A Study on the Gen Expression Data Analysis Using Fuzzy Clustering

Hang Suk, Choi[1] · Kyung Joon, Cha[2] · Hong Goo, Park[3]

요 약

Microarry 기술의 발전은 유전자의 기능과 상호 관련성 그리고 특성을 파악 가능하게 하였으며, 이를 위한 다양한 분석 기법들이 소개되고 있다. 본 연구에서 소개하는 fuzzy clustering 기법은 genome 영역의 expression 분석에 가장 널리 사용되는 기법 중 비지도학습(unsupervized) 분석 기법이다. Fuzzy clustering 기법을 효모(yeast) expression 데이터를 이용하여 분류하여 hard k-means와 비교 하였다.

주요용어 : Fuzzy Clustering, Gene Expression Analysis, Hard K-means

1. 서 론

최근 microarray 기술의 발전은 수천의 유전자들의 expression 데이터를 동시에 모니터(monitor) 할 수 있게 되었고, 복잡한 유전학의 네트워크의 연구에 큰 역할을 하게 되었으며, 유전자 구조(gene regulation)를 밝히는 기초 자료로 활용하고 있다. 유전자 기능 및 구조를 밝히기 위한 실험 중, 경계표 실험(landmark experiment)은 알려지지 않은 세포의 유전자 활동에서 패턴을 전달하는 효모 유전체에 모든 유전자를 포함하고 있는 microarray에서 효모 세포 주기에 관한 연구이다[1].

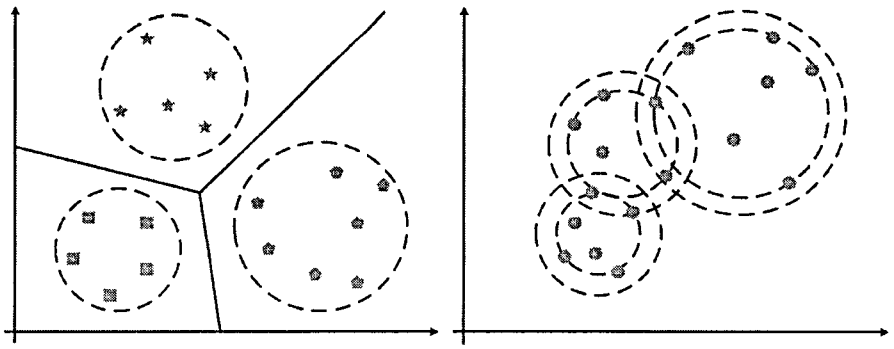
다양한 정보를 포함하고 있는 expression 데이터를 분석하기 위해 다양한 통계적 기법들이 연구되고 있으며, 그 중에서 유전자 기능을 파악하기 위해, genome 영역의 expression 분석에 가장 널리 사용되는 기법 중 하나는 다변량 분석 기법 중 비지도 학습(unsupervized) 분석인 clustering이다. 유전자 expression 데이터의 cluster 분석의 주된 목적은 유전자의 기능 분석, 유전자의 상호 관련성 분석, 각종 질병 진단 및 질병 관련 유전자 검출 등의 중요한 분야의 연구라 할 수 있다.

본 연구의 목적은 개체(object)들이 특정 cluster에 속하는 가능성을 소속의 정도로 나타내어 확률로 clustering하는 fuzzy clustering을 소개한다. 즉, 실제 자료에서 나타나기 쉬운 둘 또는 그 이상의 cluster에 속하는 개체를 확률과 퍼지정도(fuzziness)를 이용하여 개체를 clustering하는 방법이다.

1) Research Assistant Professor, Stat. Information Analysis Center, The Research Institute for Natural Sciences, Hanyang. Univ., 17 Haengdang-dong, Seongdong-Gu, Seoul, 133-791, Korea. E-mail : neuldol@ihanyang.ac.kr

2) Professor, Dept. of Mathematics, Hanyang. Univ., 17 Haengdang-dong, Seongdong-Gu, Seoul, 133-791, Korea.

3) Associate Professor, Dept. of Mathematics, Hanyang. Univ., 17 Haengdang-dong, Seongdong-Gu, Seoul, 133-791, Korea.



