Proceedings of Joint Conference of Korean Data And Information Science Society
and The Korean Data Analysis Society,
April 28 - 29, 2006, pp. 63-76

유전자 발현 자료를 이용한 군집 타당성분석 기법 비교

정윤경[1] · 백장선[2]

요 약

유전자 발현 자료(gene expression data)를 분석하기 위한 여러 가지 군집 알고리즘(clustering algorithm)과 군집 결과들을 검증하는 척도, 즉 군집 타당성분석 기법(cluster validation technique)이 제안되고 있지만, 이들 군집 타당성을 분석하는 기법들에 대한 성능의 비교·평가는 매우 드물다. 본 논문에서는 모의 생성 자료로 몇 가지 특정 상황을 연출하여 군집 타당성 분석 기법들을 비교해 보고, 실제 유전자 발현 자료 두 가지에 대해서도 이들 기법의 성능을 비교·평가해 보았다.

1. 서론

일반적으로 분포에 대한 정보가 거의 없는 복잡한(다차원) 데이터를 분석하고자 할 때는, 군집 기법(clustering technique)을 이용하여 본래의(natural) 집단 구조와 일치하는지를 확인하게 된다. 군집 분석(cluster analysis)은 크게 세 단계로 요약할 수 있는데(그림 1), 군집 타당성분석(cluster validation)은 마지막 단계에 해당하며, 특히 다음의 두 측면에 있어서 중요하다(Handl, Knowles, Kell, 2005).

- 알고리즘 개발
  군집 알고리즘이 평가·검증되어야만, 그 알고리즘의 장단점 및 성능 등을 확인할 수 있고 이로써 보다 향상된 군집 기법의 개발에 도움을 줄 수 있다.
- 군집 결과의 검증
  군집 결과의 유의성에 대한 추정치를 제공하여 군집 결과를 검증함으로써 군집 결과에 대한 신뢰 측도의 역할을 한다.

군집 타당성분석 기법은 크게 두 가지로 나눌 수 있다. 사전 정보(class)를 모르는(자율:unsupervised) 경우에 해당하는 내적 측도(Internal measure)와 사전 정보를 알고 있는(지도:supervised) 경우에 해당하는 외적 측도(External measure)가 그것이다. 사전 정보가 전혀 없는 상황에서의 정확한 군집 수를 예측하는 것이 자율 분류(unsupervised classification)문제 중 가장 기본

[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

적인 문제이자 중요한 역할이며, 많은 군집 알고리즘이 실행 전에 군집 수의 정의를 필요로 한다. 이 문제를 극복하기 위해서, 다양한 내적 측도의 군집 타당성 지수(cluster validity index)들이 군집 파티션의 질(quality)을 평가할 수 있도록 제안돼 왔다. 이러한 접근은 군집 알고리즘을 몇 번 실행하여 군집 수가 서로 다른 파티션들을 얻은 후 군집 타당성 지수를 최적화하는 군집 파티션(clustering partition)을 최적의 파티션으로 선택하는 것이다. 이와 같이 내적 측도로 구분되는 군집 타당성분석 기법의 주된 목적은 질의 척도(quality measure)가 최적인 군집 파티션을 확인하는 것이다.

