응용통계연구 제 **20**권 1호, *2007*년, *pp. 167–181*

고차원 (유전자 발현) 자료에 대한 군집 타당성분석 기법의 성능 비교

정윤경[1)] 백장선[2)]

요 약

유전자 발현 자료(gene expression data)는 전형적인 고차원 자료이며, 이를 분석하기 위한 여러 가지 군집 알고리즘(clustering algorithm)과 군집 결과들을 검증하는 군집 타당성분석 기법(cluster validation technique)이 제안되고 있지만, 이들 군집 타당성을 분석하는 기법의 성능에 대한 비교, 평가는 매우 드물다. 본 논문에서는 저차원의 모의실험 자료와 실제 유전자 발현 자료에 대하여 군집 타당성분석 기법들의 성능을 비교하였으며, 그 결과 내적 측도에서는 Dunn 지수, Silhouette 지수 순으로 뛰어났고 외적 측도에서는 Jaccard 지수가 성능이 가장 우수한 것으로 평가되었다.

주요용어: 유전자 발현 자료, 군집분석, 군집 타당성분석.

1. 서론

생명체의 유전적 특징은 한 두 유전자에 의해서 나타나는 것이 아니고, 여러 유전자간의 관계 하에서 이해될 수 있으므로, 특별히 다르게 발현된 각각의 유전자를 찾는 과정만으로는 생물체의 유전적 특징을 알아낼 수 없게 된다. 따라서 어떤 유전적 특징을 살펴보기 위해서는 유사하게 발현되는 유전자를 하나의 집단으로 모아 볼 필요가 있다. 이처럼 고차원의 유전자(변수)들을 의미 있는 최소한의 개수의 유전자들로 차원 축소하는 방법으로서도 군집 분석 방법이 사용되고 있다. 최근 유전자 발현 자료에서의 사람(환자나 정상인 등)과 유전자를 각각 행과 열로 구성하여 행렬 형태를 만들었을 때, 특성이 유사한 변수들끼리 집단으로 묶는 것 뿐만 아니라, 유전자들의 발현 강도(expression level)를 기준으로 환자와 정상인의 분류 또는 암환자의 세부적인 분류 또한 군집 분석 방법을 통하여 접근 가능하다.

이와 같이 일반적으로 분포에 대한 정보가 거의 없는 고차원의 데이터를 탐색하고자 할 때는, 군집 기법을 이용하여 집단 구조를 확인하기 마련이다. 군집 분석은 그림1.1에서 보듯이 크게 세 단계로 요약할 수 있는데 군집 타당성분석은 마지막 단계에 해당하며, 특히 알고리즘 개발과 군집결과의 검증 면에서 중요하다(Handl 등, 2005). 즉, 군집 알고리즘이 평가, 검증되었을 때, 그 알고리즘의 장단점 및 성능 등을 확인할 수 있고 이로써 보다 향상된 군집 기법의 개발에 도움을 줄 수 있다. 또한 군집 결과의 유의성에 대한 추정치를 제공하여 군집 결과를 검증함으로써 군집 결과에 대한 신뢰도를 측정할 수 있다.

1) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 통계학과, 박사과정
E-mail: joocc658@freechal.com
2) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 통계학과, 교수
E-mail: jbaek@chonnam.ac.kr

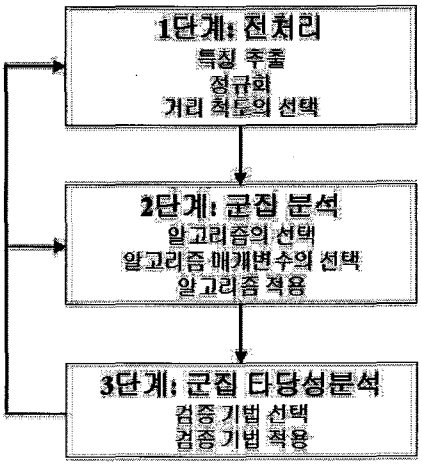


그림 1.1: 군집 분석의 주요 3단계

군집 타당성분석 기법은 크게 두 가지로 나눌 수 있다. 소속집단을 모르는(자율: unsupervised) 경우에 해당하는 내적 측도(internal measure)와 소속집단을 알고 있는(지도: supervised) 경우에 해당하는 외적 측도(external measure)가 그것이다. 소속집단에 대한 정보가 전혀 없는 상황에서의 정확한 군집 수를 예측하는 것이 자율 분류(unsupervised classification) 문제 중 가장 기본적인 문제이자 중요한 역할이며, 많은 군집 알고리즘이 실행 전에 정해진 군집 수를 필요로 한다. 이러한 문제를 해결하기 위하여 다양한 내적 측도의 군집 타당성 지수(cluster validity index)들이 군집 분할의 질을 평가할 수 있도록 제안되었다. 이러한 접근은 군집 알고리즘을 몇 번 실행하여 군집 수가 서로 다른 분할들을 얻은 후 군집 타당성 지수를 최적화하는 군집 분할을 최적의 분할으로 선택하는 것이다. 이와 같이 내적 측도로 구분되는 군집 타당성분석 기법의 주된 목적은 질적인 척도가 최적인 군집 분할을 확인하는 것이다. 내적 측도에 속하는 군집 타당성 분석 기법에는 Silhouette 방법(Rousseeuw, 1987), Dunn 지수(Bezdek 과 Pal, 1998; Dunn, 1974), Davies-Bouldin 지수(Davies 와 Bouldin, 1979), C 지수(Hubert 와 Schultz, 1976) 등이 있다. 외적 측도는 본래의 집단 구조와 군집 결과를 비교함으로써 군집 분할의 질을 평가하기 위한 척도인데, Jaccard 지수(Jaccard, 1912), Goodman-Kruskal 지수(Goodman 과 Kruskal, 1954), Rand 지수(Rand, 1971), Rand 수정 지수(Hubert 와 Arabie, 1985), 분리 지수(Pauwels 와 Frederix, 1999), Hubert의 Γ 통계량(Hubert 와 Arabie, 1985) 등이 이에 속한다. 군집 결과의 타당성을 분석하는 이들 기법 역시 성능을 비교, 평가할 필요가 있다. Bolshakova와 Azuaje의 최근 논문(Bolshakova 와 Azuaje, 2003a; 2003b)에서는 앞서 소개한 내적 측도들의 정규화와 타당성 종합 전략(validity aggregation strategy) 등을 통해 유전자 발현 자료에 대한 데이터 마이닝 성능을 향상시키는 방법들을 제시하였다. 우리는 전형적으로 고차원인 유전자 발현 자료에 대하여 군집 타당성분석 기법의 성능을 비교해 보고자 한다.

