Gath-Geva 알고리즘을 이용한 유전자 발현 데이터의 분석

박한샘[0] 유시호 조성배
연세대학교 컴퓨터과학과
sammy@sclab.yonsei.ac.kr[0] bonanza@sclab.yonsei.ac.kr sbcho@cs.yonsei.ac.kr

Analysis of Gene Expression Data Using Gath-Geva Algorithm

Han-Saem Park[0], Si-Ho Yoo and Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

다량의 유전자 발현 정보를 담고 있는 DNA 마이크로어레이 기술의 발달로 인해 대량의 생물정보를 한번의 실험을 통해 분석할 수 있게 되었다. 유전자 발현 데이터를 분석하는 방법 중 하나인 클러스터링은 비슷한 기능을 가진 유전자들을 그룹별로 묶어서 그룹 내의 유전자들의 기능을 밝히거나, 미지의 유전자를 분석하는데 이용되고 있다. 본 논문에서는 유전자 발현 데이터를 클러스터링하여 그로부터 유전 정보를 찾아내기 위한 방법으로 GG (Gath-Geva) 알고리즘을 제시한다. 퍼지 클러스터링 알고리즘중 하나인 GG 알고리즘은 대표적인 퍼지 클러스터링 방법인 퍼지 c-means 와 GK (Gustafson-Kessel) 알고리즘을 개선한 것으로, 차원이 크고 분포가 애매하여 클러스터링이 어려운 유전자 발현 데이터의 클러스터링에 적합한 알고리즘이다. 혈청(Serum) 유전자 데이터와 효모(Yeast) 세포주기 데이터를 GG 알고리즘 이용해 클러스터링 해 보고, 그 결과를 퍼지 c-means 알고리즘, GK알고리즘과 비교해 본 결과, GG 알고리즘이 유전자 발현 데이터의 클러스터링에 더 적합함을 확인하였다.

1. 서 론

한번에 수천 개 이상의 유전자 발현 정보를 획득할 수 있는 DNA 마이크로어레이 기술의 발달 등 생명공학과 정보 처리 기술이 발달함에 따라 이를 분석하기 위해 여러 가지 인공지능 기법이 필수적으로 필요하게 되었다. 마이크로어레이 기술은 대량의 생물정보를 포함한 대표적인 신기술로 질병의 진단 및 예측에 있어 여러 다른 분석 방법들과 연계하여 많은 연구가 진행되고 있다[1].

통계적 패턴분류 기법 중 하나인 클러스터링은 주어진 데이터의 집합에서 유사한 성질을 갖는 데이터를 묶어서 분할해 내는 방법이며 유전자 발현 정보 데이터와 같은, 레이블 정보가 없거나 혹은 불완전한 데이터를 분석하는데 유용한 방법이다. 클러스터링은 데이터를 각 클러스터로 분할하는 방법과 정도에 따라 하드 클러스터링 방법과 퍼지 클러스터링 방법으로 나눌 수 있다. 하드 클러스터링에서는 하나의 데이터가 하나의 클러스터의 형성에만 관여하며, 퍼지 클러스터링은 하나하나의 데이터의 각 클러스터에 대한 소속 정도에 따라 여러 개의 클러스터에 속할 수 있는 정도를 표현한다[2]. 일반적으로 실세계의 데이터는 명확하게 나누어지기 어려워서 나누어진 클러스터간의 경계가 분명하지 않기 때문에 하드 클러스터링보다는 퍼지 클러스터링이 실세계의 데이터를 분석하여 보다 의미 있는 결과를 찾아내는데 적합하다.

퍼지 클러스터링 알고리즘에도 여러 가지가 있으며, 유전자 발현 데이터의 분석에 퍼지 클러스터링을 응용한 예도 많이 있다. 하지만 그 중 많은 수의 연구가 퍼지 c-means 알고리즘을 사용하였다[3]. 퍼지 c-means 알고리즘은 대표적인 퍼지 클러스터링 알고리즘이며, 가장 많이 사용되어 왔지만 구형의 클러스터만을 형성하는 단점이 있다. 이런 퍼지 c-means 알고리즘의 단점을 보완하기 위해 GK(Gustafson-Kessel), GG(Gath-Geva) 등의 알고리즘이 제안되었다. 구형의 클러스터만을 형성하는 퍼지 c-means의 단점을 보완하기 위해 제안된 GK는 타원형 클러스터의 형성이 가능하다[4]. GG는 GK를 더 개선한 방법으로, 퍼지 c-means, GK와 달리 목적함수에 기반한 방법이 아닌 통계적인 추정법에 기반한 방법이다. 또한 퍼지 c-means나 GK가 잡아내기 힘든 분포의 데이터도 클러스터링이 가능하다[5]. 이런 점에서 고차원의 클러스터링이 어려운 분포를 하고 있는 유전자 데이터의 클러스터링에 적합하다고 할 수 있다.

