DNA microarray 자료의 군집 분석

권 성우, 한 종훈

포항 공과 대학교 화학공학과

1. 서론

생명 공학의 발전은 생명체 정보들을 대량으로 얻어내는데 큰 역할을 하고 있다. 특히, DNA에 있는 유전자 정보들을 분석해 내기 위한 DNA microarray 기술은 대량의 유전자 발현 정보를 만들어 내게 된다. 이러한 실험 데이터들을 생물학자들이 연구할 수 있는 의미 있는 정보로 조직화 하고 분석하는 과정에 통계학과 전산학이 사용되고 있다. DNA microarray 데이터에 대한 효율적인 분석 기법은 유전자의 기능 분석, 유전자의 상호 관련성 분석, 각종 질병 진단 및 질병 관련 유전자 검출 등의 중요한 분야의 연구에 크게 기여하고 있다.

본 글에서는 DNA microarray 데이터 분석에 가장 많이 사용 되고 있는 군집 분석(clustering analysis)이라는 다변량 분석 방법에 대해 설명 할 것이다. 군집 분석은 어떤 유전자나 암 환자들을 밀접한 상사성 또는 거리에 의하여 유사한 특성을 지닌 몇 개의 군집으로 집단화 하는 다변량 기법이다. 집단의 수 혹은 집단 구조에 대한 가정이 없으며, 오직 개체들 사이의 유사도에 의해 군집을 형성하고, 형성된 군집의 특성을 파악하여 군집들 사이의 관계를 분석하는 기법이다. 군집 분석의 주요한 두 가지 유형을 살펴보면 계보적 군집(hierarchical clustering)과 분리 군집 (partitioning clustering)이 있다. 또한 본 글에서는 발현 양상이 비슷한 유전자들이나 같은 종류 암 환자들을 편의상 군집이라 하고 각 각의 유전자나 암 환자 표본을 개체라는 통칭을 사용하고자 한다.

2. 본론

1. 유사도 정의

군집 분석에서 상사성이 높은 개체(유전자 혹은 환자)들은 같은 군집(발현 양상이 비슷한 유전자들 혹은 같은 종류의 암 환자)에 포함시키고, 상대적 상사성이 낮은 개체(발현 양상이 다르거나 다른 종류의 암 환자)들은 서로 다른 군집에 포함시킬 수 있도록 해주는 상사성이나 비상사성을 측정하는 유사도 정의가 필요 한다. 이러한 유사도 정의에 주로 사용되는 것은 거리인데 대표적으로 사용되는 거리 척도에는 Euclidean 거리이다.

$$d_{ij} = d(X_i, X_j) = \sqrt{(X_i - X_j)^T (X_i - X_j)}$$

한편 Euclidean 거리는 사용된 척도에 따라서 거리 순위에 상당한 영향을 미치게 된다. 이러한 문제를 극복하기 위하여 일반적으로 각 변수를 표준 편차로 나눈 표준화 변수를 사용하게 된다. 그러나 변수들 사이의 상관관계가 존재 할 때 거리는 척도의 불변성과 상관관계를 고려한 통계적 거리로 측정되어야 한다. 이러한 통계적 거리를 Mahalanobis 거리라고 한다.

$$d_{ij} = d(X_i, X_j) = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)}$$

한편 Euclidean 거리 이외에 개체들 간의 유사도 평가 기준으로는 pearson correlation coefficient, spearman correlation coefficient나 mutual information이 사용되기도 한다. 유사도에 대한 정의를 어떻게 했느냐가 군집의 질에 영향을 미치게 된다.

2. 계보적 군집 방법

계보적 군집 방법에는 상사성이 밀접한 개체(유전자나 암 환자)들을 단계적으로 발현 양상이 비슷한 유전자들이나 같은 종류의 암 환자들로 이루어진 군집을 형성해 나가는 병합적 방법 (agglomerative method)이 많이 사용 된다. 병합적 방법에서는 각 개체가 별개 군집을 형성하기 때문에 n개 군집에서 출발한다. 연속되는 각 단계 마다 가까운 두 개 군집들을 병합하면, 각 단계 마다 군집의 수가 하나씩 줄어든다. 따라서 마지막 단계에서는 모든 개체들이 하나의 군집을 형성하게 된다. 이러한 계보적 방법에 의한 군집 형성 결과는 이차원 상의 도면에 나타낸 dendrogram 형식으로 표시될 수 있다 (그림 1).