본 논문의 제2장에서는 군집 타당성분석 기법(내적 측도와 외적 측도)들을 요약, 정리

하였다. 제3장에서는 간단한 모의실험을 통해 생성된 저차원의 단순 구조를 갖는 자료에 대하여 군집 타당성분석 기법의 성능을 비교해 보고, 고차원의 실제 유전자 발현 자료에 대하여 군집 타당성분석 기법의 성능을 비교 및 평가하였다. 마지막으로 제4장에서 결론 및 토의를 기술한다.

2. 군집 타당성분석 기법

2.1. 내적 측도

2.1.1. Silhouette 지수(Silhouette Index)

임의의 군집 $X_j(j=1,\ldots,c)$에 대하여, 군집 X_j내에서 i번째 관측값의 구성 요소로서의 신뢰성 지표라 할 수 있는 Silhouette width를 개개의 관측값에 대해 다음과 같이 계산한다.

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}. \tag{2.1}$$

여기에서 $a(i)$는 X_j 내의 i번째 관측값과 i번째를 제외한 모든 관측값들 사이의 거리의 평균이고, $b(i)$는 X_j 의 i번째 관측값과 X_j를 제외한 군집 내의 관측값들과의 거리의 최소값을 나타낸다. 이 Silhouette width의 값의 범위는 $-1 \le s(i) \le 1$인데, $s(i)$가 1에 가까우면 군집화가 잘된 것이고, $s(i)$가 0에 가까우면 i번째 관측값이 가장 근접한 이웃 군집에 할당될 수도 있으며, $s(i)$가 -1에 가까우면 군집화가 잘못된 것이다.

다음으로, 군집 X_j에 대하여, 군집의 이질성(heterogeneity)과 분리된 정도(isolation property)의 특성을 나타내는 값으로서 다음과 같은 군집 Silhouette(cluster Silhouette)을 계산한다.

$$S_j = \frac{1}{m}\sum_{i=1}^{m} s(i). \tag{2.2}$$

여기에서 m은 X_j 내의 관측값 개수를 의미한다.

마지막으로, 분할 U에 대해, 다음과 같이 U에 대한 유효 타당성 지수(effective validity index)로 사용 가능한 전체 Silhouette 값(Global Silhouette value) GS_u를 계산한다.

$$GS_u = \frac{1}{c}\sum_{j=1}^{c} S_j. \tag{2.3}$$

위의 식은 가장 알맞은 군집 수를 추정하는 데 사용할 수 있으며, 가장 큰 값을 가지는 분할을 최적 분할로 고려한다.

2.1.2. Dunn 지수(Dunn's Index)

$U \leftrightarrow X : X_1 \cup \cdots \cup X_i \cup \cdots X_c$를 만족하는 임의의 분할 U에 대해, 군집들이 얼마나 잘 분류됐는지 확인할 수 있는 Dunn 지수, D를 다음과 같이 정의한다.

$$D(U) = \min_{1\le i\le c}\left\{\min_{\substack{1\le i\le c\\ j\ne i}}\left\{\frac{\delta(X_i,X_j)}{\max\limits_{1\le k\le c}\{\Delta(X_k)\}}\right\}\right\}. \tag{2.4}$$

여기에서 $\delta(X_i, X_j)$는 군집 X_i와 X_j의 군집 간 거리를, $\Delta(X_k)$는 군집 X_k의 군집 내 거리를 나타내고, c는 분할 U의 군집 수를 뜻한다. 이 척도의 주된 목적은 군집 내 거리(intracluster distance)를 최소화하는 반면, 군집 간 거리(intercluster distance)를 최대화하는 것이다. 이 값이 클수록 군집화가 잘된 것이고, D를 최대화하는 군집 수가 최적의 군집 수이다.

2.1.3. Davies-Bouldin 지수(Davies-Bouldin Index)

$U \leftrightarrow X : X_1 \cup \cdots \cup X_i \cup \cdots X_c$를 만족하는 임의의 분할 U에 대해, 군집화가 잘 된 정도를 확인할 수 있는 다음과 같은 Davies-Bouldin 지수, DB를 정의할 수 있다.

$$DB(U) = \frac{1}{c}\sum_{i=1}^{c} \max_{i \neq j} \left\{ \frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i, X_j)} \right\}. \tag{2.5}$$

여기에서 $\delta(X_i, X_j)$는 군집 X_i와 X_j의 군집 간 거리를, $\Delta(X_k)$는 군집 X_k의 군집 내 거리를 나타내고, c는 분할 U의 군집 수를 뜻한다. 이 값이 작을수록 군집 내의 밀도가 높으면서 군집의 중심이 다른 군집들의 중심과 멀리 떨어져 있음, 즉 군집화가 잘 됐음을 나타내며, DB를 최소화하는 군집 수가 최적의 군집 수이다.

