Journal of the Korean Data & Information Science Society
2014, **25**(5), 999–1009

http://dx.doi.org/10.7465/jkdi.2014.25.5.999
한국데이터정보과학회지

상대적 계층적 군집 방법을 이용한 마이크로어레이 자료의 군집분석[†]

우숙영[1] · 이재원[2] · 전명식[3]

[123]고려대학교 통계학과

접수 2014년 5월 30일, 수정 2014년 7월 24일, 게재확정 2014년 8월 2일

요 약

계층적 군집 분석은 분석 결과를 덴드로그램으로 쉽게 표시할 수 있어서 방대한 양의 마이크로어레이 자료를 탐색하기에 유용하며, 군집된 결과를 이용하여 생물학적 현상을 이해하는데 도움을 준다. 하지만, 계층적 군집방법은 두 군집간의 절대값 거리만을 고려하여 병합하기 때문에 군집 간의 상대적 비유사성은 설명하지 못하는 단점이 있다. 본 연구에서는 상대적 계층적 군집 방법을 소개하고, 마이크로어레이 자료와 같이 다양한 군집의 모양을 가진 모의실험 자료들과 실제 마이크로어레이 자료를 사용하여 상대적 계층적 군집방법과 기존의 계층적 군집 방법을 비교하였다. 두 계층적 군집 방법의 질적 평가는 오분류율, 동질성, 이질성 지표를 이용하여 수행하였다.

주요용어: 계층적 군집, 동질성, 상대적 계층적 군집, 오분류율, 이질성.

1. 서론

분자 생물학에서 유전자 분석을 위한 획기적인 발전을 이끈 마이크로어레이 (microarray) 기술은 대량의 유전자의 발현 상황을 총체적으로 탐색할 수 있게 하였고, 생명체의 유전적 특징을 한 두 서열의 독립된 유전자에 의해서가 아닌 여러 유전자들 간의 유기적인 관계 하에서 이해할 수 있게 하였다 (Lee 등, 2012). 이런 과정에서 마이크로어레이를 통해 얻어지는 대용량의 자료들을 올바르게 분석하기 위한 방법이 필요하게 되었는데, 군집 분석을 통해서 어느 정도 가능하게 되었다 (Eisen 등, 1998; Ben-Dor 등, 1999; Chen 등, 2002; Speed, 2003; Yeo, 2011; Lim 등, 2012).

군집 분석의 여러 가지 방법 중에는 거리 중심의 접근 방법과 차원 축소 중심의 방법 등의 알고리즘이 있다. 거리 중심의 접근 방법으로는 계층적 군집화 (hierarchical clustering) 또는 K-평균 군집화 (K-means clustering) 등이 있으며, 차원 축소 중심의 방법으로는 주성분 분석 (principal component analysis) 등이 있다. 이들 방법 중에 계층적 군집 분석은 가까운 거리를 차례로 묶는 군집 방법이다. 이 방법은 군집의 형성에 위계가 있어서 일단 한 군집에 속하게 된 각각의 개체는 동시에 다른 군집에

[†] 이 논문은 2014년도 고려대학교에서 지원된 연구비로 수행된 연구이며, 또한 2012년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2012R1A1A2008686).

[1] (136-701) 서울 특별시 성북구 안암로 145, 고려대학교 통계학과, 박사과정.
[2] 교신저자: (136-701) 서울 특별시 성북구 안암로 145, 고려대학교 통계학과, 교수.
E-mail: jael@korea.ac.kr
[3] (136-701) 서울 특별시 성북구 안암로 145, 고려대학교 통계학과, 교수.

속할 가능성을 잃어버리게 되는 단점이 있다. 하지만, 유전자들 간의 상호관계를 덴드로그램 (dendrogram)으로 표현하기 때문에 자료의 성격을 직관적으로 이해할 수 있게 한다 (Eisen 등, 1998; Lee 등, 2011; Lim 등, 2012).

계층적 군집 방법은 크게 분할적 방법과 병합적 방법으로 나누어 볼 수 있는데 병합적 방법은 두 군집들 내 절대값 거리 (absolute distance)만을 고려하여 가장 가까운 군집들을 병합한다. 이것은 군집 내 응집성, 즉 같은 군집 내 유사성에 대한 정보를 알려 준다 (Lance 등, 1967; Rohlf, 1973; Hartigan, 1975). 그러나 이 방법은 군집 내 유사성의 측정에는 유용하지만, 군집간의 이질성은 고려하지 못하는 단점을 갖는다. 군집 간의 이질성은 군집 간의 상대적 비유사성을 설명하는 척도이다. 만일, 군집 내 유사성과 군집 간의 비유사성을 고려하는 방법들을 결합하여 자료를 군집화 하는데 이용한다면, 기존의계층적 군집화 방법보다 더 효율적으로 자료를 병합할 수 있을 것이다.