<그림 1> 클러스터가 3인 경우 hard k-means(left)와 fuzzy clustering(right)

본 연구의 구성은 2절에서 fuzzy clustering 알고리즘 중 fuzzy c-means(FCM)에 관한 소개와 알고리즘 검증을 위한 효모(yeast) 유전자 expression 데이터를 소개한다. 그리고 3절에서 hard k-means와 FCM의 비교 결과와 토의를 할 것이다.

2. Fuzzy C-Means 알고리즘과 효모(yeast) 데이터

Clustering이란 주어진 데이터 집합의 패턴들을 비교하여 비슷한 성질을 가지는 개체들을 여러 cluster로 나누는 방법이다[2, 3]. 데이터를 분류하는 Clustering 알고리즘으로 널리 알려진 것은 hierarchical clustering, hard k-means clustering, self-organizing maps(SOM) 등이 현재 유전자 expression 데이터 분석에 이용되고 있다.

Clustering 기법 중 가장 많이 사용되는 hard k-means는 주어진 데이터 상호간의 관계가 명확하다는 가정에서 각 패턴을 분할하는 방법으로 <그림 1>의 왼쪽과 같이 경계선으로 명확하게 분류가 가능한 경우에 활용 효과가 높다. 그러나 실제 데이터에서는 대부분의 경우 데이터의 경계가 명확하지 않기 때문에, hard k-means는 실제 데이터 상호간의 그룹화에 부적절하며 실제로 clustering하는 경우 주어진 데이터 분포의 손실이 발생한다.

이를 개선하기 위해, fuzzy 이론과 소속의 정도를 나타내는 확률을 이용한 기법으로 fuzzy c-means(FCM) 알고리즘이 있다[3, 4]. FCM 알고리즘은 여러 패턴이 특정 cluster에 속하는 소속 정도(확률)를 나타냄으로써 보다 정확하게 clustering이 이루어진다.

FCM 알고리즘은 하나의 객체가 여러 클러스터에 속할 가능성을 허용하는 확률 개념을 도입한 것으로 u_{ki}를 객체 i가 클러스터 k에 속할 확률이라 할 때 다음과 같이 분류한다.

$$J_m(u,c)=\sum_{k=1}^{K}\sum_{i=1}^{N}(u_{ki})^m d^2(x_i,c_k)$$

$$\sum_{k=1}^{K}u_{ki}=1 \quad \text{for all } i=1,\cdots,n$$

여기서, K와 N은 데이터 집합에서 cluster의 수와 유전자의 개수이고, $m(\in[1,\infty])$은 퍼지정도(fuzziness)를 나타내는 변수로 m이 1에 가까울수록 고전적 cluster 알고리즘인 hard k-means에 가깝고 큰 값 일수록 각 개체가 동일한 확률로 cluster에 배정되는 fuzzy clustering으로, 일반적으로 $m=2$인 퍼지정도를 이용하여 clustering을 시행한다. $d^2(x_i,c_k)$는 cluster 중

심 c_k에서 유전자 x_i의 거리이고, 데이터로부터 각 clusters에 대한 소속정도의 합이 1이 되는 확률적 제약조건(probabilistic constraint)이 주어진다.

c_k는 cluster k의 중심좌표로 hard k-means와 달리 라그랑지 방법(Lagrange method)으로 다음을 최적화한다.

$$J_m(u,c)=\sum_{k=1}^{K}\sum_{i=1}^{N}(u_{ki})^m d^2(x_i,c_k)-\sum_{i=1}^{N}\lambda_i\left(\sum_{k=1}^{K}u_{ki}-1\right)$$

즉, 최적화하기 위하여 u_{ki}와 c_k에 대하여 각각 편미분하여 그 값을 0으로 한다. 따라서,

$$u_{ki}=\frac{1}{\sum_{a=1}^{K}\left(\frac{d^2(x_i,c_k)}{d^2(x_i,c_a)}\right)}, \qquad c_k=\frac{\sum_{i=1}^{N}u_{ki}^m x_i}{\sum_{i=1}^{N}u_{ki}^m}$$

와 같이 유도된다.

이와 같은 fuzzy c-means의 전체 알고리즘 과정의 유도는 Bezdek가 소개하였으며, 알고리즘은 다음과 같이 4단계로 진행된다.

Step 1. K, m, ε의 수렴 조건을 결정한다.