앞에서도 언급했듯이 유전자 발현 데이터의 클러스터링과 관련해 많은 연구가 진행되어 왔다. 초기에는 하드 클러스터링만을 사용한 연구들이 많았지만, 유전자 데이터를 분석하는 데에는 퍼지 클러스터링이 더 적합하다. 퍼지 클러스터링을 통해 유전자 클러스터링을 한 연구로는 혈청 데이터, 효모 세포주기 데이터를 퍼지 c-means 로 분석한 Dembele과 Kastner의 연구[3], 효모 데이터를 퍼지 k-means 로 분석한 Eisen의 연구[6] 등이 있다.

본 논문에서는 유전자 데이터를 클러스터링하는 방법으로 GG를 제시하며, 퍼지 c-means, GK와의 비교 실험을 통해 유전자 데이터 클러스터링에 GG가 적합한지 여부를 확인해 보았다.

2. 배 경

2.1 퍼지 클러스터링 알고리즘

가. 퍼지 c-means 알고리즘

퍼지 c-means 알고리즘은 가장 널리 이용되는 퍼지 클러스터링 방법이다. 이 방법은 퍼지 이론을 적용한 목적 함수의 반복 최적화에 기반을 둔 방식으로 각 데이터가 특정 클러스터에 속하는 소속 정도를 줌으로써 데이터에 대한 보다 정확한 정보를 제공한다. 주어진 데이터 집합이 $X = \{x_1, x_2, \ldots, x_n\}$이고 각 클러스터의 중심 벡터가 $V = \{v_1, v_2, \ldots, v_c\}$ 일 때, 목적 함수는 각 데이터 x_j와 각 클러스터 중심 v_j와의 거리와 클러스터에 대한 소속 정도(degree of membership) 값으로 정의된다[3].

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

여기서 u_{ij}는 x_j와 j번째 클러스터에 대한 소속 정도를 나타내며 ($c \times n$)의 소속 행렬 $U = [u_{ij}]$의 원소이다. $d^2(,)$는 유클리디안 거리의 제곱이고, 매개 변수 m은 각 데이터의 소속 정도에 대한 퍼지 값을 나타내며 1 보다 큰 값을 사용한다.

퍼지 c-means 알고리즘의 수행절차는 다음과 같다.

1) 클러스터의 수 c와 퍼지 계수 m의 값을 정한다.
2) 다음의 조건을 만족하도록 x_j의 소속 정도인 u_{ij}를 초기화 한다.

$$\sum_{i=1}^{c}\mu_{ij} = 1, 1 \le j \le n \quad (2)$$

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

$$v_i = \frac{\sum_{j=1}^{n} \mu_{ij}^{m} x_j}{\sum_{j=1}^{n} \mu_{ij}^{m}} \quad (3)$$

4) 소속 행렬 U를 계산한다.

$$\mu_{ij} = \frac{\left(\frac{1}{d^2(x_j, v_i)}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{c}\left(\frac{1}{d^2(x_j, v_k)}\right)^{\frac{1}{m-1}}} \quad (4)$$

5) 다음의 종료 조건이 만족될 때까지 3)과 4)를 반복한다. l은 반복 단계를 의미한다.

$$|\{J_m^{(l)} - J_m^{(l-1)}\}| \le \varepsilon \quad (5)$$

나. GK (Gustafson-Kessel) 알고리즘

퍼지 c-means의 식에서 유클리디안 거리 대신에 공분산 행렬로부터 유도된 다른 거리 척도를 사용하면 구형의 클러스터뿐 아니라 타원형의 클러스터로도 클러스터링이 가능하다. GK 알고리즘은 구형의 클러스터만을 형성하는 퍼지 c-means의 단점을 보완하여 Gustafson과 Kessel에 의해 고안되었다. 타원형의 클러스터 형성이 가능해짐에 따라 퍼지 c-means 보다 융통성 있는 클러스터링이 가능하다.

