*Journal of the Korean Data & Information Science Society*
*2012,* **23**(1), 39–51

한국데이터정보과학회지

# 마이크로어레이 유전자 발현 자료에 대한 군집 방법 비교[†]

임진수[1] · 임동훈[2]

[1]부산대학교 생명과학과 · [2]경상대학교 정보통계학과

접수 2011년 10월 30일, 수정 2011년 11월 17일, 게재확정 2011년 12월 12일

**요 약**

군집분석은 마이크로어레이 발현자료에서 유전자 혹은 표본들의 유사한 특성을 갖는 연관구조를 조사하는데 중요한 도구이다. 본 논문에서는 마이크로어레이 자료에서 계층적 군집방법, K-평균법, PAM (partitioning around medoids), SOM (self-organizing maps) 그리고 모형기반 군집방법 들의 성능을 3가지 군집 타당성 측도인 내적 측도, 안정적 측도 그리고 생물학적 측도를 가지고 비교분석하고자 한다. 모의실험을 통해 생성된 자료와 실제 SRBCT (small round blue cell tumor) 자료를 가지고 여러 가지 군집방법들의 성능을 비교하였으며 그 결과 모의실험 자료에서는 거의 모든 방법들이 3가지 군집측도에서 원래 자료와 일치하는 좋은 군집 결과를 나타내었고 SRBCT 자료에서는 모의실험 자료처럼 명확한 군집화 결과를 보여주지는 않으나 내적측도의 실루엣 너비 (Silhouette width) 관점에서는 PAM 방법, SOM, 모형기반 군집방법 그리고 생물학적 측도에서는 PAM 방법과 모형기반 군집방법이 모의실험 결과와 비슷한 결과를 얻었고 안정적 측도에서 모형기반 군집방법이 다른 방법들보다 좋은 군집결과를 보여주었다.

주요용어: 군집분석, 내적측도, 마이크로어레이, 생물학적 측도, 안정적 측도.

## 1. 서론

마이크로어레이 (microarray)란 슬라이드 유리와 같은 작은 고형체 위에 수천 혹은 수만 개의 DNA를 바둑판 격자처럼 배열해 놓은 칩으로 이 기술은 대량의 유전자 기능을 동시에 밝히는 기능유전체학 (functional genomics)의 중요한 도구이다 (황진수와 김지연, 2009; Deshmukh와 Purohit, 2007).

군집분석 (cluster analysis)은 탐색적 자료분석 방법으로 유사성을 갖는 자료끼리 서로 묶어서 군집을 형성해 나가는 분석방법으로 마이크로어레이 자료에서 비슷한 발현형태를 보이는 유전자나 표본끼리 군집을 형성함으로서 알려지지 않은 유전자의 기능을 유추할 수 있을 뿐만 아니라 유사한 특성을 보이는 표본들의 집단을 찾을 수 있다 (김재희와 고윤실, 2009; 여인권, 2011; 주용성 등, 2009).

본 논문에서는 마이크로어레이 자료에 대해 여러 가지 군집 방법의 성능을 비교하고 군집 타당성 측도를 통해 군집 결과를 평가하고자 한다. 본 논문에서 사용하는 SRBCT (small round blue cell tumor) 자료 (Khan 등, 2001)는 아이들에게 흔히 발생하는 악성 종양 자료로서 총 63개의 표본과 2308개의 유전자 속성을 가진 EWS (Ewing family of tumors), RMS (rhabdomyosarcoma), NB (neuroblasoma),

[†] 이 논문은 2011년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No.2011-0010089).
[1] (609-735) 부산광역시 금정구 장전동 산 30번지, 부산대학교 생명과학과, 학부생.
[2] 교신저자: (660-701) 경남 진주시 가좌동 900번지, 경상대학교 정보통계학과 및 RINS, 교수.
E-mail:dhlim@gnu.ac.kr

BL (Burkitt lymphoma)의 4종류의 암에 대해 조사한 자료이다. 본 논문에서는 계층적 군집방법 (hierarchical clustering; Eisen 등, 1998)과 비계층적 군집방법 (non-hierarchical clustering)으로 K-평균법(K-means clustering; Hartigan와 Wong, 1979), PAM (partitioning around medoids; Kaufman와 Rousseeuw, 1990) 그리고 SOM (self-organizing maps; Kohonen, 1997) 그리고 모형기반 군집방법 (model-based clustering; Fraley와 Raftery, 2002)들이 사용되고 군집결과의 타당성을 재는 측도로는 내적 측도 (internal measure)로 연결성 (connectivity; Handl 등, 2005), Dunn 지수 (Dunn, 1974), 실루엣 너비 (silhouette width; Rousseeuw, 1987)와 안정적 측도 (stability measure; Datta와 Datta, 2003; Yeung 등, 2001a)로 APN (average proportion of non-overlap), AD (average distance), ADM (average distance between means), FOM (figure of merit) 그리고 생물학적 측도 (biological measure; Datta와 Datta, 2006)로 BHI (biological homogeneity index)와 BSI (biological stability index)가 사용된다. 지금까지 마이크로어레이 자료에 대한 연구는 주로 일부의 군집방법과 일부의 군집 측도에 극한하여 제한적으로 이루어져 왔다 (정윤경과 백장선, 2007; 이경아와 김재희, 2011; He 등, 2006; Khan 등, 2001; Liu와 Ringner, 2004).

본 논문은 다음과 같이 구성되어 있다. 제 2 절에서는 여러 가지 군집 방법 들에 대해 간략하게 살펴보고 제 3 절에서는 군집 타당성 측도에 대해 살펴보고자 한다. 제 4 절에서는 모의실험을 통해 생성된 자료와 SRBCT 마이크로어레이 자료에 대해 군집방법들의 성능을 비교 평가하고 제 5 절에서 결론을 맺고자 한다.