Dunn 지수와 Davies-Bouldin 지수에 대해 각 15개씩 총 30개의 지수가 계산되었는데, 이것은 군집 간 거리와 군집 내 거리 방법들을 골고루 조합함으로써 이뤄진다(Bolshakova와 Azuaje, 2003b). 즉, 세 가지의 군집 내 거리, $\Delta_j, 1 \leq j \leq 3$와 다섯 가지의 군집 간 거리, $\delta_i, 1 \leq i \leq 5$를 이용했다. 예를 들면, DB_{32}는 군집 내 거리 계산 방법 Δ_3과 군집 간 거리 계산 방법 δ_2를 이용한 Davies-Bouldin 지수를 의미한다.

2.1.4. 내적 측도에서 사용한 거리

가장 기본이 되는 두 관측값 간의 거리 $d(x, y)$, 즉 유사성 척도로는 유클리드 거리계산법(euclidean metric)을 사용하였다.

1) 군집 간 거리

S와 T는 분할 U의 군집을 뜻하고, $d(x, y)$는 S와 T에 각각 속하는 임의의 관측값 x와 y 간의 거리를, 그리고 $|S|$와 $|T|$는 군집 S와 T에 각각 포함되는 관측값들의 개수를 나타낸다.

- 최단연결법(single linkage): 서로 다른 두 군집에 속하는 두 관측값 간의 거리 중 가장 작은 값,

$$\delta_1(S, T) = \min_{\substack{x \in S \\ y \in T}} \{d(x, y)\}. \tag{2.6}$$

- 최장연결법(complete linkage): 서로 다른 두 군집에 속하는 두 관측값 간의 거리 중 가장 큰 값,

$$\delta_2(S, T) = \max_{\substack{x \in S \\ y \in T}} \{d(x, y)\}. \tag{2.7}$$

- 평균연결법(average linkage): 서로 다른 두 군집에 속하는 모든 관측값들 간 거리의 평균,

$$\delta_3(S,T) = \frac{1}{|S||T|}\sum_{\substack{x\in S\\ y\in T}} d(x,y). \tag{2.8}$$

- 중심연결법(centroid linkage): 서로 다른 두 군집의 중심 사이의 거리,

$$\delta_4(S,T) = d(vs,vt), \quad \text{단, } vs = \frac{1}{|S|}\sum_{x\in S} x,\ vt = \frac{1}{|T|}\sum_{y\in T} y. \tag{2.9}$$

- 평균중심연결법(average of centroid linkage): 한 군집에 속하는 모든 관측값들과 다른 군집의 중심과의 거리의 평균,

$$\delta_5(S,T) = \frac{1}{|S|+|T|}\left(\sum_{x\in S} d(x,vt) + \sum_{y\in T} d(y,vs)\right). \tag{2.10}$$

2) 군집 내 거리

S는 분할 U의 군집을 뜻하고, $d(x,y)$는 S에 속하는 임의의 관측값 x와 y 간의 거리를, 그리고 $|S|$는 군집 S에 포함되는 관측값들의 개수를 나타낸다.

- 최장직경(complete diameter): 같은 군집에 속하는 관측값 간의 거리 중 가장 큰 값,

$$\Delta_1(S) = \max_{x,y\in S}\{d(x,y)\}. \tag{2.11}$$

- 평균직경(average diameter): 한 군집에 속하는 모든 관측값들 사이의 거리의 평균,

$$\Delta_2(S) = \frac{1}{|S|\cdot(|S|-1)}\sum_{\substack{x,y\in S\\ x\neq y}} d(x,y). \tag{2.12}$$

- 중심직경(centroid diameter): 한 군집에 속하는 모든 관측값들과 중심 사이의 거리의 평균의 2배,

$$\Delta_3(S) = 2\left(\frac{\sum_{x\in S} d(x,y)}{|S|}\right), \quad \text{단, } \bar{v} = \frac{1}{|S|}\sum_{x\in S} x. \tag{2.13}$$

2.2. 외적 측도

2.2.1. Jaccard 지수(Jaccard Index)

동일한 표본에 대한 집단 표지(class label) C와 군집 분석 결과 K 사이의 일치하는 정도를 측정하는 것으로 다음과 같이 계산한다.

$$J(C,K) = \frac{a}{a+b+c}. \tag{2.14}$$

표 2.1: 두 분할 비교에서의 분할표 표기법(notation)

class \ cluster	K_1	K_2	$\cdots$	K_C	합
C_1	n_{11}	n_{12}	$\cdots$	n_{1C}	$n_{1.}$
C_2	n_{21}	n_{22}	$\cdots$	n_{2C}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	$n_{R.}$
합	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.C}$	$n_{..} = n$

여기에서 a는 C에서 동일한 집단 표지를 가지고 K에서도 동일한 군집 내에 존재하는 쌍으로 이뤄진 점의 개수를, b는 C에서는 동일한 집단 표지를 가지나 K에서는 다른 군집에 속하는 쌍으로 이뤄진 점의 개수를, c는 K에서는 동일한 군집에 속하나 C에서는 다른 집단 표지를 갖는 쌍으로 이뤄진 점의 개수를 나타낸다. 이 값이 1에 가까울수록 군집화가 잘 된 것이다.

