 많은 클래스로 구성된 것으로 알려져 있기 때문에 베이지안 검증 방법은 타당한 클러스터 집단을 선택하는 것을 알 수 있다[2].

표 2는 Serum 데이터의 결과로, 기존 퍼지 클러스터 검증 방법과 베이지안 검증 방법에 대한 결과를 나타내고 있다. 표 3은 Yeast 데이터의 결과로, 표 2와 마찬가지로 기존 검증 방법들과 베이지안 검증 방법의 성능을 비교하고 있다. 클러스터 수는 Serum의 경우 6부터 15까지 설정하였고, Yeast의 경우는 10에서 20까지 설정하고 실험하였다. 각 경우에 있어서 베이지안 검증 방법이 기존의 검증 방법들에 비해 정확한 성능을 보이는 것을 알 수 있다.

표 2. 각 검증 방법들의 결과값(Serum 데이터)

	PC(max)	CE(min)	XB(min)	BS(max)
6	0.9397	0.0464	0.3721	-9.3555
7	0.9474	0.0409	0.4494	-3.4357
8	0.9507	0.0388	0.4030	-3.2499
9	0.9368	0.0486	0.6715	-3.3370
10	0.9375	0.0487	0.6144	-3.0467
11	0.9417	0.0478	0.5839	-2.3399
12	0.9356	0.0522	0.5291	-2.5074
13	0.9351	0.0531	0.4992	-2.6269
14	0.9249	0.0624	1.0484	-2.6589
15	0.9279	0.0598	0.9901	-2.5153

표 3. 각 검증 방법들의 결과값(Yeast 데이터)

	PC(max)	CE(min)	XB(min)	BS(max)
10	**0.4968**	**0.4344**	**3.0870**	-135.887
15	0.3884	0.5808	4.2287	-69.930
20	0.3240	0.6920	29.0131	-32.594

5. 결 론

본 논문에서는 퍼지 c-means 알고리즘을 통한 유전자 클러스터링의 결과를 베이지안 검증 방법을 사용하여 평가하는 방법을 제안하였다. Serum과 Yeast 데이터를 사용하여 기존 퍼지 클러스터 검증 방법들과 비교한 결과, 베이지안 검증 방법은 c의 크기에 무관한 평가가 가능하며 그 정확성 역시 기존 방법들에 비해 우월한 성능을 보임을 확인하였다.

감사의 글

본 논문은 한국전자통신연구원의 지원에 의하여 이루어진 것임.

참고 문헌

[1] H. S. Rhee and K. W. Oh, "A validity measure for fuzzy clustering and its use in selecting optimal number of clusters," in *Proc. of IEEE International Conference on Fuzzy Systems*, vol. 2, pp. 1020-1025, 1996.

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

[3] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841-847, 1991.

[4] 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.

[5] V. R. Lyer, et al., "The transcriptional program in the response of human fibroblasts to serum," *Science*, vol. 283, no. 5398, pp. 83-87, 1999.