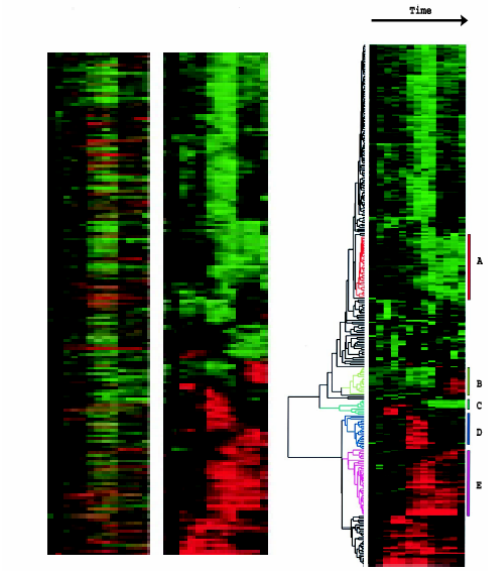

그림 1. 왼쪽이 계층적 군집 분석을 시작하기 전이고 오른쪽은 계층적 군집 분석을 수행한 후의 결과이다. 왼쪽에 비해 오른쪽이 발현 양상이 비슷한 유전자들이 군집화 된 것을 알 수가 있다.

병합적 방법으로는 두 군집 사이의 거리에 대한 정의 방법에 따라 군집 연결 방법이 달라지는데 최단 연결법, 최장 연결법, 평균 연결법, 중심 연결법 등이 있다. 최단 연결법은 우선 (NxN) 거리 행렬 D에서 우선 거리가 가장 가까운 개체가 U, V라면 두 개체를 묶어서 군집 (U V)를 형성한다. 다음 단계는 군집 (U V)와 나머지 (N-2)개의 다른 개체 중 임의의 개체 W와의 최소 거리를 다음과 같이 계산한 후

$$d_{(UV)W} = \min\{d_{uw}, d_{vw}\}$$

거리 행렬에서 거리가 가장 가까운 개체와 다시 군집을 형성한다. 이러한 과정을 반복 하면 모든 개체를 포함하는 하나의 군집을 형성하게 된다.

최장 연결법은 우선 (NxN) 거리 행렬 D에서 우선 거리가 가장 가까운 개체가 U, V라면 두 개체를 묶어서 군집 (U V)를 형성한다. 다음 단계는 군집 (U V)와 나머지 (N-2)개의 다른 개체 중 임의의 개체 W와의 최대 거리를 다음과 같이 계산한 후

$$d_{(UV)W} = \max\{d_{uw}, d_{vw}\}$$

거리 행렬에서 거리가 가장 가까운 개체와 다시 군집을 형성한다. 이러한 과정을 반복 하면 모든 개체를 포함하는 하나의 군집을 형성하게 된다.

평균 연결법은 우선 (NxN) 거리 행렬 D에서 우선 거리가 가장 가까운 개체가 U, V라면 두 개체를 묶어서 군집 (U V)를 형성한다. 다음 단계는 군집 (U V)와 나머지 (N-2)개의 다른 개체 중 임의의 개체 W와의 평균 거리를 다음과 같이 계산한 후

$$d_{(UV)W} = \frac{\sum_i \sum_j d_{ij}}{n_{(uv)} n_w}$$

거리 행렬에서 거리가 가장 가까운 개체와 다시 군집을 형성한다. 이러한 과정을 반복 하면 모든 개체를 포함하는 하나의 군집을 형성하게 된다.