본 연구에서는 마이크로어레이 자료의 다양한 분포 양상을 가정한 모의실험 자료와 실제 백혈병 유발 유전자의 마이크로어레이 자료에 Mollineda와 Vidal (2000)이 제시한 상대적 계층적 군집 (relative hierarchical clustering) 방법을 적용한 후 기존의 계층적 군집화 방법과 비교하여 그 유용성을 살펴보고자 하였다. 논문의 구성은 2절에서 상대적 계층적 군집화 방법을 소개하고 3절에서는 군집분석 평가 척도를 살펴본 후, 4절에서 모의실험 자료를 활용하여 상대적 계층적 군집 방법과 기존의 계층적 군집방법을 비교하였다. 5절에서는 실제 자료 분석을 통하여 상대적 계층적 군집 방법의 수행 능력에 대해 살펴보았다. 마지막으로 6절에서 결론을 기술하였다.

2. 상대적 계층적 군집화 방법

두 개의 군집 i, j에 대한 거리 $d(i,j)$는 군집 i와 군집 j간의 거리를 나타내고, C는 현 군집 단계에서의 군집들의 집합, $|C|$는 C에 속한 군집의 개수, 그리고 $\bar{d}_{ij}$는 군집 i에서 군집 j를 제외한 나머지 군집들과의 평균거리로 정의한다. 이때, $d(i,j)$는 최소 거리 방법 (single linkage), 최대 거리 방법 (complete linkage), 평균 거리 방법 (average linkage) 중에 하나를 사용 한다 (Hartigan, 1975; Rohlf, 1973; Lance 등, 1967).

최소거리 방법은 두 군집 사이의 거리를 계산할 때에 유사함의 척도로서 각 군집에 속하는 임의의 두 개체간의 거리를 다 계산한 후에 그 중에서 최소 거리를 사용하는 방법이다. 최대 거리 방법은 각 군집에 속하는 임의의 두 개체간의 거리를 다 계산한 후에 그 중에서 최대 거리를 사용하고, 평균 거리 방법은 각 군집에 속하는 임의의 두 개체간의 거리를 다 계산한 후에 그 평균값을 거리로 이용한다. Mollineda와 Vidal (2000)이 제시한 상대적 비유사성 (relative dissimilarity)에 대한 $D(i,j)$척도는 공식 (2.1)에서와 같이 표현하며, 상대적 비유사성의 척도인 $D(i,j)$를 이용하여 $D(i,j)$가 최소가 되는 군집들을 계층적으로 병합하는 것을 상대적 계층적 군집이라고 정의한다.

$$D(i,j) = \frac{d(i,j)}{\min\{\bar{d}_{ij}, \bar{d}_{ji}\}} \tag{2.1}$$

$$\bar{d}_{ij} = \frac{\sum_{k \in C, k \neq j} d(i,k)}{|C| - 2}$$

상대적 비유사성의 척도인 $D(i,j)$은 두 군집간의 유사성을 설명하는 거리 $d(i,j)$정보뿐만 아니라 군집 j를 제외한 나머지 군집들 간에 거리를 고려하기 때문에 나머지 군집들 간에 상대적인 비유사성도 고려하는 방법이라고 할 수 있다. 이것은 두 군집 i와 j간의 유사성이 다른 나머지 군집들과의 유사성에 비해 얼마나 믿을 수 있는지 정보를 준다. Figure 2.1과 Figure 2.2는 상대적 계층적 군집화가 기존의 계층적 군집화와 어떻게 다른지 이해를 돕기 위한 예제로 제시하였다. 일차원 공간에 6개의 개체가 4.5,

4, 4.5, 4, 4.5 간격으로 놓여 있다고 가정하고 유사한 특성을 가진 군집끼리는 같은 색깔로 표시해 놓았다. 가장 왼쪽에 놓여 있는 두 네모가 서로 유사한 특성을 가지고 있는 군이고, 왼쪽에 있는 세 번째 원부터 여섯 번째 원까지가 비슷한 특성을 가지고 있는 군이다. 이에 대한 기존의 계층적 군집 방법과 상대적 계층적 군집 방법을 적용한 결과를 각각 아래 Figure 2.1, Figure 2.2에서 덴드로그램으로 제시 하였다. 두 군집간 거리 $d(i,j)$는 유클리디안 거리 (Euclidean distance) 공식으로 계산하였고, 최대 거리 방법을 이용하였다.