## 2. 군집 방법

본 절에서는 계층적 군집방법, K-평균법, PAM 방법, SOM 방법과 모형기반 군집 방법에 대해 간략하게 살펴보고자 한다.

### 2.1. 계층적 군집 방법

계층적 군집 방법은 $N$개의 개체를 계층적인 형태로 군집을 형성해가는 방법으로 군집간의 거리 정의에 따라 최단연결법, 최장연결법, 평균연결법 등이 있다 (Eisen 등, 1998). 계층적 군집 방법은 군집 형성 방법에 따라 병합적 (agglomerative) 방법과 분할적 (divisive) 방법으로 나눈다. 병합적 방법은 가까운 개체들끼리 묶어감으로써 군집을 만들어가는 방법이고 분할적 방법은 반대로 먼 개체들을 나누어가는 방법이다.

계층적 군집 방법의 단점은 한 번 군집화된 개체는 다시 재배치가 불가능하다는 것이다.

### 2.2. K-평균법

K-평균법은 $N$개의 개체를 $K$개의 그룹으로 분할하는데 다음과 같이 목적함수가 최소가 되도록 분할하는 방법이다 (Hartigan와 Wong, 1979).

$$E = \sum_{i=1}^{K} \sum_{x \in C_i} d(x, m_i)$$

여기서 $m_i$는 군집 $C_i$의 중심이고 $d(x, m_i)$는 개체 $x$와 $m_i$와의 유클리디안 거리이다.

이 방법은 계층적 군집방법과는 달리 한 개체가 속해있던 군집에서 다른 군집으로 재배치가 가능하고 계산량도 적은 편이어서 유전자 자료와 같은 대량의 개체들에 대한 군집방법으로 유용하다.

### 2.3. PAM 방법

PAM 방법은 K-평균법과 유사하나 K-평균법은 군집의 중심값이 평균인데 반해 PAM은 군집 내 상호거리의 값이 가장 작은 것이 군집의 중심값이 된다. PAM은 K-평균법보다 계산량이 많아 자료수가 적은 경우 적합하고 또한 이상값 (outlier)과 결측값에 대해 로버스트한 방법이다 (Kaufman와 Rousseeuw, 1990).

### 2.4. SOM 방법

SOM 방법은 고차원의 데이터를 저차원의 기하학적 관계를 갖는 형태로 변환시켜 분석하는 방법이다. 즉, SOM은 데이터 공간을 입력층으로 하고 2차원 격차 맵을 출력층으로 하는 일종의 전방향 신경망 (feedforward neural network) 일종이다 (Kohonen, 1997). 입력된 데이터 개수 만큼의 뉴런 (neuron)을 가지고 있는 입력층은 연결강도벡터를 통해 출력층에 연결되어있으며 학습과정을 통해 갱신한다. 여기서 학습은 입력된 데이터들이 출력층에 투영되면서 출력층에 존재하는 뉴런들은 정해진 학습 알고리즘의 규칙적인 유사성 정도를 두고 서로 경쟁을 한다. 그 결과 승자 뉴런이 결정되면 결정된 승자 뉴런과 연결된 연결강도들은 다음과 같이 입력 데이터에 대응되도록 조절된다.

$$\boldsymbol{w}_{ij}(t+1) = \boldsymbol{w}_{ij}(t) + \alpha(\boldsymbol{y}_i(t) - \boldsymbol{w}_{ij}(t))$$

여기서 $\boldsymbol{w}_{ij}(t+1)$는 뉴런 $(i,j)$에서 $t+1$번째 새로운 연결강도 벡터이고 $\boldsymbol{w_{ij}}(t)$는 $t$번째 연결강도벡터 그리고 $\boldsymbol{y}_i(t)$는 입력패턴 벡터 그리고 $\alpha$는 학습률을 나타낸다.

### 2.5. 모형기반 군집방법

데이터 $\boldsymbol{y}$는 서로 독립인 $N$개의 다변량 관측벡터 $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N$으로 구성되어 있다고 가정한다. 데이터 $\boldsymbol{y}$가 $G$개의 군집을 형성하고 $k$번째 군집에 속하는 관측벡터 $\boldsymbol{y}_i$의 확률밀도함수는 $f_k(\boldsymbol{y}_i|\theta_k)$라고 가정할 때 혼합모형에 대한 우도함수는 다음과 같다.

$$\mathcal{L}_{MIX}(\theta_1, \cdots, \theta_G|\boldsymbol{y}) = \prod_{i=1}^{N}\sum_{k=1}^{G}\tau_k f_k(\boldsymbol{y}_i|\theta_k)$$

여기서 $\tau_k$는 관측벡터가 $k$번째 군집에 속할 확률이며 $\tau_k \geq 0,\ \sum_{k=1}^{G}\tau_k = 1$이다.

가우시안 혼합모형 하에서 $f_k(\boldsymbol{y}_i|\theta_k)$은 다음과 같이 평균 벡터 $\mu_k$와 공분산 행렬 $\Sigma_k$인 다변량 정규 확률밀도 함수를 갖는다.

$$f_k(\boldsymbol{y}_i|\mu_k, \Sigma_k) = \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{y_i} - \mu_k)^T\Sigma_k^{-1}(\boldsymbol{y_i} - \mu_k)\right\}}{\sqrt{\det(2\pi\Sigma_k)}}$$

군집의 수 $G$가 정해지면 EM 알고리즘에 의해 $\tau_k$, $\mu_k$, $\Sigma_k$가 추정된다. 모형을 선택할 때 BIC (Bayesian information criterion)값이 최대가 되는 군집의 수를 최종모형으로 선택한다 (Fraley와 Raftery, 2002).