2.2.2. Goodman-Kruskal 지수(Goodman-Kruskal Association Index)

행(C)이 주어졌을 때 열(K)에 관한 그리고 열(K)이 주어졌을 때 행(C)에 관한 예측 타당성을 갖는 측도로서, 다음과 같이 표현된다(표2.1 참조).

$$\lambda = \frac{\sum_i \max_j n_{ij} + \sum_j \max_i n_{ij} - (\max_j n_{.j} + \max_i n_{i.})}{2n - (\max_j n_{.j} + \max_i n_{i.})}. \tag{2.15}$$

이 값은 행(C)과 열(K)에 대한 정보가 있을 때, 예측 오류수의 감소량을 상대적으로 표시한 연관성 측도로서, $0 \le \lambda \le 1$의 범위에서 존재하며, 1에 가까울수록 행(C)과 열(K)의 결합도가 강한 반면, 0에 가까울수록 행(C)과 열(K)의 결합도가 약함을 의미하므로, 1에 가까운 값이 나온다면 군집화가 잘 됐다고 할 수 있다.

2.2.3. Rand 수정 지수(Adjusted Rand Index)

Rand 수정 지수는 Rand 지수를 수정한 것으로서, 우선 Rand 지수를 살펴보면, 동일한 표본에 대한 집단 표지 C와 군집 분석 결과 K 사이의 일치하는 정도를 다음 식으로 계산할 수 있는데,

$$R(C, K) = \frac{a+d}{a+b+c+d}, \tag{2.16}$$

여기에서 a는 C에서 동일한 집단 표지를 가지고 K에서도 동일한 군집 내에 존재하는 쌍으로 이뤄진 점의 개수를, b는 C에서는 동일한 집단 표지를 가지나 K에서는 다른 군집에 속하는 쌍으로 이뤄진 점의 개수를, c는 K에서는 동일한 군집에 속하나 C에서는 다른 집단 표지를 갖는 쌍으로 이뤄진 점의 개수를, d는 K에서도 다른 군집에 속하고 C에서도 다른

집단 표지를 갖는 쌍으로 이뤄진 점의 개수를 의미한다. 이 값은 0과 1사이에 있으며, 1에 가까울수록 군집화가 잘 된 것이다. 그런데, 이 Rand 지수의 기대값을 계산해 보면 일정한 상수값이 나오지 않는다는 문제점이 발견되었다. 이를 개선한 Rand 수정 지수가 개발되었는데 다음과 같이 계산할 수 있다(Yeung과 Ruzzo, 2000)(표2.1 참조).

$$AR(C,K) = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right]/\binom{n}{2}}. \tag{2.17}$$

이 값의 기대값은 0이고, 1에 가까울수록 군집화가 잘 됐다고 한다.

3. 군집 타당성분석 기법의 성능 비교

3.1. 저차원 자료를 이용한 모의실험

고차원의 유전자 발현 자료에 대하여 군집 타당성분석 기법의 성능을 비교해 보기에 앞서, 저차원의 단순한 구조를 갖는 자료를 겹침의 정도가 조금씩 다르게 생성하여 지수들을 계산하였다. 군집 방법으로는 K-평균 알고리즘을 사용했고, 모의실험 자료는 모상관계수 $\rho = 0.25$인 이변량 정규분포를 따르며 분산공분산 행렬이 동일하고, 평균벡터만 각각 다른 표본들이다. 한 표본의 크기는 25이다.

본래의 집단 수가 $2, 3, 4$일 때의 세 가지 상황을 연출하여 각 상황에 대한 내적 측도들과 외적 측도들을 계산하였다. 본래의 집단 수가 2인 경우에는 초기 상태부터 확연히 구분되어짐을 한눈에 확인할 수 있도록 하였는데, 한 집단은 고정시키고 다른 한 집단은 평균좌표를 가로, 세로 +0.6씩 총 10회 증가시켜가면서 고정된 집단으로부터 초기상태보다 더 멀어지게 하였다. 본래의 집단 수가 3인 경우에는 처음에 두 집단은 상당히 겹쳐 있고 나머지 한 집단만 다소 떨어져 있는 상태로 출발했다. 겹쳐 보이는 집단 중 한 집단을 고정시키고, 다른 두 집단의 평균좌표에 가로, 세로 +0.6씩 총 10회 증가시켜가면서 이 고정된 집단으로부터 점점 더 멀어지게 하였다. 마지막에는 고정된 집단의 평균좌표로부터 가로, 세로 +6만큼 떨어져 있는 상태가 되어 세 집단 모두가 뚜렷하게 구분 가능해진다. 본래의 집단 수가 4인 경우에는, 서로 겹쳐있는 두 집단씩 꽤 먼 두 지점에서 초기화를 하였는데, 얼핏 봐서는 네 집단이라기 보다는 마치 두 집단으로 생각하기 쉬운 상태로 시작했다. 겹쳐 있는 것들 중 한 집단씩만, 즉 명확히 구분되어지는 두 집단을 고정시키고, 이들의 평균좌표에서 가로, 세로 +0.6씩 총 10회 증가시켜가면서 이 고정된 두 집단들로부터 각각 점점 더 멀어지게 하였다. 결국은 이동 집단들이 고정된 집단들의 평균좌표로부터 가로, 세로 +6만큼 떨어져 있는 상태가 되어 네 집단 모두가 뚜렷하게 구분 가능해진다. 아래의 식은 고정된 집단으로부터 점차 이동해 가는 집단의 구조를 간단하게 표현한 것이다.