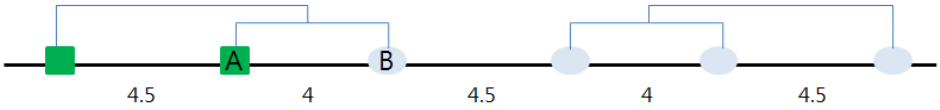

Figure 2.1 Dendrogram by hierarchical clustering

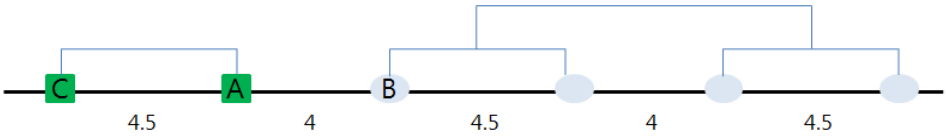

Figure 2.2 Dendrogram by relative hierarchical clustering

Figure 2.1는 기존의 계층적 군집화 방법을 적용했을 때, 네모 A와 원 B는 본래는 서로 유사한 특성이 없는 다른 군에 속해 있음에도 불구하고 같은 군집으로 묶이는 것을 보였다. 하지만, 상대적 계층적 군집화 방법을 적용하면 Figure 2.2에서 볼 수 있듯이 네모 A와 네모 C가 같은 군으로 병합하는 것을 볼 수 있다. 이는 네모 A와 원 B간에 상대적 비유사성 거리, 즉 D(A, B)는 공식 (2.1)을 통해 0.46으로 계산되는 반면에 D(A, C)는 0.43으로 계산되기 때문인데, 네모 A와 원 B간의 거리만을 이용하여 병합한 것이 아니라 두 군집들을 병합하기 전에 나머지 군집들 간에 거리를 고려하여 네모 A와 원 B간의 유사성이 다른 나머지 군집들 간의 유사성에 비해 얼마나 가까운지를 상대적으로 고려하였기 때문이다. 상대적 계층적 군집 방법은 군집내의 유사성뿐만 아니라 나머지 군집들 간의 상대적 비유사성을 동시에 고려하므로, 다른 군집에 속한 네모 A와 원 B가 같은 군으로 병합하게 되는 현상을 막을 수 있다.

3. 군집 분석 평가 척도

마이크로어레이 자료 분석에서 쓰이는 자료 행렬 $G = (g_{ij})_{n\times p}$, $i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, p$에서 관측치 g_{ij}는 i번째 샘플의 j번째 유전자 정보를 갖는다. 이때 샘플에 대하여 수행하는 군집 분석은 같은 군집 내에 속한 샘플 간의 유사성은 높고 서로 다른 군집에 속하는 샘플 간에는 유사성을 작게 하는 것을 목표로 한다. 유사한 특성을 보이는 샘플들을 함께 묶어 군집분석을 수행한 후, 군집 분석의 결과에 대한 타당성 평가는 오분류율(PIGP; percentage of incorrectly grouped points), 동질성 (homogeneity) 그리고 이질성 (separation)을 이용하여 평가 할 수 있다.

오분류율 (PIGP; percentage of incorrectly grouped points)

상대적 계층적 군집 방법과 기존의 계층적 군집 방법을 적용한 각각의 결과들은 군집 오류 매트릭스 (GCM; grouping confusion matrix)를 가지고 오분류율을 계산하여 평가하였다. 오분류율은 마이크로어레이 자료 분석에서 군집 i에 있어야 하는 샘플이 군집 j로 얼마나 많이 병합 되었는지를 나타내는 지표로 식은 다음과 같다 (Mollineda, 2000).

$$\text{오분류율(PIGP; \%)} = [\text{오분류된 샘플 수}/\text{전체 샘플 수}] * 100 \tag{3.1}$$

동질성 (homogeneity)

동질성은 한 군집에 속한 샘플간의 유사성을 평가하는 지표로 각 샘플과 샘플이 속한 군집 간의 중심 사이의 평균거리를 계산하여 평가한다. 식은 다음과 같다 (Datta, 2003).

$$H = \frac{1}{N}\sum_{i} D(g_i, C(g_i)) \tag{3.2}$$

여기서 g_i는 i번째 샘플, $C(g_i)$는 g_i가 속한 군집의 중심, N은 샘플의 총 수, 그리고 D는 상대적 계층적 군집방법 혹은 기존의 계층적 군집방법에 해당하는 거리함수이다. H값이 작을수록 군집이 잘 되었다고 평가한다. 이를 통해 같은 군집 안에 있는 샘플들 간의 근접성 혹은 유사성을 평가하여 응집성에 관한 정보를 확인할 수 있다.

이질성 (separation)

이질성은 군집 중심들 간의 가중 평균 거리로 계산하며, 식은 다음과 같다 (Datta, 2003).

$$S = \frac{1}{\sum\limits_{i\neq j} N_{ci}N_{cj}} \sum_{i} N_{ci}N_{cj}D(C_i.C_j) \tag{3.3}$$

C_i와 C_j는 i번째와 j번째 군집의 중심이고, N_{c_i}와 N_{c_j}는 i번째와 j번째의 군집에 있는 샘플의 개수이다. S값이 클수록 군집이 잘 되었다고 평가하고, 이를 통해 다른 군집들 간의 비유사성에 대한 정보를 얻을 수 있다.