$$A_i = \sqrt[p]{\det(S_i) S_i^{-1}} \quad (6)$$

$$d_{F_i}^{2}(x_j, v_i) = (x_j - v_i) A_i (x_j - v_i)^T \quad (7)$$

GK 알고리즘의 수행절차는 퍼지 c-means와 거의 같은데, 거리 계산식 (6), (7)이 다르며 퍼지 c-means 수행절차의 식(3)과 식(4) 사이에 공분산을 구하기 위한 다음의 식이 하나가 더 필요하다. 식(6)의 p는 데이터의 차원수이고, 식(7), (8)의 x는 데이터를, v는 클러스터의 중심을 각각 의미한다.

$$S_i = \frac{\sum_{j=1}^{d} \mu_{ij}^{w} (x_j - v_i)(x_j - v_i)^T}{\sum_{j=1}^{d} \mu_{ij}^{w}} \quad (8)$$

2.2 퍼지 클러스터링 검증 방법

가. Partition Coefficient(PC)

가장 많이 사용되고 있는 검증 방법 중의 하나이다. 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. 클러스터 내의 데이터들이 얼마나 조밀하게(compactly) 모여 있는지를 기준으로 삼는다[6].

$$PC(U;c) = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c}\left(u_{ij}\right)^2}{n} \quad (9)$$

나. Classification Entropy(CE)

Partition coefficient와 마찬가지로 가장 많이 사용되고 있는 퍼지 클러스터링 평가 척도 중의 한 방법이다. 이 방법은 최종 결과값인 CE값이 작을수록 클러스터가 잘 형성된 것으로 본다. 이 방법 역시 클러스터 내의 데이터들이 얼마나 조밀하게 잘 모여 있는지를 기준으로 삼는다[7].

$$CE(U;c) = \frac{-\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij} \log_a u_{ij}}{n} \quad (10)$$

3. 제안하는 방법

3.1. GG (Gath-Geva) 알고리즘

GG 알고리즘은 분포한 데이터의 밀도뿐 아니라 넓이도 고려하여, 밀도는 좀 작지만 넓게 분포한 클러스터도 클러스터링이 가능하도록 GK를 개선한 방법이다. 또한 클러스터링이 어려운 서로 중첩된 클러스터도 찾아 낼 수 있으며, 거리 계산식에 지수함수가 포함되어 거리에 따른 소속정도의 차이가 커지기 때문에 노이즈에 강한 장점을 갖는다[4].

$$d_{F_i}^{2}(x_j, v_i) = \frac{\sqrt{\det(S_i)}}{P_i} \exp\left(\frac{1}{2}(x_i - v_i)^T S_i^{-1}(x_i - v_i)\right) \quad (11)$$

목적함수는 퍼지 c-means, GK와 동일하고 $d^2(,)$만 다른데, 거리 계산할 때에 앞에서 계산된 사전 확률(priori probability)과 공분산 행렬을 모두 사용한다. 공분산 행렬 S_i의 경우 GK와는 약간 다른 식을 사용한다.

GG 알고리즘의 수행절차는 다음과 같다.

1) ~ 3) 퍼지 c-means, GK와 동일하다.

4) 사전 확률을 계산한다.

$$P_i = \frac{1}{n}\sum_{j=1}^{n} \mu_{ij} \quad (12)$$

5) 클러스터의 공분산 행렬 S_i를 계산한다.

$$S_i = \frac{\sum_{j=1}^{d} \mu_{ij} (x_j - v_i)(x_j - v_i)^T}{\sum_{j=1}^{d} \mu_{ij}^{w}} \quad (13)$$

6) 식 (4)를 이용해서 소속 행렬 U를 계산한다.

7) 종료조건은 퍼지 c-means와 동일하다.

3.2. 유전자 데이터의 클러스터링을 위한 GG 알고리즘

위에서 설명한 것처럼 GG 알고리즘은 퍼지 c-means나 GK 알고리즘을 여러 가지로 개선한 방법이며, 유전자 데이터의 클러스터링을 위해 매우 적합한 알고리즘이다.

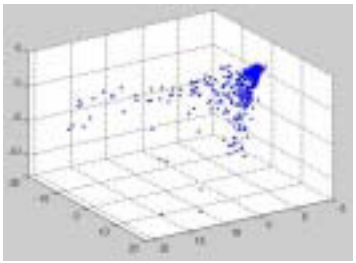

그림 1. 혈청 데이터의 분포

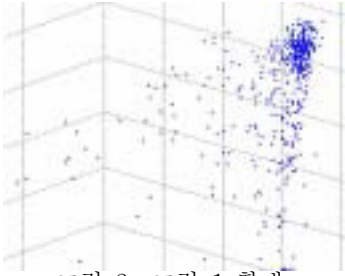

그림 2. 그림 1 확대

그림 1은 17차원의 혈청 유전자 데이터로부터 주성분 분석(PCA: Principal Component Analysis)을 이용해 주요 성분을 뽑아내서 3차원 데이터의 분포로 나타낸 그래프이다. 특별히 나뉘어 진 클러스터를 찾아보기 어렵다. 그림 2는 그림 1에서 데이터가 모여 있는 부분을 확대한 것이다. 그림 1보다는 클러스터링을 하기에 나아 보이지만 여전히 클러스터의 경계가 명확하지 않고 애매한 분포를 하고 있음을 확인할 수 있다.