$$(X_{L1}, X_{L2}) \sim N(\underline{\mu}_L, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{pmatrix} 1/3, & 1/12 \\ 1/12, & 1/3 \end{pmatrix}, \quad L = 1, \ldots, 10,$$

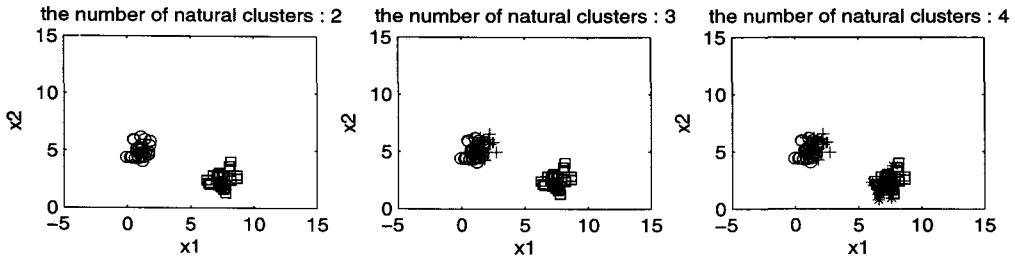
$$\underline{\mu}_L = \underline{\mu}_0 + L \times 0.6 \times \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \underline{\mu}_0 = \begin{pmatrix} \mu_{01} \\ \mu_{02} \end{pmatrix}.$$

L은 이동해 가는 각 단계를 나타내고, $\underline{\mu}_0$은 고정된 집단의 중심 위치(평균좌표)이고 $\underline{\mu}_L$은 이동하는 집단의 중심 위치이다. 예를 들어 평균이 $\underline{\mu}_0 = (1,5)'$인 고정된 집단을 기준으로 5단계에 있는 경우를 생각해보면, $\underline{\mu}_0$로부터 5(단계) $\times$ 0.6(증분) = 3만큼 이동했으므로 평균좌표는 $\underline{\mu}_5 = (4,8)'$이 된다. 그림3.1은 이러한 총 10 단계 중 초기 상태($L = 1$), 4단계($L = 4$), 7단계($L = 7$)와 이동의 최종 단계($L = 10$)를 각각 보여주고 있다.

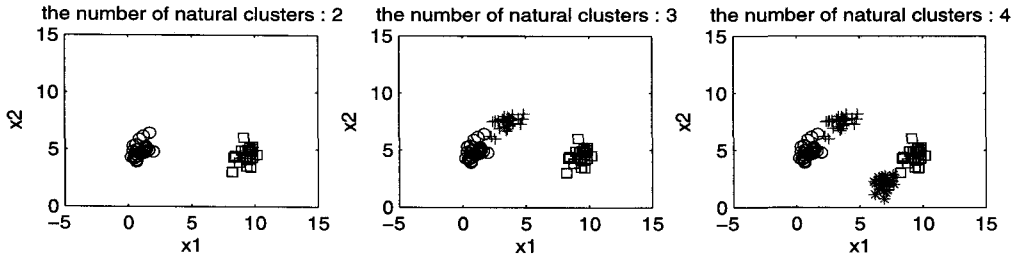
세 가지 상황의 열 단계 각각에 대하여, 최적의 군집 수 예측을 위한 군집 분석을 $k = 2, \ldots, 6$으로 10번씩 수행해서 얻은 10개의 내적 측도의 평균을 계산하고, 본래의 집단 구조와의 비교를 위한 군집 분석을 $k = 2, \ldots, 6$로 각각 10번씩 수행하여 얻은 10개의 외적 측도의 평균을 계산한다. 이 과정을 총 100번 반복하여 얻은 결과의 평균을 계산(한 예로 제시한 것이 표 3.1이다)한 후, 이들의 본래의 집단 수에 최초 도달한 단계(최초도달단계)와 본래의 집단 수와 일치하는 단계의 총 횟수(일치단계횟수)를 표 3.2에 요약하였다. 최초도달단계 값이 작을수록, 일치단계횟수 값이 클수록 성능이 뛰어나다고 할 수 있다.

내적 측도에 대한 비교 결과를 정리해 보면, 표 3.2에서 보듯이 본래의 집단 수가 3일 때 본래의 집단 수에 제일 먼저 도달한 것은 Dunn 지수였다. 표 3.1을 보면, 본래의 집단 수가 4일 때도 본래의 집단 수에 제일 먼저 도달한 것 역시 Dunn 지수였고 최초 도달 이후에도 본래의 집단 수를 유지하는 모습을 보여준 반면, 다른 지수(Silhouette 지수와 Davies-Bouldin 지수)들은 Dunn 지수보다 본래의 집단 수에 늦게 도달했을 뿐만 아니라, 그 이후에도 심한 변동이 있었다. 즉, 겹치는 부분이 완전히 없어지고 확연하게 네 집단으로 구분되는 자료 구조에서 최적의 군집 수를 3으로 예측하는 오류를 범한 것을 볼 수 있다. 이 결과를 종합해 보면, 저차원 자료에 대한 가장 우수한 성능을 가진 내적 측도를 예측해 보면 Dunn 지수라 할 수 있겠다.