## 3. 군집 타당성 측도

### 3.1. 내부 측도

(1) 연결성

$nn_{i(j)}$를 개체 $i$로부터 $j$번째 가까이 위치한 개체라 할 때 다음과 같이 정의할 수 있다.

$$x_{i,nn_{i(j)}} = \begin{cases} 0, & \text{개체 } i\text{와 } nn_{i(j)}\text{는 같은 집락에 속할 때} \\ 1/j, & \text{기타} \end{cases}$$

따라서 $N$개 개체들에 대한 $K$개 군집 $\boldsymbol{C} = \{C_1, \cdots, C_K\}$에 대한 연결성은 다음과 같다.

$$Conn(\boldsymbol{C}) = \sum_{i=1}^{N} \sum_{j=1}^{M} x_{i,nn_{i(j)}}$$

여기서 $M$은 개체로부터 측정되는 변수의 수이다. 연결성 값은 0과 $\infty$사이이며 값이 작을수록 좋은 군집의 결과를 나타낸다.

(2) 실루엣 너비

개체 $i$에 대한 실루엣 너비를 다음과 같이 정의한다.

$$S(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$$

여기서 $a(i)$는 $i$번째 개체와 똑같은 군집에 있는 다른 개체들 간의 평균 거리이고 $b(i)$는 $i$번째 개체와 다른 군집에 있는 개체들 간의 거리의 최소값이다. $-1 \leq S(i) \leq 1$이고 $S(i)$가 1에 가까우면 군집화가 잘된 것이고 $S(i)$가 0에 가까우면 $i$번째 개체가 다른 가장 가까운 이웃 군집에 할당될 수도 있다는 뜻이고 $S(i)$가 -1에 가까우면 군집화가 잘못된 것을 의미한다.

$N$개 개체들에 대한 평균 실루엣 너비는 다음과 같다.

$$\bar{S} = \frac{1}{N} \sum_{i=1}^{N} S(i)$$

이 값은 군집화가 잘 되었는지를 측정하는 값으로 큰 값을 갖는 경우 최상의 군집화를 나타낸다.

(3) Dunn 지수

Dunn 지수는 최대의 군집내 거리 (intra-cluster distance)에 대한 최소의 군집간 거리 (inter-cluster distance)의 비로 나타낸다.

$$D(\boldsymbol{C}) = \frac{\min_{C_k, C_l \in \boldsymbol{C}, C_k \neq C_l} (\min_{i \in C_k, j \in C_l} dist(i,j))}{\max_{C_m \in \boldsymbol{C}} diam(C_m)}$$

여기서 $dist(i,j)$는 개체 $i$와 $j$사이의 거리이고 $diam(C_m)$는 군집 $C_m$에 있는 개체들 간의 최대 거리를 나타낸다. 같은 군집에 속해있는 두 개체간의 거리가 작을수록, 다른 군집에 속해있는 두 개체간의 거리가 클수록 $D(\boldsymbol{C})$값은 커지므로 이 값이 클수록 군집화가 잘 되었다고 판단할 수 있다.

### 3.2. 안정적 측도

(1) APN (average proportion of non-overlap)

APN은 원래의 자료로부터 얻어진 군집화와 하나의 변수를 제외한 자료로부터 얻어진 군집화를 비교하여 같은 군집에 속해 있지 않은 개체들의 평균비율을 측정한 측도로서 다음과 같이 정의된다.

$$APN(\boldsymbol{C}) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\left(1 - \frac{n(C^{i,l} \cap C^{i,0})}{n(C^{i,0})}\right)$$

여기서 $N$은 개체의 수이고 $M$은 전체 변수의 수 그리고 $C^{i,0}$은 원래의 자료로부터 얻어진 군집화에서 개체 $i$를 포함하는 군집이고 $C^{i,l}$은 $l$번째 변수를 제외한 자료로부터 얻어진 군집화에서 개체 $i$를 포함하는 군집이다. APN의 범위는 0과 1사이이고 0에 가까울수록 일관성이 있는 군집화를 나타낸다.

(2) AD (average distance)

AD는 원래의 자료로부터 얻어진 군집화와 하나의 변수를 제외한 자료로부터 얻어진 군집화를 비교하여 같은 군집에 속해 있는 개체들 간의 평균거리를 측정한 측도로서 다음과 같이 정의된다.

$$AD(\boldsymbol{C}) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\frac{1}{n(C^{i,0})n(C^{i,l})}\left[\sum_{i\in C^{i,0}, j\in C^{i,l}} dist(i,j)\right]$$

AD의 범위는 0과 $\infty$사이이고 0에 가까울수록 일관성이 있는 군집화를 나타낸다.

(3) ADM (average distance between means)

ADM는 원래의 자료로부터 얻어진 군집화와 하나의 변수를 제외한 자료로부터 얻어진 군집화를 비교하여 같은 군집에 속해 있는 개체들의 군집중심 간의 평균거리를 측정한 측도로서 다음과 같이 정의된다.

$$ADM(\boldsymbol{C}) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M} dist(\bar{x}_{C^{i,l}}, \bar{x}_{C^{i,0}})$$

여기서 $\bar{x}_{C^{i,0}}$은 원래의 자료로부터 얻어진 군집화에서 개체 $i$를 포함하는 군집에서 개체들의 평균이고 $\bar{x}_{C^{i,l}}$는 $l$번째 변수를 제외한 자료로부터 얻어진 군집화에서 개체 $i$를 포함하는 군집에서 개체들의 평균이다. ADM의 범위는 0과 $\infty$사이이고 0에 가까울수록 군집화의 결과가 일관성이 있다고 판단할 수 있다.

(4) FOM (figure of merit)

FOM은 원래의 자료로부터 군집화가 되었을 때 제외한 변수에 있는 자료들의 평균 군집내 분산을 측정한 측도로 $K$개의 군집에서 제외된 변수 $l$에 대한 FOM은 다음과 같이 정의한다.

$$FOM(l, \boldsymbol{C}) = \sqrt{\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in C_k(l)} dist(x_{i,l}, \bar{x}_{C_{k(l)}})^2}$$

여기서 $x_{i,l}$은 $l$번째 제외된 변수의 $i$번째 개체이고 $\bar{x}_{C_k(l)}$은 원래의 자료로부터 얻어진 군집 $C_k(l)$내의 평균 개체이다. 모든 제외된 변수들에 대한 FOM은 다음과 같다.

$$FOM = \sum_{l=1}^{M} FOM(l, \boldsymbol{C})$$

FOM의 범위는 0과 $\infty$사이이고 0에 가까울수록 좋은 군집화를 나타낸다.