4. 모의 실험 자료 분석

마이크로어레이 자료와 같이 다양한 군집 패턴들을 가정하여 생성된 모의실험 자료를 가지고 상대적 계층적 군집방법을 기존의 계층적 군집방법과 비교하여 평가하였다. 모의실험에 사용된 자료들은 정규분포를 따르게 하였고, 각 군집마다 유전자 발현 정도값 x, y를 갖는 200개의 mRNA 샘플로 가정하였다. 각 군집에 속하는 임의의 두 샘플간의 거리는 유전자 발현정도 값을 가지고 유클리디안 거리 공식을 이용하여 계산하였고, 최대거리 방법을 적용하였다. 첫 번째 자료는 세 군집의 모양이 둥근 형태의 분포 (round-shaped distribution)를 이루도록 생성한 것이며, 군집들의 중심을 연결하면 삼각형 모양을 띠는 자료이다. 두 번째 자료는 군집들이 가로로 긴 형태 (longish-shaped distribution)를 보이도록 하였고, 세 번째 자료는 첫 번째 자료와 같이 세 군집의 모습이 둥근 형태를 갖지만, 각 군집의 중심을 연결하면 일직선이 되는 자료 형태가 되도록 하였다. 네 번째 자료는 각 군집의 중심을 연결하면 일직선이 되지만, 가운데에 위치한 군집은 가로로 긴 형태의 분포를 보이며 양쪽에 있는 다른 두 군집은 둥근 형태를 보이도록 하였다. 두 계층적 군집 방법의 질적 평가는 오분류율 척도로 비교하였다.

4.1. 세 군집의 모양이 모두 둥근 분포를 보이는 자료

세 군집의 모양이 모두 공 모양으로 둥글고, 군집의 중심을 연결하면 삼각형 모양을 띠는 자료를 생성하기 위해, 각 군집별로 Table 4.1과 같은 평균과 표준편차를 사용하였다. 각 군집의 샘플들은 이차원 공간 내에서 정규분포로부터 무작위로 생성되었고, 각 군집들은 통계적으로 서로 독립적이다. 이렇게 생성된 자료의 각 군집에 따른 산점도를 Figure 4.1에 제시하였다.

Table 4.1 Means, standard deviations in data sets with round-shaped distributions

class	mean ($\bar{x}$)	mean ($\bar{y}$)	standard deviation ($\bar{\sigma}_x$)	standard deviation ($\bar{\sigma}_y$)
class1 (N=200)	-2	2	0.2	0.2
class2 (N=200)	0	-2	0.2	0.2
class3 (N=200)	2	2	0.2	0.2

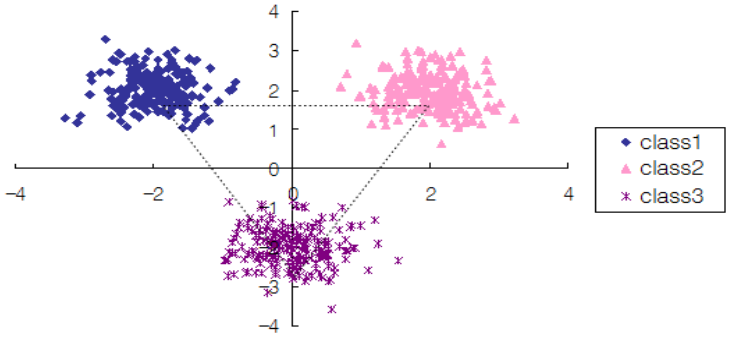

Figure 4.1 Distribution of classes in round-shaped data

Table 4.2에서 보면 두 군집방법을 적용한 경우에 오분류율이 모두 0%였다. 군집 모양이 둥글고, 이상치도 보이지 않은 경우에서는 상대적 계층적 군집화와 기존의 계층적 군집화 방법 모두 만족할 만한 결과를 보이는 것을 알 수 있다.

Table 4.2 Comparison between hierarchical clustering and relative hierarchical clustering in round-shaped data

hierarchical clustering						relative hierarchical clustering					
PIGP	GCM					PIGP	GCM				
	before clustering						before clustering				
			class1	class2	class3				class1	class2	class3
0%	after	class1	200	0	0	0%	after	class1	200	0	0
		class2	0	200	0			class2	0	200	0
		class3	0	0	200			class3	0	0	200

4.2. 두 군집의 모양이 길쭉한 형태를 갖는 자료

두 군집의 모양이 가로로 긴 모양의 분포를 보이는 자료를 생성하기 위해서 각 군집 별로 Table 4.3과 같은 평균과 표준편차를 사용하였다. 각 군집의 샘플들은 이차원 공간 내에서 정규분포를 통해 무작위로 생성하였고, 각 군집들은 통계적으로 독립적이다. Figure 4.3은 이렇게 생성된 자료로 표시한 각 군집별 산점도이다.