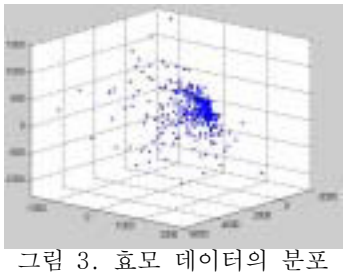

그림 3. 효모 데이터의 분포

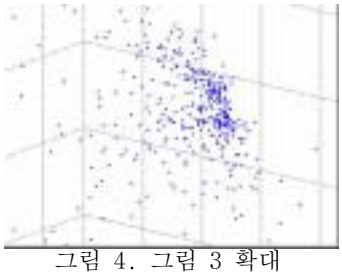

그림 4. 그림 3 확대

그림 3은 마찬가지로 주성분 분석방법을 이용해 21차원의 효

모 데이터로부터 추출해 낸 3차원의 효모 데이터의 분포이며, 그림 4는 그림 3의 데이터가 모여 있는 부분을 확대 한 것이다. 역시 구형이나 타원형의 클러스터를 잡아내기가 쉽지 않은 분포를 보이고 있다.

앞에서 언급했듯이, GG는 다른 방법으로는 찾기 어려운 서로 중첩된 클러스터를 찾아 낼 수 있으며, 노이즈에 강한 장점을 갖는다. 때문에 구형의 클러스터를 찾기 좋은 퍼지 c-means, 타원형의 클러스터를 찾을 수 있는 GK로는 찾기 어려운 경계가 명확하지 않은 클러스터를 찾는 데 장점을 갖는다. 유전자 데이터의 출처(source)인 마이크로어레이 칩은 유용한 자료이지만 아직까지는 적지 않은 노이즈를 포함하고 있는 점 또한 GG를 이용하면 좋은 이유라고 할 수 있다[8].

GG는 또한 퍼지 c-means, GK와는 달리 목적함수에 기반한 방법이 아닌, 통계적 추정방법인 퍼지 최대우도 추정법(Fuzzy Maximum Likelihood Estimation)을 이용한 방법이다. 사전확률을 계산하여, 그 확률 값에 따라 거리 및 소속정도가 바뀌게 된다[4].

4 실험 및 결과

4.1. 실험 데이터

가. 효모(Yeast) 세포주기 데이터

효모의 세포주기 데이터는 두 개의 세포 주기를 거치는 동안의 약 6000개의 유전자들의 발현 정도를 나타내는 데이터이다. 실험 과정에서 의미 있는 발현 변화를 보이는 420개의 유전자가 선택되었고, 본 논문에서는 의미 있는 421개의 유전자의 세포 주기의 17개의 시점에서 선택된 발현 데이터를 이용하여 클러스터링 하였다.

출처 http://www.yscdp.stanford.edu/yeast_cell_cycle

나. 혈청(Serum) 데이터

혈청 유전자 데이터는 517개 유전자의 19가지 실험 조건에 따른 유전자 발현 정도 값을 갖는다. 이 데이터는 사람의 섬유아세포에 혈청을 농축 시키고 농축 시간에 따라 다양하게 나타나는 각 유전자의 발현 정도를 측정하였다[3, 9].

출처 http://genome-www.stanford.edu/serum

4.2. 실험 결과

혈청 유전자 데이터와 효모 유전자 데이터를 퍼지 c-means, GK, GG를 이용하여 클러스터수를 바꿔가며 클러스터링 해 보고, PC와 CE값을 계산해서 방법별로 비교해 보았다.

가 혈청(Serum) 데이터 클러스터링 결과

표 1은 클러스터 수를 2개~12개로 바꾸어 가며 혈청 데이터를 퍼지 c-means, GK, GG로 클러스터링 한 결과의 PC와 CE값을 계산하여 정리한 것이다. C#은 클러스터 수를 의미한다.