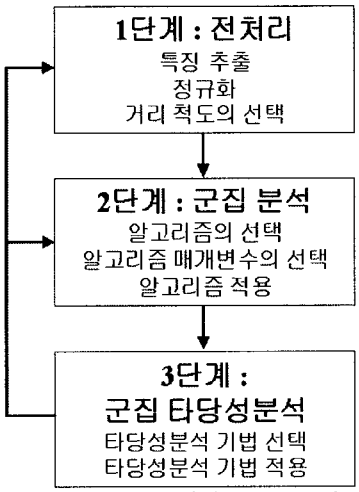

그림 1. 군집 분석의 주요 3단계

내적 측도에 속하는 군집 타당성 분석 기법에는 Silhouette 방법(Silhouette validation method) (Rousseeuw, 1987), Dunn 지수(Dunn's validation index)(Bezdek, Pal, 1998; Dunn, 1974), Davies-Bouldin 지수(Davies-Bouldin validation index)(Davies, Bouldin, 1979), C 지수(C index)(Hubert, Schultz, 1976) 등이 있다. 이에 대응되는 외적 측도는 본래의 집단 구조와 군집 결과를 비교함으로써 군집 파티션의 질을 평가하기 위한 척도인데, Jaccard 지수(Jaccard index)(Jaccard, 1912), Goodman-Kruskal 지수(Goodman-Kruskal association index)(Goodman, Kruskal, 1954), Rand 지수(Rand index)(Rand, 1971), Rand 수정 지수(Adjusted Rand index)(Hubert, Arabie, 1985), 분리 지수(Isolation index)(Pauwels, Frederix, 1999), Hubert의 Γ 통계량(Hubert's Γ statistics)(Hubert, Arabie, 1985) 등이 이에 속한다. 군집 결과의 타당성을 분석하는 이들 기법 역시 성능을 비교 · 평가할 필요가 있다. Bolshakova와 Azuaje의 최근 논문(Bolshakova, Azuaje, 2003a; Bolshakova, Azuaje, 2003b)에서는 앞서 소개한 군집 타당성 분석 기법들의 정규화(normalization)와 타당성 종합 전략(validity aggregation strategy) 등을 통해 유전자 발현 자료에 대한 데이터마이닝 성능을 향상시키는 방법들을 제시하기도 했다. 우리는 몇 가지 특정 상황의 모의 생성 자료와 실제 자료를 대상으로 외적 측도를 이용해서 내적 측도에 의한 군집 타당성분석 기법의 성능을 비교해 보고자 한다.

본 논문에서는 군집 분석과 군집 타당성분석 기법(내적 측도와 외적 측도)에 대해 살펴보고, 모의 생성 자료를 가지고 본래의 집단 구조를 모른다고 가정한 채 내적 측도로 최적의 군집 수를 예측하여 군집 분석을 실시한 후 본래의 집단 구조와 군집 결과(cluster result)를 비교하는 외적 측도를 통해 내적 측도의 성능을 평가하고, 실제 자료로써 유전자 발현 자료에도 같은 방법을 적용하여 군집 타당성분석 기법의 성능을 평가하고자 한다. 본 논문에서 사용한 군집 알고리즘은 K-평균 알고리즘이다.

2. 군집 타당성분석 기법

2.1 내적 측도

2.1.1 Silhouette 지수(Silhouette index)

주어진 군집, X_j (j=1,…,c)에 대하여, X_j의 각각의 샘플에 대하여, 군집 X_j 내에서의 i번째 샘플의 구성 요소로서의 신뢰성 지표라고 할 수 있는 질의 척도로서, 다음과 같은 Silhouette width를 정의한다.

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, \tag{1}$$

여기에서 $a(i)$는 X_j 내의 i번째 샘플과 모든(i번째를 제외한) 샘플들 사이의 거리의 평균이고, $b(i)$는 X_j의 i번째 샘플과 X_j를 제외한 군집 내의 샘플들과의 거리의 최소값을 나타낸다. 이 Silhouette width의 값의 범위는 $-1 \le s(i) \le 1$인데, $s(i)$가 1에 가까우면 군집화가 잘된 것("well-clustered")이고, $s(i)$가 0에 가까우면 i번째 샘플이 가장 근접한 이웃 군집에 할당될 수도 있으며, $s(i)$가 -1에 가까우면 군집화가 잘못된 것("misclassified")이다.

주어진 군집, X_j(j=1,…,c)에 대하여, 군집의 이질성(heterogeneity)과 분리된 정도(isolation property)의 특성을 나타내는 값으로서 다음과 같은 군집 Silhouette(cluster Silhouette)를 계산할 수 있는데,

$$S_j=\frac{1}{m}\sum_{i=1}^{m}s(i), \tag{2}$$

여기에서 m은 X_j 내의 샘플 개수를 의미한다.