Table 4.3 Means, standard deviations in data sets with longish-shaped distributions

class	mean($\bar{x}$)	mean($\bar{y}$)	standard deviation($\bar{\sigma}_x$)	standard deviation($\bar{\sigma}_y$)
class1 (N=200)	4	5	2	0.3
class2 (N=200)	4	2	2	0.3

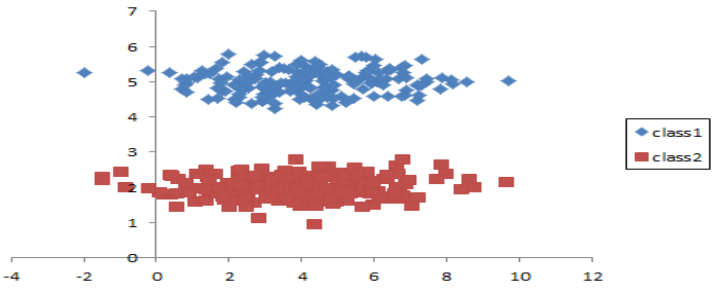

Figure 4.2 Distribution of classes in longish-shaped data

Table 4.4를 보면 두 군집방법을 적용한 경우 상대적 계층적 군집방법을 이용한 경우는 오분류율이 35.5%였고, 계층적 군집방법을 이용한 경우는 오분류율이 44.3%였다. 긴 형태의 분포를 보이는 자료에서는 상대적 계층적 군집 방법을 이용할 경우에 기존의 계층적 군집 방법을 이용할 때 보다 더 바람직한 군집화가 된 것으로 평가하였다.

Table 4.4 Comparison between hierarchical clustering and relative hierarchical clustering in longish-shape data

hierarchical clustering					relative hierarchical clustering				
PIGP	GCM				PIGP	GCM			
44.3%		before clustering			35.5%		before clustering		
	after		class1	class2		after		class1	class2
		class1	74	51			class1	77	19
		class2	126	149			class2	123	181

4.3. 세 군집의 모양이 둥근 형태를 띠며, 군집의 중심이 일렬로 위치한 경우

각 군집 별로 Table 4.5와 같은 평균과 표준편차를 사용하여 세 군집의 모습이 모두 둥근 형태를 갖지만, 각 군집의 중심을 연결하면 일직선이 되는 자료를 생성하였다. 각 군집의 샘플들은 이차원 공간 내에서 정규분포를 통해 무작위로 생성하였고, 각 군집들은 통계적으로 독립적이다. Figure 4.3은 이에 대한 각 군집별 산점도이다.

Table 4.5 Means, standard deviations in round-shaped data with the same mean

class	mean ($\bar{x}$)	mean ($\bar{y}$)	standard deviation ($\bar{\sigma}_x$)	standard deviation ($\bar{\sigma}_y$)
class1 (N=200)	-2	2	0.3	0.2
class2 (N=200)	0	2	0.3	0.2
class3 (N=200)	2	2	0.3	0.2

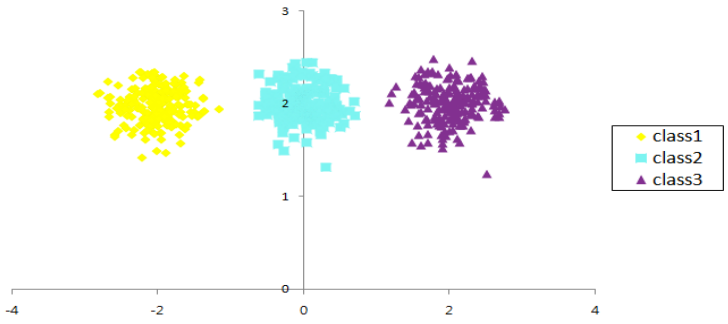

Figure 4.3 Distribution of classes in round-shaped data with the same mean

Table 4.6 Comparison between hierarchical clustering and relative hierarchical clustering in round-shaped data with the same mean

hierarchical clustering						relative hierarchical clustering					
PIGP	GCM					PIGP	GCM				
0.5%			before clustering			0.2%			before clustering		
	after		class1	class2	class3		after		class1	class2	class3
		class1	199	2	0			class1	199	0	0
		class2	1	198	0			class2	1	200	0
		class3	0	0	200			class3	0	0	200

Table 4.6에서 제시한 바와 같이 상대적 계층적 군집방법을 이용한 경우 오분류율이 0.2%였고, 기존의 계층적 군집방법을 이용한 경우에는 0.5%의 오분류율을 보여, 세 군집의 모양이 모두 둥글고 군집의 중심이 일렬로 위치한 자료에서는 두 계층적 방법 모두에서 오분류율의 크기도 작았다.