$\sum_{k=1}^{K}u_{ki}=1$을 만족하는 소속 행렬(membership matrix) U를 임의로 초기화한다.

Step 2. 각 cluster의 중심 좌표 c_k를 계산한다.

Step 3. 객체 i가 cluster k에 속할 확률 u_{ki}를 계산한다.

Step 4. 객체 i를 u_{ki}가 가장 큰 cluster k에 속하게 만들고 clustering을 수행한다.
앞의 cluster 결과와 비교하여 동일하거나 변동이 ε보다 작으면 정지하고 그렇지 않으면 Step 1을 반복 수행한다.

FCM 알고리즘을 적용하기 위해 본 연구에서 사용되는 expression 데이터는 Chu 등이 연구한 효모(yeast) 데이터이다[5]. Saccharnoyces cerevisiae는 포자형성(sporulation)의 동시발생(synchrony)이 최대인 $t=0$에서 포자형성(sporulation) 배양기(medium)(SPM) 전달에 의해서 동시에 발생하고, RNA 데이터는 SPM이 전달된 후 $t=0, 0.5, 2, 5, 7, 9, 11.5$ 시간에 추출하였다. RNA는 oligo(dT) 섬유질(column)을 정제(purification)하여 실험이 진행되었다. 각 유전자들의 mRNA expression 수준은 SPM의 전달 전에 조절하여 사용되었다.

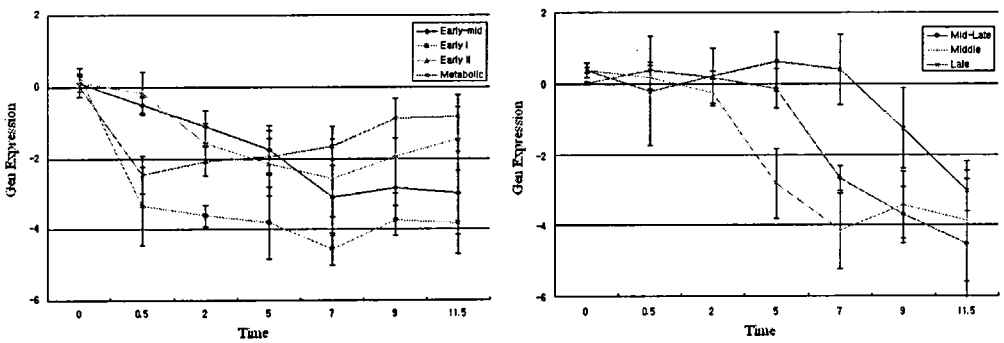
전체 데이터는 약 6000개의 유전자들의 expression profiles이 포함되어 있다[6]. 본 연구에서는 포자형성동안 mRNA의 중요한 증가를 보이는 Chu 등의 방법과 같이 유전자를 뽑아내는 방법을 사용하여 분석을 실시하였다[5]. FCM과 hard k-means 비교에 사용되는 유전자는 40이며 연구에 사용된 유전자는 Kupiec 등에 의하여 그 기능이 밝혀진 것이다[7]. 분석에 사용된 데이터는 유전자 특성에 따라 Early I, Early II, Early-Mid, Middle, Mid-Late, Late 그리고 Metabolic으로 나누었다.

<표 1> $t=2$(left)와 $t=7$(right)에서의 유전자 특성별 분산 분석 결과

유전자 특성	평균±표준편차	Duncan grouping
Late	0.2263±0.7656	A
Mid-Late	0.1769±0.1856	A
Middle	-0.2380±0.3791	A
Early-Mid	-1.0968±0.4465	B
Early II	-1.5636±0.5661	B C
Metabolic	-2.0700±0.4142	C
Early I	-3.6330±0.3042	D

유전자 특성	평균±표준편차	Duncan grouping
Late	0.3935±0.9773	A
Metabolic	-1.6416±0.5392	B
Early II	-2.5630±1.1153	B C
Mid-Late	-2.6749±0.3695	B C
Early-Mid	-3.1088±1.0656	D C
Middle	-4.1662±1.0637	D E
Early I	-4.5654±0.4453	E

Duncan's multiple range test by α=0.05



<그림 2> Mitchell에 의해 특성화한 유전자들의 시간대별 평균 변화량

효모(yeast) 데이터 분석을 위한 fuzzy c-means 알고리즘과 hard k-means 알고리즘은 R(version 2.0)을 이용하여 코딩하였으며, 데이터의 통계분석은 SAS(Statistical Analysis System, version 8.2, USA)를 이용하였고, 유전자 특성에 따라 시간대 별로 유의성 검정은 분산분석(ANOVA)을 실시하였다.