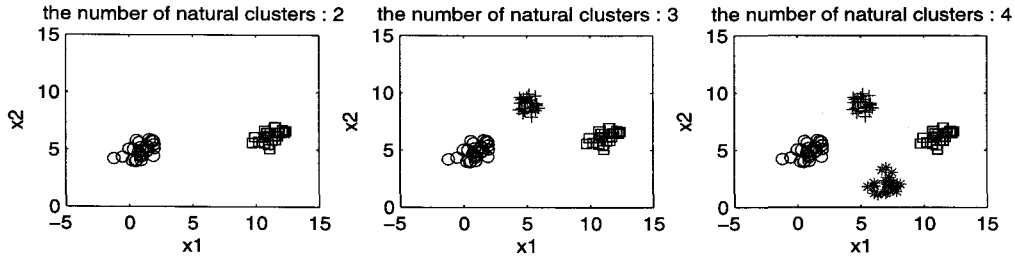
외적 측도에 대한 결과로는, 표3.2에서 보듯이, 본래의 집단 수가 3일 때 본래의 집단 수에 최초로 도달한 지수는 Goodman-Kruskal 지수와 Rand 수정 지수였는데 이 외에는 세 지수가 비슷하다가 본래의 집단 수가 4일 때 월등한 성능으로 두드러진 지수는 Jaccard 지수였다. 따라서 외적 측도 중에서는 Jaccard 지수가 저차원 자료에 대하여 다른 두 지수보다는 좀 더 나은 성능을 보여주었다.



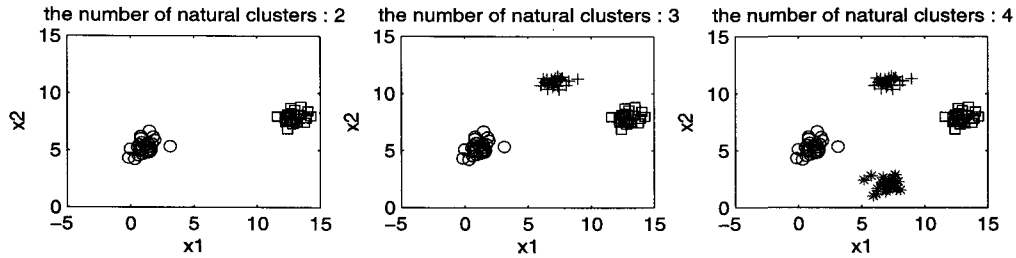
(a) 초기 상태(L=1)의 모의생성자료



(b) 4단계(L=4)의 모의생성자료



(c) 7단계(L=7)의 모의생성자료



(d) 최종 단계(L=10)의 모의생성자료

그림 3.1: 모의실험자료의 각 단계별 분포도(the number of natural clusters:본래의 집단 수)

표 3.1: 모의실험에서 얻은 내적 측도 평균(본래의 집단 수 : 4)

		$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
$L=1$	Silhouette(GS)	**0.8307**	0.5461	0.4458	0.4014	0.3807
	Dunn(D)	**5.7658**	1.4027	1.2613	1.1681	1.1289
	Davies-Bouldin(DB)	**0.5903**	2.8482	3.2099	3.2263	3.0676
$L=2$	Silhouette(GS)	**0.7839**	0.5884	0.4894	0.4390	0.4115
	Dunn(D)	**4.5673**	1.4367	1.3082	1.1687	1.1290
	Davies-Bouldin(DB)	**0.7463**	2.3332	2.7223	2.8192	2.7642
$L=3$	Silhouette(GS)	**0.7350**	0.6437	0.5502	0.4868	0.4472
	Dunn(D)	**3.7217**	1.4814	1.5166	1.1314	1.0888
	Davies-Bouldin(DB)	**0.8985**	1.6561	2.1137	2.3257	2.4195
$L=4$	Silhouette(GS)	0.6795	**0.6830**	0.5951	0.5280	0.4835
	Dunn(D)	**3.0955**	1.5514	1.7050	1.0395	1.0298
	Davies-Bouldin(DB)	**1.0747**	1.1228	1.6363	1.9355	2.1072
$L=5$	Silhouette(GS)	0.6271	**0.7051**	0.6355	0.5572	0.5091
	Dunn(D)	**2.6405**	1.6071	2.0911	0.9928	0.9900
	Davies-Bouldin(DB)	1.3388	**0.9502**	1.3964	1.7604	1.9528
$L=6$	Silhouette(GS)	0.5537	**0.7010**	0.6667	0.5887	0.5292
	Dunn(D)	2.2580	1.6301	**2.6062**	1.0248	0.9874
	Davies-Bouldin(DB)	1.8438	**1.1431**	1.2579	1.6356	1.8945
$L=7$	Silhouette(GS)	0.5226	0.6978	**0.7006**	0.6137	0.5449
	Dunn(D)	1.9918	1.6513	**3.3686**	1.0861	1.0443
	Davies-Bouldin(DB)	1.8608	1.1630	**1.1159**	1.5412	1.8132
$L=8$	Silhouette(GS)	0.5203	0.6987	**0.7216**	0.6329	0.5553
	Dunn(D)	1.9320	1.6750	**3.9506**	1.1353	1.0661
	Davies-Bouldin(DB)	1.6817	1.1311	**1.0307**	1.4893	1.8232
$L=9$	Silhouette(GS)	0.5341	**0.7180**	0.7153	0.6264	0.5561
	Dunn(D)	2.0551	1.6976	**3.8068**	1.0792	1.0577
	Davies-Bouldin(DB)	1.6656	**1.0146**	1.0486	1.5170	1.8149
$L=10$	Silhouette(GS)	0.5578	**0.7335**	0.6897	0.6100	0.5484
	Dunn(D)	2.2504	1.7012	**3.2105**	0.9700	0.9984
	Davies-Bouldin(DB)	1.4839	**0.9406**	1.1552	1.5443	1.8018