4.4. 세 군집 중에 한 군집만이 길쭉한 자료인 경우

각 군집의 중심을 연결하면 일직선이지만 가운데에 위치한 군집은 가로로 길쭉한 형태를 띠며 양쪽에 있는 다른 두 군집은 둥근 모양을 한 자료를 생성하기 위해서 각 군집 별로 Table 4.7과 같은 평균과 표준편차를 사용하였다. 각 군집의 샘플들은 이차원 공간 내에서 정규분포를 통하여 무작위로 생성되었고, 각 군집들 간에는 통계적으로 독립적이었다. Figure 4.4은 각 군집별로 그린 산점도이다.

Table 4.7 Means, standard deviations in data sets with round and longish-shaped distributions

class	mean ($\bar{x}$)	mean ($\bar{y}$)	standard deviation ($\bar{\sigma}_x$)	standard deviation ($\bar{\sigma}_y$)
class1 (N=200)	-2	2	0.2	0.2
class2 (N=200)	0	2	0.6	0.1
class3 (N=200)	2	2	0.2	0.2

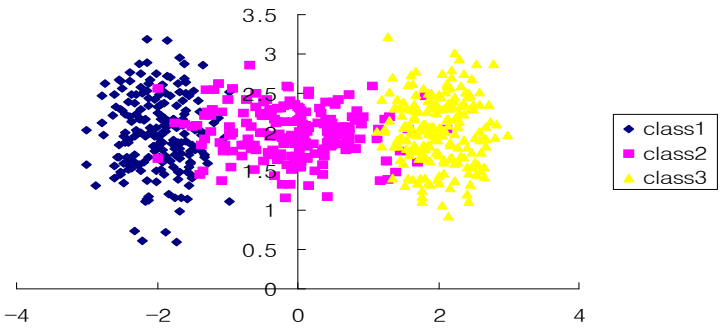

Figure 4.4 Distribution of classes in round and longish-shaped data

Table 4.8에서 보듯이 상대적 계층적 군집방법을 이용한 경우에 오분류율은 8.67%였고, 기존의 계층적 군집방법을 시행한 경우에는 오분류율이 11.33%였다. 비록 class1을 class2로 오분류한 비율은 계층적 군집방법을 사용한 것 보다 상대적 계층적 군집 방법을 사용한 것이 높을지라도, 세 군집 중에 한 군집만이 길쭉한 자료 분포를 가진 경우에는 길쭉한 모양의 군집인 class2에 속한 데이터들이 상대적 계층적 군집방법을 적용한 결과에서 더 잘 묶인 것을 볼 수 있다.

Table 4.8 Comparison between hierarchical clustering and relative hierarchical clustering in round and longish-shaped data

hierarchical clustering						relative hierarchical clustering					
PIGP	GCM					PIGP	GCM				
	before clustering						before clustering				
			class1	class2	class3				class1	class2	class3
11.33%	after	class1	195	16	0	8.67%	after	class1	183	6	0
		class2	5	169	30			class2	17	187	22
		class3	0	15	170			class3	0	7	178

이와 같이 모의실험을 위해 생성한 자료를 분석한 결과에서 세 군집의 모양이 모두 둥근 모습을 갖고, 이상치도 없으며, 군집의 중심을 연결하면 삼각형 모양을 띠는 분포를 보이는 자료에서는 두 군집 방법

모두 오분류율이 0%로 좋은 결과를 보여주었다. 하지만, 군집 모양이 긴 모양의 분포를 보이는 경우에는 상대적 계층적 군집화 방법이 기존의 계층적 군집방법 보다 본래의 군집 모습으로 잘 병합하고 있음을 알 수 있었다. 결과를 제시하지 않았으나 평균 거리 연결법을 이용하여도 유사한 결과를 보여주었다.

5. 실제자료 분석

마이크로어레이 자료는 집단의 수 혹은 집단 구조에 대한 가정이 없으며, 오직 개체들 사이의 유사성에 의해 군집을 형성하기 때문에 마이크로어레이 자료의 올바른 통계학적인 분석을 위해서는 형성된 군집의 특성을 파악하여 군집들 사이의 관계를 파악할 수 있는 군집 분석이 유용하다 (Ben-Dor 등, 1999; Chen 등, 2002; Eisen 등, 1998; Speed, 2003). 본 연구에서 사용한 실제자료는 백혈병 (leukemia)을 유발하는 유전자의 실제 마이크로어레이 자료를 이용하였다. 38개의 B세포 급성 림프구성 백혈병 (B-cell ALL; B-cell acute lymphoblastic leukemia) 샘플, 9개의 T세포 급성 림프구성 백혈병 (T-cell ALL) 샘플, 25개의 급성 골수성 백혈병 (AML; acute myeloid leukemia) 샘플자료로 이루어진 3571개의 유전자 발현자료 (gene expression data; http://www.genome.wi.mit.edu/MPR)를 활용하였고, 각 유전자들의 측정 척도가 동일하기 때문에 분산이 1이 되기 위해 표준화할 필요는 없었다. 결측치에 대한 처리는 가장 가까운 k개의 이웃을 택한 후, 이들 k개의 관찰치들을 사용하여 결측치를 추정하는 방법인 k-최근접이웃 (k-nearest neighbor) 알고리즘 (k=5)을 사용하였다. 상대적 계층적 군집 방법과 기존의 계층적 군집 방법을 적용한 결과를 오분류율, 동질성, 이질성 지표로 비교하였다. 군집 간 거리는 유클리디안 공식을 사용하였고, 최대 연결방법을 적용하였다.