### 3.3. 생물학적 측도

(1) BHI (biological homogeneity index)

BHI는 군집들이 어느정도 생물학적으로 균일한가를 측정하는 측도이다. $\mathbf{B} = \{B_1, \cdots, B_F\}$는 $F$개의 기능적으로 분류된 클래스 집합이라 하고 $B(i)$는 유전자 $i$를 포함하는 기능적인 클래스라 하자. 그러면 주어진 통계적인 군집 분할 $\boldsymbol{C} = \{C_1, \cdots, C_K\}$에 대하여 생물학적 클래스 $\mathbf{B}$에 대한 BHI는 다음과 같이 정의된다.

$$BHI(\boldsymbol{C}, \boldsymbol{B}) = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{n_k(n_k-1)} \sum_{i \neq j \in C_k} I(B(i) = B(j))$$

여기서 $I(B(i) = B(j))$는 지시함수이고 $n_k = n(C_k \cap B)$이다. BHI의 값의 범위는 [0,1]이고 값이 클수록 생물학적으로 균일하다는 것을 나타낸다.

(2) BSI (biological stability index)

BSI는 다른 안정성측도와 비슷하게 군집이 생물학적으로 어느 정도 안정한가를 측정하는 측도로서 다음과 같이 정의된다.

$$BSI(\boldsymbol{C}, \boldsymbol{B}) = \frac{1}{F}\sum_{k=1}^{F} \frac{1}{n(B_k)(n(B_k)-1)M} \sum_{l=1}^{M} \sum_{i \neq j \in B_k} \frac{n(C^{i,0} \cap C^{j,l})}{n(C^{i,0})}$$

BSI의 값의 범위는 [0,1]이고 큰 값은 생물학적으로 안정하다는 것을 나타낸다.

## 4. 실험 및 논의사항

본 절에서는 고차원의 유전자 발현자료에 대하여 군집 타당성 분석을 수행하기 앞서 모의실험을 통해 생성된 저차원과 고차원의 인공적인 자료와 실제 SRBCT 자료를 가지고 5가지 군집방법들의 성능을 여러 가지 타당성 측도를 가지고 비교하고자 한다. 모의실험은 R 프로그램을 이용하며 clValid (Brock 등, 2008) 패키지를 이용하여 군집 타당성 측도를 계산하고자 한다.

### 4.1. 모의실험에 의한 군집분석

(1) 저차원의 모의실험

4개의 군집에서 각각 3변량 정규분포를 따르고 공분산이 존재하는 경우와 존재하지 않는 경우 각각 나누어서 관측벡터 100개씩을 발생시킨다. 즉,

$$\boldsymbol{y_i} \sim N_3(\mu_i, \Sigma_k), i = 1, 2, 3, 4;\ k = 1, 2,$$

여기서

$$\mu_1 = \begin{pmatrix} 2 \\ 6 \\ 11 \end{pmatrix}, \mu_2 = \begin{pmatrix} 5 \\ 9 \\ 14 \end{pmatrix}, \mu_3 = \begin{pmatrix} 8 \\ 12 \\ 17 \end{pmatrix}, \mu_4 = \begin{pmatrix} 11 \\ 15 \\ 20 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 2 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 3 \end{pmatrix}.$$

표 4.1은 각 군집방법들에 대해 10번 반복 모의실험을 통해 얻은 내적 측도값이며 $\Sigma_1$와 $\Sigma_2$인 경우 얻어진 자료이다.

공분산이 0인 경우 연결성 측도 관점에서 보면 계층적 군집방법을 제외한 K-평균법, PAM, SOM 그리고 모형기반 군집방법이 군집의 수 $k$=4인 경우 좋은 군집화를 나타내고 있고 계층적 군집방법인 경우는 군집의 개수가 적을수록 좋은 결과를 보여주고 있다.

**표 4.1** 저차원 모의실험 자료에 대해 얻은 내적 측도 값

| 군집방법 | 군집측도 | 군집의 개수 ($k$) | | |
|---|---|---|---|---|
| | | 3 | 4 | 5 |
| 계층적 | 연결성 | 5.0702, 21.5544 | 6.8853, 30.1956 | 9.8143, 35.7619 |
| | Dunn 지수 | 0.1097, 0.0373 | 0.1579, 0.0450 | 0.1579, 0.0516 |
| | 실루엣 너비 | 0.5213, 0.4465 | 0.5541, 0.4152 | 0.5057, 0.3948 |
| K-평균법 | 연결성 | 31.9056, 45.4151 | 5.8548, 46.9893 | 35.1968, 73.2563 |
| | Dunn 지수 | 0.0438, 0.0248 | 0.1579, 0.0666 | 0.0506, 0.0358 |
| | 실루엣 너비 | 0.5062, 0.4626 | 0.5567, 0.4447 | 0.4748, 0.3817 |
| PAM | 연결성 | 30.7222, 41.7873 | 5.8548, 47.0746 | 34.0000, 55.3833 |
| | Dunn 지수 | 0.0438, 0.0346 | 0.1579, 0.0533 | 0.0700, 0.0388 |
| | 실루엣 너비 | 0.5069, 0.4623 | 0.5567, 0.4445 | 0.4621, 0.3868 |
| SOM | 연결성 | 31.5210, 41.7139 | 5.8548, 52.0988 | 32.6940, 75.3488 |
| | Dunn 지수 | 0.0438, 0.0345 | 0.1579, 0.0488 | 0.0613, 0.0499 |
| | 실루엣 너비 | 0.5068, 0.4623 | 0.5567, 0.4443 | 0.4641, 0.3688 |
| 모형기반 | 연결성 | 9.2643, 92.6127 | 5.8548, 56.5345 | 17.7452, 91.2060 |
| | Dunn 지수 | 0.0860, 0.0203 | 0.1579, 0.0522 | 0.0996, 0.0178 |
| | 실루엣 너비 | 0.5168, 0.4306 | 0.5567, 0.4158 | 0.4588, 0.3741 |

Dunn 지수와 실루엣 너비 관점에서 군집은 비슷한 결과를 보여주고 있고 $k$=4인 경우 모든 군집방법들에서 좋은 군집화를 보여주고 있다.