$U \leftrightarrow X : X_1 \cup \cdots \cup X_i \cup \cdots \cup X_c$를 만족하는 모든 파티션에 대해, 최종적으로 다음과 같이 U에 대한 유효 타당성 지수(effective validity index)로 사용 가능한 전체 Silhouette 값(Global Silhouette value)을 계산한다.

$$GS_u=\frac{1}{c}\sum_{j=1}^{c}S_j. \tag{3}$$

위의 식은 가장 알맞은 군집 수를 추정하는 데에 적용될 수 있다고 증명된 바 있으며, 이 경우 가장 큰 값을 가지는 파티션을 최적 파티션으로 고려한다.

2.1.2 Dunn 지수(Dunn's index)

$U \leftrightarrow X : X_1 \cup \cdots \cup X_i \cup \cdots \cup X_c$를 만족하는 모든 파티션에 대해, 군집들이 조밀(compact)하게 잘 분류됐는지를 확인할 수 있는 Dunn 지수, D를 다음과 같이 정의할 수 있다.

$$D(U) = \min_{1 \le i \le c} \left\{ \min_{\substack{1 \le j \le c \\ j \ne i}} \left\{ \frac{\delta(X_i, X_j)}{\max_{1 \le k \le c} \{\Delta(X_k)\}} \right\} \right\}, \tag{4}$$

$\delta(X_i, X_j)$: 군집 X_i와 X_j의 군집 간 거리(intercluster distance)

$\Delta(X_k)$: 군집 X_k의 군집 내 거리(intracluster distance)

c : 파티션 U의 군집 수

이 척도의 주된 목적은 군집 내 거리를 최소화하는 반면, 군집 간 거리를 최대화하는 것이다. 이 값이 클수록 군집화가 잘된 것이고, D를 최대화하는 군집 수가 최적의 군집 수이다.

2.1.3 Davies-Bouldin 지수(Davies-Bouldin index)

$U \leftrightarrow X : X_1 \cup \cdots \cup X_i \cup \cdots \cup X_c$를 만족하는 모든 파티션에 대해, 군집들이 밀도 높게(compact) 잘 분류됐는지를 확인할 수 있는 다음과 같은 Davies-Bouldin 지수, DB를 정의할 수 있다.

$$DB(U) = \frac{1}{c} \sum_{i=1}^{c} \max_{i \ne j} \left\{ \frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i, X_j)} \right\}, \tag{5}$$

$\delta(X_i, X_j)$: 군집 X_i와 X_j의 군집 간 거리

$\Delta(X_k)$: 군집 X_k의 군집 내 거리

c : 파티션 U의 군집 수

이 값이 작을수록 군집 내의 밀도가 높으면서 군집의 중심이 다른 군집들의 중심과 멀리 떨어져 있으며, 즉 군집화가 잘 된 것이며, DB를 최소화하는 군집 수가 최적의 군집 수이다.

군집 간 거리와 군집 내 거리를 계산하기 위해서 여러 가지 방법을 사용할 수 있는데, Dunn 지수와 Davies-Bouldin 지수에 대해 각 15개씩 총 30개의 지수가 계산되었는데, 이것은 군집 간 거리와 군집 내 거리 방법들을 다양하게 조합함으로써 이뤄진다(Bolshakova, Azuaje, 2003b). 3개의 군집 내 거리, Δ_j $1 \le j \le 3$와 5개의 군집 간 거리, δ_i $1 \le i \le 5$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. 이들 군집 내 거리와 군집 간 거리의 수학적 정의는 다음, 다음 섹션에서 설명하고 있고, 모의 생성 자료에서

는 Dunn 지수와 Davies-Bouldin 지수의 이들 군집 내 거리 및 군집 간 거리 계산 방법의 조합에 대한 평균을 기준으로 결과를 정리하였다(표 2).