Figure 5.1은 PCA biplot을 사용하여 백혈병 자료를 이차원과 삼차원 그래프를 보여준 것으로 이차원 그래프에서는 급성 림프구성 백혈병 (ALL)과 급성 골수성 백혈병 (AML)이 잘 구분되어져 있음을 색깔을 통해 쉽게 확인할 수 있고, 급성 림프구성 백혈병 (ALL)과 급성 골수성 백혈병 (AML)은 서로 군집의 분포 양상이 길쭉한 모양을 띄고 있음을 알 수 있다. 동일한 백혈병 자료를 사용하여 상대적 계층적 군집화와 기존의 계층적 군집화 방법으로 분석한 결과를 오분류율, 군집 오류 매트릭스, 동질성과 이질성 등의 지표로 평가한 결과를 Table 5.1과 Table 5.2에 제시하였다.

Table 5.1 Comparison between hierarchical clustering and relative hierarchical clustering using homogeneity and separation

	hierarchical clustering	relative hierarchical clustering
homogeneity	6.60	2.94
separation	0.14	0.16

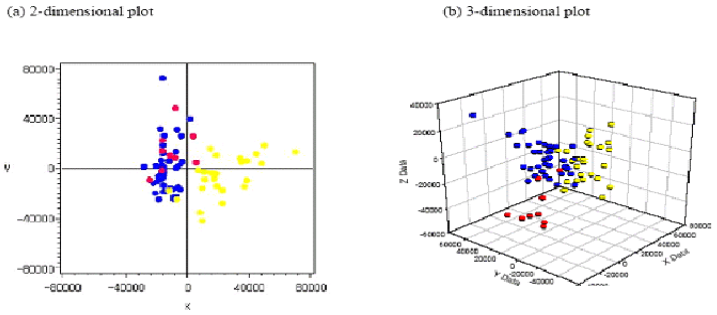

Figure 5.1 PCA biplot (row plot) of leukemia gene expression data set. 2-dimensional (a) and 3-dimensional (b) plots. B-cell ALL: blue dots; T cell ALL: red dots; AML: yellow dots

Table 5.2 Comparison between hierarchical clustering and relative hierarchical clustering using PIGP and GCM in leukemia data

hierarchical clustering						relative hierarchical clustering					
PIGP	GCM					PIGP	GCM				
36%	before clustering					23.6%	before clustering				
	after		B-cell	T-ALL	AML		after		B-cell	T-ALL	AML
		B-cell	37	7	18			B-cell	30	0	0
		T-ALL	1	2	0			T-ALL	7	8	9
		AML	0	0	7			AML	1	1	16

Table 5.1은 상대적 계층적 군집방법에서 기존의 계층적 군집방법 보다 동질성 값이 작고 군집이 잘 형성되었음을 보여주고 있다. 이질성 평가에서는 상대적 계층적 군집방법에서 기존의 계층적 군집방법의 경우 보다 약간 큰 값을 보이나 차이가 크지 않았다. Table 5.2는 두 군집 방법에 대한 오분류율을 비교한 결과로 상대적 계층 적 군집방법에서는 오분류율이 23.6%, 기존의 계층적 군집방법에서는 36%의 오분류율을 보여 상대적 계층적 군집 방법을 적용했을 경우에 계층적 군집방법보다 오분류율이 낮은 것을 확인할 수 있었다. 특히, 길쭉한 모양의 분포를 보이는 급성골수성백혈병 (AML)의 경우에 모의실험 자료 분석에서와 마찬가지로 상대적 계층적 군집방법이 기존의 계층적 군집방법 보다 군집이 잘 이루어졌다.

6. 결론

본 연구에서는 다양한 군집의 모양을 가진 모의생성 자료들과 백혈병 유발 유전자에 대한 마이크로어레이 자료를 실제 분석 자료로 활용하여 상대적 계층적 군집방법과 기존의 계층적 군집방법을 통한 결과를 비교 분석하였다. 이를 위해 다양한 분포 형태를 가정한 모의자료를 통해 어떤 조건에서 상대적 계층적 군집 방법이 기존의 계층적 군집 방법보다 수행 능력이 좋을지 오분류율, 동질성, 이질성 등의 질적 평가 지표를 사용하여 평가하였다. 그 결과 군집의 형태가 둥근 분포를 보이는 경우에는 두 군집화 방법 모두 만족할 만한 결과를 보여 주었으나, 자료의 분포 형태에서 길쭉한 모양의 분포를 갖는 군집이 포함되어 있을 경우에는 상대적 계층적 군집 방법이 기존의 계층적 군집 방법보다 오분류율과 동질성 측면에서 만족스러운 결과를 보였다.