중심 연결법은 두 군집 사이의 거리는 두 군집의 중심간의 거리로 계산 된다. 만약 군집 C_1에 속하는 개체의 수가 N_1 일 때 군집 C_1의 중심은

$$\overline{X}_1 = \sum_{i \in c_i} \frac{X_i}{n_1}$$

이 된다. 군집 C_2에 속하는 N_2개체의 중심을 $\overline{X}_2$라고 하면 두 군집 사이의 거리는

$$d_{(C_1)(C_2)} = P(\overline{X}_1, \overline{X}_2) = \left|\overline{X}_1 - \overline{X}_2\right|^2$$

이 된다. 여기서 P는 근접 척도로 Euclidean 거리의 자승이 사용된다. 만약 두 군집이 결합되면 새로운 군집의 중심은 가중 평균을 이용하여 구하고 각 군집 사이의 거리를 구한후 중심 거리가 가장 가까운 개체와 다시 새로운 군집을 형성한다. 이러한 과정을 반복하면 모든 개체를 포함하는 하나의 군집을 형성하게 된다.

3. 분리 군집 방법

계보적 군집 방법에서는 일단 어떤 개체가 특정한 군집에 할당되면 다른 군집에 다시 할당될 수 없는 단점을 가지고 있다. 반면에 분리 군집 방법은 어떤 개체가 초기 할당에서 잘못되었다 하더라도 다시 할당할 수 있는 방법이다. 이 분리 군집 방법은 미리 설정된 기준의 최적화에 근거해서 자료를 분리하게 된다. 또한 이 방법을 사용할 때에는 최종 군집의 수가 알려져 있고 또한 미리 설정될 수 있다고 가정한다.

분리군집 방법으로 가장 대표적인 k-평균 군집 방법은 다음과 같다. n개 개체(유전자나 암 환자)가 p차원 다변량 개체라고 했을 때 각 개체는 초기에 설정된 k개의 발현 양상이 비슷한 유전자들이나 같은 종류의 암 환자들로 이루어진 군집중 어느 한 군집에 할당된다고 가정하자. 이때 i번째 개체의 j번째 변수를 X(i,j)로 표시하고, c번째 군집에 속한 nc개 개체들의 j번째 변수에 대한 평균을 $\overline{X}$ (c,j)로 표시했을 때 i번째 개체와 c번째 군집 사이의 Euclidean 거리는 다음과 같이 표시 할 수 있다.

$$D(i,c)=\left(\sum_{j=1}^{p}\left[X(i,j)-\overline{X}(c,j)\right]^2\right)^{0.5}$$

또한 각 개체를 c번째 군집에 재 할당할 때 오차 자승 합 E는

$$E=\sum_{i=1}^{n}\left[D(i,c(i))\right]^2$$

이 된다. 위 식에서 c(i)는 군집 c는 i번째 개체를 포함하고 있다는 것이고, D(i,c(i))는 i번째 개체와 그 개체를 포함하고 있는 군집 사이의 Euclidean 거리를 표시한다. 따라서 분리 군집 방법은 각 개체를 어느 한 군집으로부터 다른 군집으로 움직일 때 오차 자승 합을 계산하여 비교하면서 더 이상 움직일 개체가 없을 때인 오차 자승 합이 최소화 하는 곳까지 반복한다.

4. 최근 제안 되는 군집화 방법

최근 들어 많은 전산학자와 통계 학자, 생물학자들이 개인 별로 또는 팀을 형성하여 군집화 방법을 개발하고 있으며 논문들이 발표되고 있다. 대표적인 논문들에 대해서 아래의 표에 설명 했다.