공분산이 존재하는 경우 Dunn 지수에 대해서는 공분산이 0인 경우처럼 $k$=4에서 좋은 군집화를 보여주고 있으나 실루엣 너비 측면에서 공분산이 0인 경우와 같은 결과를 낳으나 실루엣 너비 측면에서는 $k$=3에서 좋은 군집화를 보였다.

표 4.2는 각 군집방법들에 대해 모의실험을 통해 자료를 10번 반복 생성하여 얻은 안정적 측도값 들이다.

**표 4.2** 저차원 모의실험 자료에 대한 안정적 측도 값

| 군집방법 | 군집측도 | 군집의 개수 ($k$) | | |
|---|---|---|---|---|
| | | 3 | 4 | 5 |
| 계층적 | APN | 0.1897, 0.1596 | 0.0775, 0.1315 | 0.0837, 0.1844 |
| | AD | 3.7273, 4.3792 | 2.4277, 3.5674 | 2.4209, 3.1626 |
| | ADM | 1.7857, 2.0546 | 0.3540, 1.5436 | 0.3766, 0.9549 |
| | FOM | 1.5952, 1.9267 | 1.2131, 1.6998 | 1.2133, 1.4800 |
| K-평균법 | APN | 0.1216, 0.1302 | 0.0542, 0.1310 | 0.1794, 0.2850 |
| | AD | 3.3471, 3.6707 | 2.3549, 2.9997 | 2.3638, 3.0276 |
| | ADM | 0.8522, 0.6717 | 0.2455, 0.5657 | 0.6422, 1.1262 |
| | FOM | 1.6037, 1.7440 | 1.1575, 1.4484 | 1.1592, 1.4530 |
| PAM | APN | 0.1337, 0.1396 | 0.0630, 0.1342 | 0.1328, 0.2859 |
| | AD | 3.3800, 3.7043 | 2.3761, 3.0109 | 2.3189, 3.0186 |
| | ADM | 0.9551, 0.8437 | 0.2795, 0.6070 | 0.4271, 1.2201 |
| | FOM | 1.6100, 1.7579 | 1.1754, 1.4522 | 1.1673, 1.4951 |
| SOM | APN | 0.1336, 0.1696 | 0.0542, 0.1353 | 0.1892, 0.2614 |
| | AD | 3.3832, 3.7770 | 2.3549, 3.0077 | 2.3773, 3.0007 |
| | ADM | 0.9487, 1.1248 | 0.2455, 0.5918 | 0.6988, 1.1931 |
| | FOM | 1.6103, 1.7747 | 1.1575, 1.4539 | 1.1858, 1.4866 |
| 모형기반 | APN | 0.1877, 0.2485 | 0.0542, 0.0933 | 0.0660, 0.1551 |
| | AD | 3.7430, 4.1759 | 2.3549, 2.9882 | 2.3494, 2.9782 |
| | ADM | 1.8343, 1.7759 | 0.2455, 0.4169 | 0.3244, 0.7363 |
| | FOM | 1.5942, 1.8072 | 1.1575, 1.3870 | 1.1589, 1.3746 |

공분산 존재 여부에 관계없이 모든 군집방법들이 거의 4개의 안정적 측도 면에서 $k$=4인 경우 일관성이 있는 군집결과를 보여주고 있다.

표 4.3은 각 군집방법들에 대해 얻은 생물학적 측도를 계산한 표이다. BHI와 BSI 계산에서 기능적 군집의 수는 4개로 정의하여 얻었다.

**표 4.3** 저차원 모의실험 자료에 대한 생물학적 측도 값

| 군집방법 | 군집측도 | 군집의 개수 ($k$) | | |
|---|---|---|---|---|
| | | 3 | 4 | 5 |
| 계층적 | BHI | 0.8259, 0.7611 | 0.9901, 0.8874 | 0.7921, 0.9097 |
| | BSI | 0.8072, 0.6803 | 0.9179, 0.8286 | 0.9086, 0.8145 |
| K-평균법 | BHI | 0.6889, 0.6282 | 0.9901, 0.8503 | 0.9921, 0.8727 |
| | BSI | 0.8113, 0.7637 | 0.9380, 0.7739 | 0.8126, 0.6475 |
| PAM | BHI | 0.6936, 0.6389 | 0.9901, 0.8765 | 0.9709, 0.8615 |
| | BSI | 0.6936, 0.7634 | 0.9901, 0.7779 | 0.9709, 0.6627 |
| SOM | BHI | 0.6896, 0.6282 | 0.9901, 0.8499 | 0.9877, 0.7962 |
| | BSI | 0.7980, 0.7609 | 0.9381, 0.7663 | 0.8048, 0.6323 |
| 모형기반 | BHI | 0.8226, 0.6354 | 0.9901, 0.9573 | 0.9516, 0.9623 |
| | BSI | 0.8092, 0.7255 | 0.9380, 0.8876 | 0.9242, 0.8411 |

공분산 존재 여부에 관계없이 모든 군집방법들이 $k$=4인 경우 BHI, BSI 값들을 보면 대부분 큰 값을 갖고 있기 때문에 군집이 생물학적으로 안정되었다고 볼 수 있다.

