Proceedings of KIIS Spring Conference 2008 Vol. 18, No. 1.

다차원 데이터의 일부 차원을 반영한 지역 클러스터링

Local Clustering for Multidimensional Data

이선아[1] · 황경순[1] · 이건명[1] · 이찬희[2]
Sun A Lee, Kyung Soon Hwang, Keon Myung Lee, Chan Hee Lee
[1]충북대학교 전기전자컴퓨터공학부, 충북BIT지방연구중심대학사업단
E-mail: kmlee@cbnu.ac.kr
[2]충북대학교 생명과학부

요 약

다차원 데이터들에 대한 거리기반 클러스터링에서는 데이터의 전체 차원을 고려한 거리 정보를 이용하여 근접한 것들을 인접하게 만든다. 마이크로어레이 데이터의 경우에는 일부 차원 관점에서 유사한 지역 클러스터를 찾는 것이 분석에서 유용한 경우가 있다. 이 논문에서는 마이크로어레이 데이터에 대한 지역 클러스터를 찾는 방법을 제안한다.

키워드 : 클러스터링, 마이크로어레이, 생명정보학

1. 서 론

생명현상의 분석 및 규명을 위해 분자생물학적인 수준에서 사용되는 도구로서 마이크로어레이가 있다. 마이크로어레이는 유리, 필터 또는 실리콘 판 위에 유전자를 검출할 수 있는 많은 수의 프로우브를 붙여 놓거나 합성하여 놓아서, 동시에 많은 유전자에 대한 발현량을 측정할 수 있도록 한 것이다.[2] 하나의 마이크로어레이 실험에서 발행하는 데이터는 프로우브에 해당하는 유전자의 발현량을 나타내는 벡터 형태로 주어지고, 여러 실험을 하게 되면, 이들 벡터를 열로 하는 행렬로 표현된다. 마이크로어레이 데이터 분석에서 클러스터링 기법이 대표적으로 많이 사용된다. 행에 대응하는 유전자나 열에 대응하는 샘플에 대해서 클러스터링하게 된다. 이때 거리 정보를 이용하게 되는 데 전체 행이나 열에 있는 값을 이용하게 된다. 그런데 마이크로어레이 데이터 분석에서는 유사한 발현정도를 보이는 유전자 집단 및 샘플 집단을 추출하여 추가적인 분석을 하는 것이 유용한 경우들이 있다.

2. 기존 클러스터링 기법

거리기반 클러스터링 기법을 마이크로어레이 데이터에 적용하면 (그림 1)과 같은 형태로 결과가 얻어진다. 그런데 거리 기반 방법에서는 전체 차원에 대한 거리를 고려하기 때문에 그림에서 사각형으로 표시된 바와 같이 지역적으로 유사한 특성을 갖는 부분들이 분리되어 나타날 수 있다. 실제 분석에서는 이러한 유사한 부분을 추출하는 것이 유용한 정보를 제공하는 경우가 있다. 전체 차원을 고려하는 클러스터링 방법은 전체 유전자 또는 전체 샘플들에 대한 유사도의 관점에서 이들 개체들을 클러스터링하지만, 지역적인 클러스터를 효과적으로 찾아주기 어렵기 때문에 이를 위한 방법이 요구된다. 이 논문에서는 이를 지역적인 클러스터를 찾는 방법을 제안한다.

3. 제안한 지역 클러스터링 기법

먼저 샘플집단과 유전자 집단에 대해서 계층적 클러스터링을 수행한다. 입력 마이크로어레이 데이터 행렬은 M으로 나타내고, 클러스터링 결과로 정렬된 발현정도값 행렬은 E으로 나타낸다. 행렬에서 각 행은 하나의 유전자에 대응하여, 각 열은 하나의 샘플에 대응하고, 원소 e_{ij}는 i번째 유전자 g_i의 j번째 샘플에서 발현정도를 나타낸다.