2.2. 외적 측도

2.2.1 Jaccard 지수(Jaccard index)

동일한 데이터셋(dataset)에 대한 집단 표지(class label) C와 군집 분석 결과 K 사이의 일치하는 정도를 측정하는 것으로 다음과 같이 계산한다. 이 값이 1에 가까울수록 군집화가 잘 된 것이다.

$$J(C,K)=\frac{a}{a+b+c}, \tag{6}$$

a : C에서 같은 집단 표지를 가지고, K에서도 같은 군집 내에 존재하는 쌍으로 이뤄진 점의 개수
b : C에서는 같은 집단 표지를 가지나, K에서는 다른 군집에 속하는 쌍으로 이뤄진 점의 개수
c : K에서는 같은 군집에 속하나, C에서는 다른 집단 표지를 갖는 쌍으로 이뤄진 점의 개수

2.2.2 Goodman-Kruskal 지수(Goodman-Kruskal Association index)

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

$$\lambda=\frac{\sum_i \max_j n_{ij}+\sum_j \max_i n_{ij}-(\max_j n_{\cdot j}+\max_i n_{i\cdot})}{2n-(\max_j n_{\cdot j}+\max_i n_{i\cdot})}. \tag{7}$$

이 값은 행(R)과 열(C)에 대한 정보가 있을 때, 예측 오류수의 감소량을 상대적으로 표시한 연관성 측도로써, $0\leq\lambda\leq1$의 범위에서 존재하며, 1에 가까울수록 행(R)과 열(C)의 결합도가 강한 반면, 0에 가까울수록 행(R)과 열(C)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 것으로, 이 값은 $0\leq R(U,V)\leq1$의 범위에서 존재하고, 1에 가까운 값일수록 군집화가 잘 된 것이다. 그런데, 이 Rand 지수의 문제점이 발견되었고, 이를 개선한 Rand 수정 지수가 개발되었는데 다음과 같이 계산할 수 있으며,

$$ARI(U,V)=\frac{\sum_{i,j}\binom{n_{ij}}{2}-\left[\sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{n_{i\cdot}}{2}+\sum_j\binom{n_{\cdot j}}{2}\right]-\left[\sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}\right]/\binom{n}{2}}, \tag{8}$$

이 값은 0과 1 사이의 범위에 존재하고, 1에 가까울수록 군집화가 잘 됐다고 한다.

2.3. 군집 타당성분석 기법에서 사용된 거리법(metric)

두 샘플 간의 거리, $d(x,y)$는 유클리드, 맨하탄, 체비쉐프 거리 등 여러 가지 거리 계산법이 있으나, 본 논문에서는 유클리드 거리를 사용하였다.

2.3.1 군집 간 거리

Dunn 지수와 Davies-Bouldin 지수를 계산하는 데에 다섯 가지의 거리 계산법을 사용했다. 여기에서 S와 T는 파티션 U의 군집을 의미하고, $d(x,y)$는 S와 T에 각각 속하는 임의의 샘플 x와 y 간의 거리를, 그리고 $|S|$와 $|T|$는 군집 S와 T에 각각 포함되는 샘플들의 개수를 의미한다.

1) 최단연결법(Single linkage) : 서로 다른 두 군집에 속하는 두 샘플 간의 거리 중 가장 작은 값

$$\delta_1(S,T)=\min\left\{d(x,y)\right\}_{x\in S,\, y\in T}. \tag{9}$$

2) 최장연결법(Complete linkage) : 서로 다른 두 군집에 속하는 두 샘플 간의 거리 중 가장 큰 값

$$\delta_2(S,T)=\max\left\{d(x,y)\right\}_{x\in S,\, y\in T}. \tag{10}$$

3) 평균연결법(Average linkage) : 서로 다른 두 군집에 속하는 모든 샘플들 간 거리의 평균

$$\delta_3(S,T)=\frac{1}{|S||T|}\sum_{\substack{x\in S\\ y\in T}} d(x,y) \tag{11}$$

4) 중심연결법(Centroid linkage) : 서로 다른 두 군집의 중심 사이의 거리

$$\delta_4(S,T)=d(vs,vt),\quad vs=\frac{1}{|S|}\sum_{x\in S}x,\quad vt=\frac{1}{|T|}\sum_{y\in T}y. \tag{12}$$