(2) 고차원의 모의실험

고차원의 정규분포를 갖는 자료를 생성하여 군집방법들 간 비교하고자 한다. 여기서는 차원이 100인 4개의 군집을 생각하였으며 각 군집은 다음과 같이 정규분포를 갖는 10개의 자료들로 구성되어있다.

$$\boldsymbol{y_1} \sim N(1,1), \boldsymbol{y_2} \sim N(2,1), \boldsymbol{y_3} \sim N(3,1), \boldsymbol{y_4} \sim N(4,1).$$

표 4.4, 표 4.5 그리고 표 4.6은 고차원 자료에 대해 얻은 군집 타당성 측도값 들이다.

**표 4.4** 고차원 모의실험 자료에 대해 얻은 내적 측도 값

| 군집방법 | 군집측도 | 군집의 개수 ($k$) | | |
|---|---|---|---|---|
| | | 3 | 4 | 5 |
| 계층적 | 연결성 | 3.7989 | 5.6502 | 10.8512 |
| | Dunn 지수 | 0.7143 | 0.8420 | 0.7963 |
| | 실루엣 너비 | 0.1983 | 0.1667 | 0.1429 |
| K-평균법 | 연결성 | 3.7989 | 5.6502 | 13.9153 |
| | Dunn 지수 | 0.7143 | 0.8420 | 0.7963 |
| | 실루엣 너비 | 0.1983 | 0.1667 | 0.1402 |
| PAM | 연결성 | 4.9288 | 5.6502 | 18.1379 |
| | Dunn 지수 | 0.6757 | 0.8420 | 0.7858 |
| | 실루엣 너비 | 0.1909 | 0.1667 | 0.1319 |
| SOM | 연결성 | 4.7836 | 5.6502 | 13.8933 |
| | Dunn 지수 | 0.6685 | 0.8420 | 0.7901 |
| | 실루엣 너비 | 0.1976 | 0.1667 | 0.1336 |
| 모형기반 | 연결성 | 3.0375 | 5.6502 | 15.5953 |
| | Dunn 지수 | 0.7119 | 0.8420 | 0.7589 |
| | 실루엣 너비 | 0.1990 | 0.1667 | 0.1341 |

표 4.4에서 보면, Dunn 지수 관점에서는 군집의 수 $k$=4인 경우 좋은 군집화를 보이고 있고 연결성과 실수엣 너비 측도에서는 군집의 수가 적을수록 좋은 군집 결과를 보여주고 있다.

**표 4.5** 고차원 모의실험 자료에 대한 안정적 측도 값

| 군집방법 | 군집측도 | 군집의 개수 ($k$) | | |
|---|---|---|---|---|
| | | 3 | 4 | 5 |
| 계층적 | APN | 0.0463 | 0.0000 | 0.0252 |
| | AD | 14.0747 | 12.7594 | 12.5775 |
| | ADM | 0.9887 | 0.0000 | 0.6900 |
| | FOM | 1.0740 | 0.9966 | 1.0022 |
| K-평균법 | APN | 0.0463 | 0.0000 | 0.0486 |
| | AD | 14.0747 | 12.7594 | 12.5742 |
| | ADM | 0.9887 | 0.0000 | 0.8932 |
| | FOM | 1.0740 | 0.9966 | 0.9984 |
| PAM | APN | 0.0151 | 0.0027 | 0.0153 |
| | AD | 13.9941 | 12.7708 | 12.4066 |
| | ADM | 0.2689 | 0.0489 | 0.2188 |
| | FOM | 1.0726 | 0.9980 | 0.9954 |
| SOM | APN | 0.1641 | 0.0000 | 0.0714 |
| | AD | 14.9257 | 12.7594 | 12.6472 |
| | ADM | 3.1423 | 0.0000 | 0.9824 |
| | FOM | 1.0778 | 0.9966 | 0.9971 |
| 모형기반 | APN | 0.0100 | 0.0000 | 0.0333 |
| | AD | 13.8941 | 12.7594 | 12.4770 |
| | ADM | 0.2138 | 0.0000 | 0.5733 |
| | FOM | 1.0630 | 0.9966 | 1.0009 |

표 4.5에서 보면, 모든 군집방법들이 $k$=4인 경우 상대적으로 작은 값을 갖고 있어 일관성이 있는 군집화를 나타내고 있다.

**표 4.6** 고차원 모의실험 자료에 대한 생물학적 측도 값

| 군집방법 | 군집측도 | 군집의 개수 ($k$) | | |
|---|---|---|---|---|
| | | 3 | 4 | 5 |
| 계층적 | BHI | 0.8246 | 1.0000 | 0.9000 |
| | BSI | 0.9538 | 1.0000 | 0.9486 |
| K-평균법 | BHI | 0.8246 | 1.0000 | 1.0000 |
| | BSI | 0.9538 | 1.0000 | 0.9091 |
| PAM | BHI | 0.7557 | 1.0000 | 1.0000 |
| | BSI | 0.9229 | 0.9973 | 0.8659 |
| SOM | BHI | 0.7281 | 1.0000 | 1.0000 |
| | BSI | 0.8408 | 1.0000 | 0.8942 |
| 모형기반 | BHI | 0.8246 | 1.0000 | 1.0000 |
| | BSI | 0.9900 | 1.0000 | 0.8863 |

표 4.6에서 보면, 모든 군집방법들이 $k$=4에서 1의 값을 갖고 있어 생물학적으로 군집이 안정되었다고 알 수 있다.

### 4.2. SRBCT 자료에 대한 군집분석

본 논문에서 다루고자 하는 SRBCT자료에 대해 전처리 과정 (preprocessing)으로 log 변환을 수행하였으며 Quantile 표준화 작업 (Deshmukh와 Purohit, 2007)을 거친 다음 2308개 유전자 중에서 분산분석을 통해 유의확률이 0.01이하인 유의한 유전자 937개를 선택하였다.