<표 1>는 유전자 특성별 검정에서 $t=2$와 $t=7$인 시점에서 유전자 특성별 차이가 있는지를 검정하였다. 그 결과 유전자 특성별 차이가 있는 것으로 나타났으며 시간대별로 다르게 나타나는 것으로 분석되었다. $t=2, 7$ 이외의 다른 시간대에서도 특성별 차이가 다르게 나타나 유전자 특성별 차이가 뚜렷한 것으로 나타났다.

<그림 2>에서 유전자 특성별로 expression은 시간에 따라 감소하고 있으며 감소 변화량의 차이가 크게 나타났다. <그림 2>의 왼쪽의 경우 유전자 별로 $t=0.5$까지 감소하였으며 그 후 변화가 작거나 증가를 보였으며, 오른쪽의 경우 $t=2$까지는 변화가 없다가 $t=5$ 이후 큰 폭으로 낮아지는 것으로 나타났다.

3. 결론 및 토의

본 연구에서는 각 cluster가 확연하게 분류되는 경우 효용이 높은 hard k-means(HKM)와 실생활에 활용 가능한 각 cluster에 속하는 가능성을 확률 및 퍼지정도(fuzziness)로 표현한 fuzzy c-means(FCM)을 비교 하였다.

<표 2> FCM과 HKM의 분류 정확도

	normal cluster	mis-cluster	정확도
Fuzzy c-means	38	3	0.925
Hard k-means	30	7	0.825

분산분석(ANOVA)에서 시간대별로 유의한 차이를 보인 효모(yeat)의 expression 데이터를 이용하여 fuzzy c-means와 hard k-means을 적용한 clustering 결과 <표 2>와 같이 분류 정확도는 40개 유전자 중 3개를 오류분한 FCM=0.925이었고, 40개 유전자 중 7개를 오분류한 HKM=0,825으로 나타났다.

효모(yeast) 데이터를 분산분석결과 뚜렷한 분류 ($t=2$)인 경우와 경계가 모호한 경우 ($t=7$)로 나타났는데, 대부분의 유전자 expression 데이터의 경우 각 cluster의 경계가 명확하지 않은 경우가 많이 존재하고 있기 때문에 기존에 활용되는 hard k-means의 경계 영역에 속하는 데이터의 분류 문제를 fuzzy clustering으로 해결 할 수 있을 것으로 본다.

향후 대용량의 자료를 이용한 정확한 FCM의 적용이 필요 할 것이며, 시간에 따른 유의한 변수를 결정하는 기법 및 필터링이 적용된다면 좀더 빠른 시간에 clustering이 이루어 질 것이다.

참고문헌

DeRisi, J.L. Iyer, V.R. and Brown, P.O. (1970), Exploring the metabolic and genetic control of gene expression on a genomic scale, Science, Vol.275, pp. 680-686.

Rhee, H. and Oh, K. (1996), A design and analysis of objective function based unsupervised neural networks of fuzzy clustering, Neural Processing Letters, Vol. 4, pp. 83-95.

Bezdek, J. (1980), A convergence theorem for the fuzzy ISODATA clustering algorithms, IEEE Trans. on Patt. Anla. and Mach. Intell, Vol. PAMI-2, No. 1, pp. 1-8.

Pal, N. and Bezdek, J. (1995), On Cluster Validity for the Fuzzy C-Means Model, IEEE Trans. on Fuzzy Sys., Vol. 3, No. 3.

Chu, S. Derisi, J. and Mulholland, J. et al. (1998), The transcriptional program of sporulation in budding yeast, Science 282, pp. 699-705.

The data set is available at http://cmgm.stanford.edu/pbrown/sporulation/

Kupiec, M., Ayers, B., et al. (1997), The molecular and cellular biology of the yeast Saccaromyces, Cold Spring Harbor, pp. 889-1036.