5) Average of centroid linkage: 한 군집에 속하는 모든 샘플들과 다른 군집의 중심과의 거리의 평균

$$\delta_5(S,T)=\frac{1}{|S|+|T|}\left(\sum_{x\in S} d(x,vt)+\sum_{y\in T} d(y,vs)\right). \tag{13}$$

2.3.2 군집 내 거리

Dunn 지수와 Davies-Bouldin 지수를 계산하는 데에 세 가지의 거리 계산법이 사용되었다. 여기에서 S는 파티션 U의 군집을 의미하고, $d(x,y)$는 S에 속하는 임의의 샘플 x와 y 간의 거리를, 그리고 $|S|$는 군집 S에 포함되는 샘플들의 개수를 의미한다.

1) Complete diameter : 같은 군집에 속하는 샘플 간의 거리 중 가장 큰 값

$$\Delta_1(S)=\max_{x,y\in S}\left\{d(x,y)\right\}. \tag{14}$$

2) Average diameter : 한 군집에 속하는 모든 샘플들 사이의 거리의 평균

$$\Delta_2(S)=\frac{1}{|S|\cdot(|S|-1)}\sum_{\substack{x,y\in S\\ x\neq y}} d(x,y) \quad (15)$$

3) Centroid diameter : 한 군집에 속하는 모든 샘플들과 중심 사이의 거리의 평균의 2배

$$\Delta_3(S)=2\left(\frac{\sum_{x\in S} d(x,\bar{v})}{|S|}\right),\quad \bar{v}=\frac{1}{|S|}\sum_{x\in S} x \quad (16)$$

3. 모의 생성 자료를 이용한 비교 실험

3.1 실험 자료 생성

본 논문에서 사용하는 첫 번째 모의 생성 자료는 두 변수가 양의 상관관계를 갖는 2차원 다변량 정규분포를 하며 공분산행렬이 똑같은 네 집단으로, 한 집단 당 25개씩 총 100 X 2 크기의 행렬로 생성시켰다. 두 번째 자료는 세 변수가 서로 양의 상관관계를 갖는 3차원 다변량 정규분포를 하며 공분산행렬이 똑같은 네 집단의 자료로, 한 집단 당 25개씩 총 100 X 3 크기로 생성시켰다.

본래의 집단 수가 2, 3, 4일 때의 3가지 상황을 연출하여 각각에 대하여 내적 측도들과 외적 측도들을 계산하였다. 본래의 집단 수가 2인 경우에는 확연히 구분되어짐을 한눈에 확인할 수 있도록 하였는데, 한 집단은 고정시키고 다른 한 집단은 평균에 해당하는 좌표(평균좌표)를 가로, 세로 0.6씩 총 10회 증가시켜가면서 고정된 집단으로부터 점점 더 멀어지게 하였다. 본래의 집단 수가 3인 경우에는 명확히 구분되어지는 두 집단을 고정시키고, 그 중 한 집단의 평균좌표에서 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 3 집단 모두가 뚜렷하게 구분 가능해진다. 본래의 집단 수가 4인 경우에는 명확히 구분되어지는 두 집단을 고정시키고, 이들의 평균좌표에서 가로, 세로 0.6씩 총 10회 증가시켜가면서 이 고정된 두 집단들로부터 각각 점점 더 멀어지게 하였다. 마지막에는 고정된 집단들의 평균좌표로부터 가로, 세로 6만큼 떨어져 있는 상태가 되어 4 집단 모두가 뚜렷하게 구분 가능해진다. 아래의 식은 고정된 집단으로부터 점차 이동해 가는 집단의 구조를 간단하게 표현한 것이다. L은 이동해 가는 각 단계를 나타내고, m1은 이동하는 집단의 초기 중심 위치이다.

$$(X_{L1}, X_{L2}) \sim N(\underline{\mu}_L, \Sigma), \qquad L=1,\ldots,10, \quad (17)$$

$$\underline{\mu}_L=\underline{\mu}_1+L\times 0.6\times\begin{pmatrix}1\\1\end{pmatrix}, \qquad \underline{\mu}_1=\begin{pmatrix}\mu_{11}\\ \mu_{12}\end{pmatrix}. \quad (18)$$

다음 그림은 이러한 총 10 단계 중 첫 번째 단계(L=1)와 마지막 단계(L=10)를 보여주고 있다(그림 2-그림 3).