모형기반 군집방법에서 실험에 사용된 공분산행렬의 모형은 “EI”라고 표기되는 동등구형볼륨 모형 (equal volume spherical model; Yeung 등, 2001b)이 사용되었다.

표 4.7은 각 군집방법들에 대해 SRBCT 자료로부터 얻은 내부 측도 값들이다.

**표 4.7** SRBCT 데이터에 대한 내적 측도 값

| 군집방법 | 군집측도 | 군집의 개수 ($k$) | | |
|---|---|---|---|---|
| | | 3 | 4 | 5 |
| 계층적 | 연결성 | 11.1218 | 15.8286 | 17.6976 |
| | Dunn 지수 | 0.6530 | 0.6147 | 0.6731 |
| | 실루엣 너비 | 0.1533 | 0.1306 | 0.1717 |
| K-평균법 | 연결성 | 11.1218 | 17.8254 | 23.5115 |
| | Dunn 지수 | 0.6530 | 0.5398 | 0.6876 |
| | 실루엣 너비 | 0.1533 | 0.1711 | 0.1830 |
| PAM | 연결성 | 14.4528 | 20.3560 | 27.9127 |
| | Dunn 지수 | 0.5956 | 0.5581 | 0.5469 |
| | 실루엣 너비 | 0.1558 | 0.1837 | 0.1666 |
| SOM | 연결성 | 12.6548 | 20.8365 | 24.9734 |
| | Dunn 지수 | 0.6201 | 0.5581 | 0.5469 |
| | 실루엣 너비 | 0.1462 | 0.1851 | 0.1240 |
| 모형기반 | 연결성 | 12.6548 | 12.9909 | 19.1532 |
| | Dunn 지수 | 0.6201 | 0.6201 | 0.6207 |
| | 실루엣 너비 | 0.1462 | 0.1875 | 0.1851 |

연결성 관점에서 보면 군집의 개수가 작을수록 좋은 군집결과를 보여주는 반면에 계층적 군집방법과 K-평균법을 제외한 PAM, SOM 그리고 모형기반 군집방법은 실루엣 너비 관점에서 군집의 개수 $k$= 4개인 경우 좋은 군집화를 나타내고 있다.

표 4.8은 각 군집방법들에 대해 SRBCT 자료로부터 얻은 안정적 측도 값들이다.

**표 4.8** SRBCT 자료에 대한 안정적 측도 값

| 군집방법 | 군집측도 | 군집의 개수 ($k$) | | |
|---|---|---|---|---|
| | | 3 | 4 | 5 |
| 계층적 | APN | 0.0000 | 0.0021 | 0.0000 |
| | AD | 35.9063 | 34.9988 | 31.9115 |
| | ADM | 0.0000 | 0.0761 | 0.0000 |
| | FOM | 0.8559 | 0.8486 | 0.7886 |
| K-평균법 | APN | 0.0000 | 0.0016 | 0.0000 |
| | AD | 35.9063 | 33.1267 | 31.5125 |
| | ADM | 0.0000 | 0.0698 | 0.0000 |
| | FOM | 0.8559 | 0.8043 | 0.7794 |
| PAM | APN | 0.0003 | 0.0000 | 0.0000 |
| | AD | 35.8257 | 33.0134 | 31.4399 |
| | ADM | 0.0114 | 0.0000 | 0.0000 |
| | FOM | 0.8501 | 0.8002 | 0.7772 |
| SOM | APN | 0.2968 | 0.2447 | 0.1882 |
| | AD | 38.7055 | 35.6217 | 33.5347 |
| | ADM | 11.5695 | 9.9810 | 8.5348 |
| | FOM | 0.8524 | 0.8146 | 0.7917 |
| 모형기반 | APN | 0.0006 | 0.0006 | 0.0040 |
| | AD | 35.6327 | 32.8309 | 31.4240 |
| | ADM | 0.0315 | 0.0315 | 0.1734 |
| | FOM | 0.8470 | 0.7986 | 0.7766 |

표 4.8로부터 모형기반 군집방법이 다른 방법들보다 일관성면에서 비교적 좋은 군집결과를 보여주고 있고 다음으로 계층적 군집방법, K-평균법 그리고 PAM 방법이 비슷한 결과를 보여주고 있다. SOM

방법은 APN과 ADM 측면에서 다른 값들보다 너무 큰 값을 가지고 있어 일관성이 떨어짐을 알 수 있다.

표 4.9는 각 군집방법들에 대해 SRBCT 자료로부터 얻은 생물학적 측도 값들이다. BHI와 BSI 계산에서 기능적 군집의 수는 4개의 암에 대한 자료이므로 4개로 정의하여 얻었다.

**표 4.9** SRBCT 자료에 대한 생물학적 측도 값

| 군집방법 | 군집측도 | 군집의 개수 ($k$) | | |
|---|---|---|---|---|
| | | 3 | 4 | 5 |
| 계층적 | BHI | 0.8114 | 0.6324 | 0.8000 |
| | BSI | 0.9783 | 0.9514 | 0.9533 |
| K-평균법 | BHI | 0.8114 | 0.8714 | 0.9500 |
| | BSI | 0.9783 | 0.8750 | 0.8585 |
| PAM | BHI | 0.6782 | 0.8585 | 0.8487 |
| | BSI | 0.8697 | 0.8796 | 0.7426 |
| SOM | BHI | 0.6049 | 0.8737 | 0.8989 |
| | BSI | 0.6701 | 0.7459 | 0.6561 |
| 모형기반 | BHI | 0.7998 | 0.9762 | 0.9692 |
| | BSI | 0.9779 | 0.9779 | 0.8520 |

표 4.9로부터 PAM 방법과 모형기반 군집방법 만이 BHI와 BSI 관점에서 원래 자료에서 클래스의 수와 일치하는 결과를 보였고 계층적 군집방법은 $k$=3인 경우 K-평균법과 SOM 방법은 BHI 관점에서 군집이 $k$=5인 경우 생물학적으로 안정하다는 것을 보여주고 있다.