표 3.2: 모의실험 결과

		본래의 집단 수: **2**		본래의 집단 수:**3**		본래의 집단 수:**4**	
		최초 도달 단계	일치 단계 횟수	최초 도달 단계	일치 단계 횟수	최초 도달 단계	일치 단계 횟수
내적	Silhouette(GS)	1	10	6	5	7	2
	Dunn(D)	1	10	**5**	**6**	**6**	**5**
측도	Davies-Bouldin(DB)	1	10	6	4	7	2
외적	Jaccard(J)	1	10	2	9	2	**8**
	Goodman-Kruskal(GK)	1	10	**1**	9	2	4
측도	Adjusted Rand(AR)	1	10	**1**	9	2	3

최초도달단계 : 본래의 집단 수에 최초 도달한 단계(L)
일치단계횟수 : 본래의 집단 수와 일치하는 단계(L)의 총횟수

3.2. 유전자 발현 자료에 대한 군집 타당성분석 기법의 성능 비교

유전자 발현 자료는 Fort 와 Lambert-Lacroix(2005)에서 분석한 세 가지 자료들이다. 첫 번째 자료는 전처리 과정을 거쳐 유전자 3571개의 발현 강도로 표시되는 총 72개의 관측값들로 이뤄져 있고, 이 중 25개의 관측값은 급성 림프모구 백혈병(Acute Myeloid Leukaemia: AML)이고 47개의 관측값은 급성 골수성 백혈병(Acute Lymphoblastic Leukae-mia: ALL)으로서, 본래의 집단 수가 두 개인 백혈병 자료이다. 두 번째 자료는 전처리 과정을 거쳐 유전자 1224개의 발현 강도로 표시되는 총 62개의 관측값들로 구성되며, 이 중 22개의 관측값은 정상(normal)이고 40개의 관측값은 암(tumor)인, 본래의 집단 수가 두 개인 결장(colon) 자료이다. 그리고, 세 번째 자료는 전처리 과정을 거쳐 유전자 5966개의 발현 강도로 표시되는 총 102개의 관측값들로 구성되며, 이 중 50개의 관측값은 정상이고 52개의 관측값은 암으로, 본래의 집단 수가 두 개인 전립선(prostate) 자료이다.

군집 알고리즘은 모의실험에서 썼던 K-평균 알고리즘과 계층적 군집 알고리즘, 퍼지 C-평균, 그리고 자기 조직화 지도(self-organizing map: som)의 총 네 가지 알고리즘을 이용했다. 세 가지 자료에 대하여 최적의 군집 수 예측을 위한 네 가지 군집 분석을 $k = 2, \ldots, 6$으로 각각 10번씩 반복 수행하여 얻은 10개의 내적 측도의 평균들을 계산하고, 본래의 집단 구조와의 일치도를 얻기 위한 네 가지 군집 분석을 $k = 2, \ldots, 6$로 각각 10번씩 수행하여 얻은 10개의 외적 측도의 평균을 계산했다.

표 3.3, 표 3.4와 표 3.5는 세 가지 자료에 대해 각각 네 가지 군집 알고리즘을 수행하고 얻은 결과로부터 내적 측도의 최적의 군집 수가 본래의 집단 수와 일치하는지 여부와 외적 측도의 본래의 집단 구조와 가장 일치하는 군집 수가 본래의 집단 수와 일치하는지 여부를 정리하였다. 표 3.3를 보면, 백혈병 자료에서는 내적 측도 중 Dunn 지수가, 외적 측도 중 Jaccard 지수가 가장 우수한 성능을 가진 것으로 나타났다. 즉 내적 측도 중 Dunn 지수는

표 3.3: 백혈병 자료에 대한 결과

군집 알고리즘		K-평균	계층적	fcm	som	계
내적	Silhouette(GS)	×	○	×	○	2
	Dunn(D)	○	○	○	○	4
측도	Davies-Bouldin(DB)	×	×	○	×	1
외적	Jaccard(J)	○	○	○	○	4
	Goodman-Kruskal(GK)	○	○	×	×	2
측도	Adjusted Rand(AR)	○	○	×	×	2

계 : 각 측도의 최적 분할이 본래의 집단과 일치한 결과를 보여준 군집 방법의 수

네 가지 군집 방법 중 어느 것을 적용하든지 모두 본래의 집단 수를 정확하게 예측하였다. 반면에 Silhouette 지수나 Davies-Bouldin 지수는 네 방법 중 두 방법에서만 정확한 집단 수를 예측하였다. 외적 측도 중에서는 Jaccard 지수만이 네 방법 모두에서 정확한 집단 수를 예측하였다. 표 3.4를 보면, 결장 자료에서는 내적 측도 중 Silhouette 지수와 Dunn 지수가 동등하게, 그리고 외적 측도 중 Jaccard 지수가 가장 우수한 성능을 가진 것으로 나타났다. 표 3.5를 보면, 전립선 자료에서는 결장 자료에서와 마찬가지로 내적 측도 중 Silhouette 지수와 Dunn 지수가 동등하게, 그리고 외적 측도 중 Jaccard 지수가 가장 우수한 성능을 가진 것으로 나타났다.