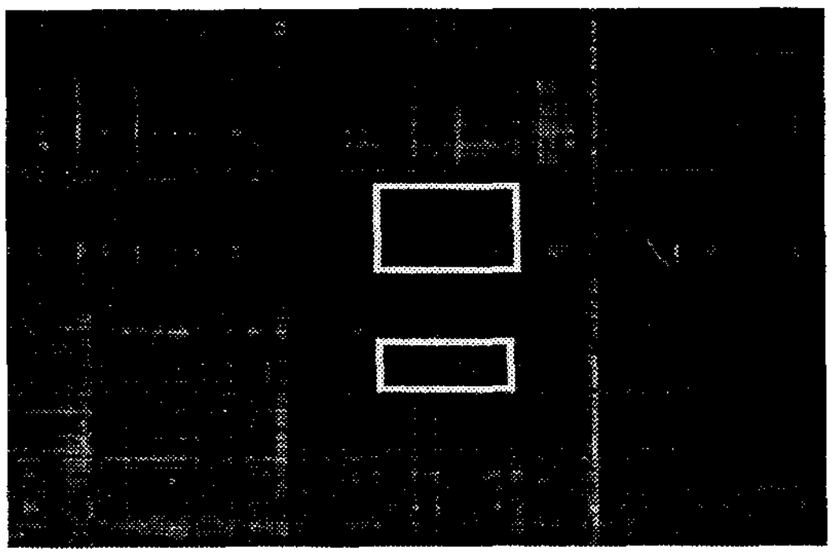

그림 1. 마이크로어레이 데이터의 클러스터링결과

$$E=\begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1m} \\ e_{21} & e_{22} & \cdots & e_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nm} \end{pmatrix}$$

행에 대응하는 유전자 이름은 $G=(g_1, g_2, \cdots, g_n)$로 나타내고, 열에 대응하는 샘플 이름은 $S=(s_1, s_2, \cdots, s_m)$으로 나타낸다. 편의상 유전자 i의 전체 샘플에 대한 발현정도는 $G_i=(e_{i1}, e_{i2}, \cdots, e_{im})$로 나타

내고, 샘플 j의 전체 유전자 집합에 대한 발현정도는 $S_j=(e_{1j},e_{2j},\cdots,e_{nj})^t$로 나타낸다.

클러스터링 결과 E로부터 관심 영역을 선정한다. 관심영역(interest region)은 비슷한 발현정도를 갖는 클러스터링 결과에서 사각형 영역으로서, 해당 부분과 유사한 유전자 및 샘플 집단을 군집화하기 위한 기준(reference)에 대한 정보를 제공하는 역할을 한다. 선택된 관심영역을 다음과 같은 부분 행렬$B_{[a,b:c,d]}$로 나타낸다.

$$B_{[a,b:c,d]}=\begin{pmatrix} e_{ac}\cdots e_{ad} \\ \vdots \ \ddots \ \vdots \\ e_{bc}\cdots e_{bd} \end{pmatrix}$$

선정된 관심영역에 대한 최소값(min), 1첫번 퀀타일(Q_1), 중간값(med), 3번째 퀀타일(Q_3), 최대값(max)를 계산한다.

$$\begin{aligned} \min &= \min\{e_{ij}|e_{ij}\in B_{a,b:c,d}\}, \\ Q_1 &= Q_1\{e_{ij}|e_{ij}\in B_{a,b:c,d}\}, \\ med &= med\{e_{ij}|e_{ij}\in B_{a,b:c,d}\}, \\ Q_3 &= Q_3\{e_{ij}|e_{ij}\in B_{a,b:c,d}\}, \\ \max &= \max\{e_{ij}|e_{ij}\in B_{a,b:c,d}\} \end{aligned}$$

관심영역의 중간값 med을 지역클러스터링의 기준값 m로 설정하고, 기준값에 대한 유사도 반경 r을 중간값에서 Q_1과 Q_2간의 거리의 최대값으로 설정한다.

$$m=med,\quad r=\max\{m-Q_1,Q_3-m\}$$

관심영역의 확장을 통해서 클러스터의 크기를 확장하기 위해서 다음과 같은 열(샘플)확장과 행(유전자)확장을 반복한다. 관심영역 $B_{[a,b:c,d]}$을 현재 지역클러스터 행렬 CS로 설정하고, CS와 중첩되는 유전자 집합을 $G_{cs}=(g_a,\cdots,g_b)$로, CS와 중첩되는 샘플 집합을 $S_{cs}=(s_c,\cdots,s_d)$라 나타낸다. 열확장을 위해서는 현재 CS에 속하지 않은 열 S_j들의 CS에 속하는 유전자 $(g_a,\cdots,g_b)$에 대응하는 발현정도값 $(e_{aj},\cdots,e_{bj})^t$들의 기준값으로 부터의 최대 거리를 구하여, 거리가 $r\cdot\tau$이내이면 해당 열을 후보 열(샘플) 집합 S_C에 추가한다. S_C은 공집합(ϕ)으로 초기화된 것이다.

$$\text{if } \max_{g_k\in G_{cs}}|e_{kj}-m|\le r\cdot\tau,\ S_C\leftarrow S_C\cup\{s_j\}$$

여기에서 τ는 사용자에 의해서 지정할 수 있는 식별하는 클러스터의 동질성의 정도를 결정하는 파라미터이다. τ을 크게 할수록 동질성이 작은 것들이 포함될 수 있는 여지가 커지게 된다.