## 5. 결론

지금까지 마이크로어레이 자료에 대한 연구는 주로 일부의 군집방법과 일부의 군집 측도에 극한하여 제한적으로 이루어져 왔다.

본 논문에서는 마이크로어레이 자료에 대해 많이 사용하는 계층적 군집방법과 비계층적 군집방법으로 K-평균법, PAM 그리고 SOM 그리고 모형기반 군집방법 들의 성능을 3 가지 타당성 측도인 내적 측도, 안정적 측도 그리고 생물학적 측도를 가지고 비교분석하였다.

우리는 실제 SRBCT 자료와 모의실험 통해 생성된 자료를 가지고 군집방법 성능을 비교하였다.

먼저 모의실험에서 거의 모든 군집 방법들이 3가지 측도 하에서 원래 자료수와 일치하는 좋은 군집화를 나타내고 있음을 알 수 있다. 그리고 SRBCT 자료에서는 모의실험 자료처럼 명확한 군집화 결과를 보여주지는 않으나 내적 측도인 연결성 면에서 보면 군집의 개수가 작을수록 좋은 군집결과를 보여주는 반면에 실루엣 너비 관점에서 보면 PAM, SOM 그리고 모형기반 군집방법은 군집의 개수 $k$= 4개인 좋은 군집화를 나타내고 있다. 안정성 측도 면에서 모형기반 군집방법이 다른 방법들보다 비교적 좋은 군집결과를 보여주고 있고 SOM 방법은 다른 방법들 보다 다소 일관성이 떨어짐을 알 수 있다.
생물학적 측도 면에서 PAM 방법과 모형기반 군집방법은 원래 클래스의 수와 일치하는 결과를 보였다.

## 참고문헌

김재희, 고윤실 (2009). 군집분석 비교 및 한우 관능평가 데이터 군집화. <응용통계연구>, **22**, 745-758.
여인권 (2011). 우리나라 기상자료에 대한 군집분석. <한국데이터정보과학회지>, **22**, 941-949.
이경아, 김재희 (2011). 효모 마이크로어레이 유전자 발현 데이터에 대한 군집화 비교. <한국데이터정보과학회지>, **22**, 741-753.

정윤경, 백장선 (2007). 고차원(유전자 발현) 자료에 대한 군집 타당성 분석 기법의 성능비교. <응용통계연구>, **20**, 167-181.

주용성, 정형주, 김병준 (2009). 한국 기상자료의 군집분석: 베이지안 모델기반 방법의 응용. <한국데이터정보과학회지>, **20**, 57-64.

황진수, 김지연 (2009). 마이크로어레이 자료에서 서포트 벡터 머신과 데이터 뎁스를 이용한 분류방법의 비교연구. <한국데이터정보과학회지>, **20**, 311-319.

Brock, G., Pihur, V., Datta, S. and Datta, S. (2008). clValid: An R package for cluster validation. *Journal of Statistical Software*, **25**, 1-21

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

Datta, S. and Datta, S. (2006). Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics*, **7**, 397.

Deshmukh, S. R. and Purohit, S. G. (2007). *Microarray data: Statistical analysis using R*, Alpha Science International Ltd, Oxford.

Dunn, J. C. (1974). Well separated clusters and fuzzy partitions. *Journal on Cybernetics*, **4**, 95-104.

Eisen, M. B., Spellman, T. P., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 863-14868.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611-631.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*, John Wiley & Sons, New York.

Khan, J., Wei, S., Ringer, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Anyonescu, C. R., Peterson, C. and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, **7**, 673-679

Kohonen, T. (1997). *Self-organizing maps*, Springer-Verlag, New York.

Handl, J., Knowles, J. and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**, 3201-3212.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, **28**, 100-108.

He, Y., Pan, W. and Lin, J. (2006). Cluster analysis using multivariate normal mixture models to detect differential gene expression with microarray data. *Computational Statistics & Data Analysis*, **51**, 641-658

Liu, Y. and Ringner, M. (2004). Multiclass discovery in array data. *BMC Bioinformatics*, **5**, 70-79.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53-65.

Yeung, K. Y., Haynor, D. R. and Ruzzo, W. L. (2001a). Validating clustering for gene expression data. *Bioinformatics*, **17**, 309-318.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L. (2001b). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977-987.

*Journal of the Korean Data & Information Science Society*
*2012,* **23**(1), 39–51

한국데이터정보과학회지

# Comparison of clustering methods of microarray gene expression data[†]

Jin Soo Lim[1] · Dong Hoon Lim[2]

[1]Department of Biological Sciences, Busan National University
[2]Department of Information Statistics, Gyeongsang National University

Received 30 October 2011, revised 17 November 2011, accepted 12 December 2011

**Abstract**

Cluster analysis has proven to be a useful tool for investigating the association structure among genes and samples in a microarray data set.We applied several cluster validation measures to evaluate the performance of clustering algorithms for analyzing microarray gene expression data, including hierarchical clustering, K-means, PAM, SOM and model-based clustering. The available validation measures fall into the three general categories of internal, stability and biological. The performance of clustering algorithms is evaluated using simulated and SRBCT microarray data. Our results from simulated data show that nearly every methods have good results with same result as the number of classes in the original data. For the SRBCT data the best choice for the number of clusters is less clear than the simulated data. It appeared that PAM, SOM, model-based method showed similar results to simulated data under Silhouette with of internal measure as well as PAM and model-based method under biological measure, while model-based clustering has the best value of stability measure.

*Keywords*: Biological measure, cluster analysis, internal measure, microarray, stability measure.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No.2011-0010089).
[1] Undergraduate student, Department of Biological Sciences, Busan National University, Busan 609-735, Korea.
[2] Corresponding author: Professor, Department of Information Statistics, Gyeongsang National University, Jinju 660-701, Korea. E-mail:dhlim@gnu.ac.kr