표 3.4: 결장 자료에 대한 결과

군집 알고리즘		K-평균	계층적	fcm	som	계
내적	Silhouette(GS)	○	○	○	○	4
	Dunn(D)	○	○	○	○	4
측도	Davies-Bouldin(DB)	×	×	○	○	2
외적	Jaccard(J)	○	○	○	○	4
	Goodman-Kruskal(GK)	×	×	×	×	0
측도	Adjusted Rand(AR)	×	○	×	×	1

계 : 각 측도의 최적 분할이 본래의 집단과 일치한 결과를 보여준 군집 방법의 수

표 3.5: 전립선 자료에 대한 결과

군집 알고리즘		K-평균	계층적	fcm	som	계
내적	Silhouette(GS)	○	○	○	○	4
	Dunn(D)	○	○	○	○	4
측도	Davies-Bouldin(DB)	×	×	○	○	2
외적	Jaccard(J)	○	○	○	○	4
	Goodman-Kruskal(GK)	×	×	×	×	0
측도	Adjusted Rand(AR)	×	×	×	×	0

계 : 각 측도의 최적 분할이 본래의 집단과 일치한 결과를 보여준 군집 방법의 수

4. 결론 및 토의

저차원 자료에 대한 모의실험에서는 세 가지 내적 측도 중 Dunn 지수가 Silhouette이나 Davies-Bouldin 지수보다는 최적의 군집 수 예측력이 뛰어난 것으로 결과가 나왔으며, 세 가지 외적 측도 중에서는 Jaccard 지수가 Goodman-Kruskal 지수나 Rand 수정 지수보다는 조금 더 나은 성능을 보여주었다.

고차원의 세 가지 유전자 발현 자료에 대해서 종합적으로 결론을 내려보면, 세 가지 내적 측도 중 Dunn 지수, Silhouette 지수 순으로 가장 성능이 뛰어나고, 외적 측도에서는 Jaccard 지수가 다른 두 지수보다도 매우 훌륭한 성능을 보여주어 저차원 자료의 모의실험 결과와 일치하였다.

본 논문에서는 실제 유전자의 발현 자료를 이용하여 개체(관측값, 즉 정상 또는 암)에 대한 분류를 시행함으로써, 군집 타당성분석 기법의 성능을 비교, 평가하였다. 그러나 이처럼 차원이 매우 높은 유전자 발현 자료에 대한 군집분석의 앞 단계인 전처리 과정에서 차원 축소(특징 추출)가 유의하게 잘 이루어지면 군집분석의 성능을 전반적으로 높일 수 있으므로, 보다 효율적인 특징 추출 방법을 적용한다면 모든 군집 타당성분석 기법의 성능도 향상되리라 기대된다.

감사의 글

심사위원님들의 충고 덕분에 내용이 더 충실하게 되었습니다. 세심하게 지적해 주신 심사위원님들께 진심으로 감사의 뜻을 표합니다.

참고문헌

Bezdek, J. C. and Pal, N. R. (1998). Some new indexes of cluster validity, *IEEE Transactions on Systems, Man and Cybernetics, Part B:Cybernetics*, **28**, Issue 3, 301–315.

Bolshakova, N. and Azuaje, F. (2003a). Improving expression data mining through cluster validation, Conference Proceedings. *4th International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine 2003*, 19–22.

Bolshakova, N. and Azuaje, F. (2003b). Cluster validation techniques for genome expression data classification, *Signal Processing*, **83**, 825–833.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure, *IEEE Transactions on Pattern Recognition and Machine Intelligence*, **1**, 224–227.

Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions, *Journal Cybernet*, **4**, 95–104.

Fort, G. and Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression, *Bioinformatics*, **21**, 1104–1111.

Goodman, L. and Kruskal, W. (1954). Measures of associations for cross-validations, *Journal of the American Statistical Association*, **49**, 732–764.

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

Hubert, L. and Arabie, P. (1985). Comparing partitions, *Journal of Classification*, **2**, 193–218.

Hubert, L. and Schultz, J. (1976). Quadratic assignment as a general data-analysis strategy, *The British Journal of Mathematical & Statistical Psychology*, **29**, 190–241.

Jaccard, P. (1912). The distribution of flora in the alpine zone, *New Phytologist*, **11**, 37–50.

Pauwels, E. J. and Frederix, G. (1999). Finding salient regions in images: nonparametric clustering for image segmentation and grouping, *Computer Vision and Image Understanding*, **75**, 73–85.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association*, **66**, 846–850.

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. and Ruzzo, W. L. (2000). An Empirical Study on Principal Component Analysis for Clustering Gene Expression Data, *Technical Report UW-CSE-2000-11-03*, Department of Computer Science and Engineering, University of Washington.

[2006년 8월 접수, 2006년 11월 채택]

Comparison of the Cluster Validation Methods for High-dimensional (Gene Expression) Data

Yunkyoung Jeong[1)] Jangsun Baek[2)]

ABSTRACT

Many clustering algorithms and cluster validation techniques for high-dimensional gene expression data have been suggested. The evaluations of these cluster validation techniques have, however, seldom been implemented. In this paper we compared various cluster validity indices for low-dimensional simulation data and real gene expression data, and found that Dunn's index is the most effective and robust, Silhouette index is next and Davies-Bouldin index is the bottom among the internal measures. Jaccard index is much more effective than Goodman-Kruskal index and adjusted Rand index among the external measures.

Keywords: Gene expression data, cluster analysis, cluster validation.

1) Graduate Student, Department of Statistics, Chonnam National University, 300, Yongbong-dong, Buk-gu Gwangju, 500-757, Korea
E-mail: joocc658@freechal.com

2) Professor, Department of Statistics, Chonnam National University, 300, Yongbong-dong, Buk-gu, Gwangju 500-757, Korea
E-mail: jbaek@chonnam.ac.kr