실제 연구에서 다루게 되는 자료의 군집 분포 형태는 본 모의실험 및 실제 자료에서 살펴본 군집 형태보다 훨씬 다양하다. 상대적 계층적 군집 방법의 상대적 비유사성이 유사성과 상반되게 측정되는 경우는 없는지에 대한 고찰이 추후 필요하며, 복잡한 형태의 다양한 자료에서 두 군집화 방법을 비교하는 연구를 계속 진행해야 한다. 또한, 군집간의 거리를 유클리디언 거리가 아닌 마할라노비스 거리, 민코프스키 거리 등을 가지고 최대 연결 방법 외에 최소 거리 방법, 평균 거리 방법을 적용하여, 군집들이 어떤 형태일 때 어떤 거리 방법과 연결 방법을 이용하여 군집 방법을 적용하면 보다 효율적인 결과를 이끌어내는지 살펴보는 후속 연구가 필요하다. 이와 같은 추후 연구를 통해서 자료의 형태에 따라 가장 적절한 군집 방법을 선택하는 가이드라인을 제시할 수 있다면 마이크로어레이 자료와 같이 복잡한 연구 자료를 올바르게 탐색함에 있어 유용한 정보를 제공할 것이다.

References

Ben-Dor, A., Shamir, R. and Yakhini, Z. (1999). Clustering gene expression patterns. *Journal of Computational Biology*, **6**, 281-297.

Chen, G., Jaradat S. A., Banerjee, N., Tanaka T. S., Ko M. S. H. and Zhang, M. Q. (2002) Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. *Statistica Sinica*, **12**, 241-262.

Datta, S. and Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, **19**, 459-466.

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

Hartigan, J. A. (1975). *Clustering algorithms*, Wiley, New York.

Lance, G. N. and Williams W. T. (1967). A General theory of classificatory sorting strategies: 1. Hierarchical system. *Computer Journal*, **9**, 373-380.

Lee, S. H. and Lee, K. H. (2012). Detecting survival related gene sets in microarray analysis. *Journal of the Korean Data & Information Science Society*, **23**, 1-11.

Lim, J. S. and Lim, D. H. (2012). Comparison of clustering methods of microarray gene expression data. *Journal of the Korean Data & Information Science Society*, **23**, 39-51.

Mollineda, R. A. and Vidal E. (2000). *Pattern recognition and applications*, IOS Press, Amsterdam.

Rohlf, F. J. (1973). Hierarchical clustering using the minimum panning tree. *Computer Journal*, **16**, 93-95.

Speed, T. (2003). *Statistical analysis of gene expression microarray data*, CRC Press, Boca Raton, Florida.

Yeo, I. (2011). Clustering analysis of Korea's meteorological data. *Journal of the Korean Data & Information Science Society*, **22**, 941-949

Journal of the Korean Data & Information Science Society
2014, **25**(5), 999–1009

http://dx.doi.org/10.7465/jkdi.2014.25.5.999
한국데이터정보과학회지

Microarray data analysis using relative hierarchical clustering[†]

Sook Young Woo[1] · Jae Won Lee[2] · Myoungshic Jhun[3]

[123]Department of Statistics, Korea University

Received 30 May 2014, revised 24 July 2014, accepted 2 August 2014

Abstract

Hierarchical clustering analysis helps easily exploring massive microarray data and understanding biological phenomena with dendrogram. But, because hierarchical clustering algorithms only consider the absolute similarity, it is difficult to illustrate a relative dissimilarity, which consider not only the distance between a pair of clusters, but also how distant are they from the rest of the clusters. In this study, we introduced the relative hierarchical clustering method proposed by Mollineda and Vidal (2000) and compared hierarchical clustering method and relative hierarchical method using the simulated data and the real data in the various situations. The evaluation of the quality of two hierarchical methods was performed using percentage of incorrectly grouped points (PIGP), homogeneity and separation.

Keywords: Hierarchical clustering, homogeneity, percentage of incorrectly grouped points, relative hierarchical clustering, separation.

[†] This research was supported by Korea University Grant in 2014. It was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2008686).
[1] Ph.D candidate, Department of Statistics, Korea University, Seoul 136-701, Korea.
[2] Corresponding author: Professor, Department of Statistics, Korea University, Seoul 136-701, Korea. E-mail: jael@korea.ac.kr
[3] Professor, Department of Statistics, Korea University, Seoul 136-701, Korea.