논문	발표 년도	사용 기법	실험 데이터	소프트 웨어	비고
Eisen et al	1998	계층적 군집화 분석	budding yeast cell cycle	공개	Cluster와 TreeView 프로그램
Tamayo et al	1999	SOM(self organizing map)	Budding yeast cell cycle	공개	

Tavazoie et al	1999	K 평균 군집화 분석	budding yeast cell cycle	공개	
Ben-Dor et al	1999	CAST		공개	Biocluster 프로그램
Sharan et al	2000	CLICK	Budding yeast cell cycle	라이센스 등록후 제공	계층적 군집화 분석과 비교 결과 제시

표 1 연구 동향

Eisen등은 통계학에서 널리 쓰이는 계층적 군집화를 DNA microarray에 적용하여 좋은 연구 업적을 남겼고, 이 연구의 산출물인 Cluster 와 TreeView라는 소프트웨어는 많은 사람에 의해 사용되고 있다. 이외에도 Hartuv는 그래프 이론을 바탕으로 군집화 방법을 제안하였고. Ben-Dor와 Shamir등도 역시 그래프를 응용하여 CAST라는 방법을 제안하고 프로그램으로 구현하였다. Tamayo등은 SOM(self –organizing maps)라는 방법을 적용하였으며 이외의 다수가 현재 군집화 방법의 개발 또는 평가 논문을 제시하고 있다.

5. 군집 분석시 유의 사항

군집을 형성하는 방법에는 여러 방법이 있지만 최적 방법에 대한 기준은 명확하지 않다. 따라서 군집 분석을 할 때 유의해야 할 사항을 정리하면 다음과 같다.

하나. 만약 집단들 사이에 상당한 차이가 없다면 현실적으로 군집 분석에서 매우 명확한 결과를 기대할 수 없을 것이다. 특히 만약 개체들이 비선형 형태로 분포되어 있다면 명확한 군집들을 찾아내는 것은 어려울 것이다.

둘. 군집 분석은 이상치에 대하여 상당히 민감한 결과를 보인다. 따라서 군집 분석을 시행하기 전에 자료에 대하여 이상치 존재 여부를 살펴야 한다고 판단된다.

셋, 군집 분석에 사용될 변수들의 측정 척도가 서로 다른 경우에는 분석 전에 표준화하여야 한다.

넷, 군집의 타당성을 검토하기 위하여 두 가지 방법이 사용된다. 첫 번째 방법은 자료에 대하여 여러 가지 군집 방법을 적용한 후 그 결과들이 유사한가를 검토하는 방법이다. 두 번째 방법은 자료를 랜덤 하게 이등분하고 각각에 대하여 여러 가지 군집 방법을 실시한 후 두 개의 결과가 유사한지 여부를 검토하여 일치도가 높은 군집 방법을 선택하는 방법이다.

3. 요약 및 연구 전망

DNA microarray 실험의 일반화와 유전자를 이용한 연구의 발전으로 인하여 데이터들은 급속히 증가되고 있다. 이러한 방대한 양의 정보들을 바탕으로 의미 있는 정보를 획득하는

데 있어서는 군집화 분석이 중요한 위치를 차지하고 있다. 그래프 이론에서의 접근, 신경망을 이용한 학습 기법에서의 접근, 확률 통계학적인 접근 방식으로 여러 군집화 분석이 연구되어 왔다. 특히 계층적 군집화와 분리 군집화 분석의 경우는 단순하면서도 어느 정도 의미있는 결과를 도출하는 알고리즘으로써 많이 활용되고 있으며 CLICK과 CLIFF는 군집화의 정확성을 더욱 높이는 결과를 가져왔다.

이 분야에 있어서 연구는 지속적인 발전이 있으리라고 본다. 생물학이라는 특수 영역에 의존된 데이터라는 특색과 유전 자수에 비해 실험 횟수는 매우 적은 특수한 환경 제약이 있기 때문에 이를 극복하기 위한 여러 가지 heuristic한 방법들이 계속해서 연구되리라고 전망된다.

참고 문헌

M. B. Eisen, P. T. Spellman, P.O. Brown and D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proceedings of National Academy of Sciences of the USA*, 95, 14863-14868, December 1998

P.Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander and T. R. Golub, interpreting patterns of gene expression with self-organizing maps: Method and application to Hematopoeitic differentiation, *Proceedings of National Academy of Sciences of the USA*, 96, 2907-2912, March 1999

R. Sharan and R. Shamir, CLICK: A clustering algorithm with applications to gene expression analysis, *In Proceedings ISMB 2000*, 2000

R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell,* 2, 6573, 1998