행확장을 위해서는 현재 CS에 속하지 않은 행 G_i들의 CS에 속하는 샘플 $(s_c,\cdots,s_d)$에 대응하는 발현정도값 $(e_{ic},\cdots,e_{id})$들의 기준값으로 부터의 최대 거리를 구하여, 거리가 $r\cdot\tau$이내이면 해당 행을 후보 행의 집합 G_C에 추가한다. G_C은 공집합(ϕ)으로 초기화된 것이다.

$$\text{if } \max_{s_k\in S_{cs}}|e_{ik}-m|\le r\cdot\tau,\ G_C\leftarrow G_C\cup\{g_i\}$$

(그림 2)는 관심영역 $B_{[a,b:c,d]}$에 대해서 행확장을 통해 유전자의 집합 G_C가 추가되어 RE 영역이 확장 후보지역으로 편입되고, 열확장을 통해 샘플의 집합 S_C가 추가되어 CE 영역이 확장 후보지역으로 편입된 것을 보이고 있다. 그림에서 RE와 CE 영역의 원소의 값 e_{ij}는 모두 $|e_{ij}-m|\le r\cdot\tau$의 조건을 만족하지만, U 영역의 원소들은 이 조건을 만족한다는 보장이 없기 때문에 이 부분을 $B_{[a,b:c,d]}$에 유사한 것들이 $B_{[a,b:c,d]}$, RE, CE에 가까워지도록 행과 열을 재배치할 필요가 있다. 이를 위해 U에 대응하는 행렬 D를 다음과 같이 정의한다.

$$U=(e_{st})_{p\times q}\Rightarrow D=(d_{st})_{p\times q},\ \text{if}$$
$$|e_{st}-m|>r\cdot\tau,\ d_{st}=1,\ \text{otherwise}\ d_{st}=0.$$

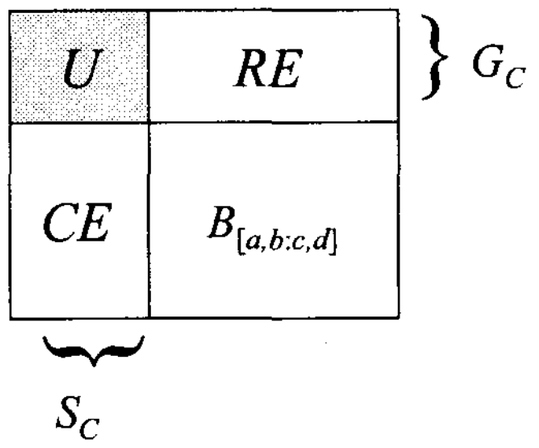

그림 2. 행확장 및 열확장의 결과

행렬 D의 각 행 s에 대해서 원소의 합 rc_s과, 각 열 t에 대한 원소의 합 cc_t를 계산한다.

$$rc_s=\sum_{t=1}^{q}d_{st}\ ,\qquad cc_t=\sum_{s=1}^{p}d_{st}$$

G_C에 속하는 행들을 rc_s값이 작아지는 순으로 정렬하고, S_C에 속하는 열들을 cc_t값이 작아지는 순으로 정렬한다. 이 과정을 거치고 나면, U 영역의 원소들은 오른쪽 하단으로 갈수록 $B_{[a,b:c,d]}$과 유사한 것들이 많아지고, 왼쪽 상단으로 갈수록 유사도가 떨어지는 것이 많아지게 되도록 정렬된다.

4. 결론

클러스터링은 유사한 특성을 갖는 개체를 군집으로 묶는 역할을 하는 것으로, 다차원 데이터가 주어지면 전체 차원에서 데이터간의 거리를 이용하여 클러스터링을 한다. 마이크로어레이 데이터 분석에서는 일부 차원만을 고려한 지역 클러스터를 찾는 것이 유용한 경우들이 있고, 이 논문에서는 효과적으로 지역 클러스터를 찾는 방법을 제안하였다. 향후 계층적 클러스터링을 제공하는 도구에 제안한 기법을 통합하여 생명과학자들을 위한 분석도구로 개발할 예정이다.

참 고 문 헌

[1] D. Nam, S.-Y. Kim, Gene-Set Approach for Expression Pattern Analysis, Briefing in Bioinformatics, Jan., 2008.

[2] S. Draghici,"Data Analysis Tools for DNA Microarrays," Chapman & Hall/CRC, 2003.

[3] D. W. Mount, "Bioinformatics: Sequence and Genome Analysis," Cold Spring Habor Lab Press, 2004.

[4] G. B. Forgel, D. W. Corne, "Evolutionary Computation in Bioinformatics," Morgan Kaufmann Publishers, 2003