표 1. 혈청 데이터 클러스터링 결과

	퍼지 c-means		GK		GG	
C#	PC	CE	PC	CE	PC	CE
2	0.941050	0.445610	0.957834	0.033238	0.999547	0.000345
3	0.922499	0.060756	0.847608	0.107245	0.999854	0.000097
4	0.872628	0.100212	0.798629	0.164987	0.999114	0.000793
5	0.833843	0.134836	0.779208	0.177488	0.989824	0.007438
6	0.847554	0.123697	0.782093	0.183229	0.999599	0.000319
7	0.846281	0.125561	0.762771	0.201828	0.997044	0.001043
8	0.833934	0.139291	0.705648	0.257718	0.989464	0.007834
9	0.831488	0.144416	0.672471	0.289250	0.999644	0.000386
10	0.795724	0.172061	0.760178	0.205925	0.998716	0.001267
11	0.778153	0.189716	0.725033	0.250905	0.992544	0.005530
12	0.776899	0.190776	0.745418	0.231450	0.999443	0.000600

앞에서도 설명했듯이 PC 값은 1에 가까울수록, CE값은 0에 가까울수록 클러스터링이 잘 된 것을 의미한다. 표 1을 보면 PC, CE값 모두 클러스터 수에 관계없이 GG가 압도적인 수치를 나타내고 있음을 확인할 수 있다. 한 가지 의외인 점은 GK의 수치가 퍼지 c-means의 수치를 밑돌고 있다는 점이다.

나 효모(Yeast) 데이터 클러스터링 결과

표 2는 클러스터 수를 2개~12개로 바꾸어 가며 효모 데이터를 퍼지 c-means, GK, GG로 클러스터링 한 결과의 PC와 CE값을 계산하여 정리한 것이다.

표 2. 효모 데이터 클러스터링 결과

	퍼지 c-means		GK		GG	
C#	PC	CE	PC	CE	PC	CE
2	0.893244	0.075222	0.927733	0.055053	0.999832	0.000179
3	0.896706	0.079186	0.914167	0.070186	0.999516	0.000534
4	0.897333	0.083606	0.849829	0.126044	0.998349	0.001631
5	0.862156	0.113285	0.826193	0.145355	0.999747	0.000273
6	0.858523	0.120882	0.814494	0.156762	0.999155	0.000894
7	0.840357	0.135851	0.804848	0.174925	0.999666	0.000410
8	0.829024	0.147410	0.813390	0.159451	0.999561	0.000521
9	0.806473	0.166808	0.787836	0.189317	0.999288	0.000738
10	0.794134	0.181408	0.755624	0.210941	0.999245	0.000751
11	0.808641	0.168289	0.765409	0.205102	0.999782	0.000315
12	0.789135	0.186557	0.712562	0.254758	0.993246	0.002668

마찬가지로 PC 값은 1에 가까울수록, CE값은 0에 가까울수록 클러스터링이 잘 된 것을 의미한다. 표 2에서도 역시 PC, CE값 모두 GG의 압도적인 수치를 확인할 수 있다.

5. 결 론

본 논문에서는 데이터의 특성상 차원이 크고 클러스터의 형성이 쉽지 않은 분포를 하는 유전자 발현 데이터를 클러스터링하는 방법으로 GG 알고리즘을 제시하였으며, 주요 퍼지 클러스터링 알고리즘인 퍼지 c-means, GK와의 비교 실험을 통해 PC, CE값을 비교해 본 결과 GG가 훨씬 좋은 결과를 보임을 확인하였다.

감사의 글

본 논문은 보건복지부의 보건의료기술 진흥사업의 지원에 의하여 이루어진 것임.

참고 문헌

[1] M. B. Eisen and P. O. Brown, "DNA arrays for analysis of gene expression," *Methods Enzymbol*, vol. 303, pp. 179-205, 1999.

[2] T. Van Le, "Evolutionary fuzzy clustering," *In Proc. IEEE Int. Conf. on Evolutionary Computation*, vol. 2, pp. 753-758, 1995.

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

[4] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, N.Y. Plenum, 1981.

[5] F. Hoppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis*, Wiley, pp. 43-39, 1999.

[6] 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.

[7] N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," *IEEE Trans. on Fuzzy Systems*, vol. 3, pp. 370-379, 1995.

[8] 박종화, *생명정보공학 심층 보고서*, KOSEN/OSTIN, 2000.

[9] 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 Int. Conf. on Fuzzy Systems*, vol. 2, pp. 1020-1025, 1996.