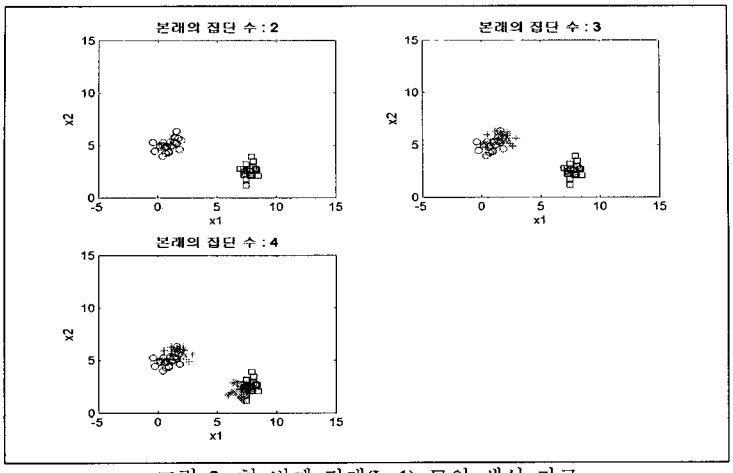

그림 2. 첫 번째 단계(L=1) 모의 생성 자료

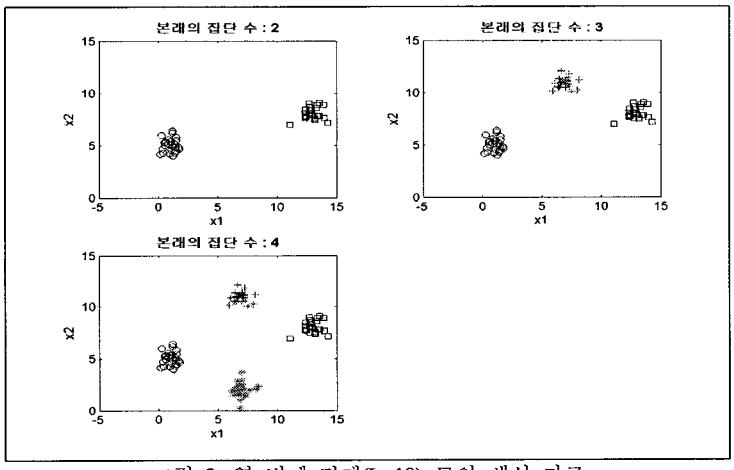

그림 3. 열 번째 단계(L=10) 모의 생성 자료

3.2 실험 방법 및 결과

이렇게 25개씩 4집단짜리 크기 100의 모의 자료를 생성시켜서 최적의 군집 수 예측을 위한 군집 분석을 k=2,…,6에 대해 각각 10번씩 수행한 결과(내적 측도)의 평균을 얻고 다시 예측된 군집 수에서의 군집 타당성분석을 위한 군집 분석을 k=2,…,6에 대해 각각 10번씩 수행한 결과(외적 측도)의 평균을 얻는다. 이 과정을 총 100번 반복하여 얻은 결과의 평균을 계산(한 예로 제시한 것이 표 1이다)한 후, 이들의 본래의 집단 수에 최초 도달한 단계(L)와 본래의 집단 수와 일치하는 단계(L)의 총 횟수를 세어 정리해보았다(표 2).

3.2.1 내적 측도

본래의 집단 수가 3일 때, 자료 형태 변화에 따른 내적 측도 중 본래의 집단 수에 제일 먼저 도

표 1. 2차원 다변량정규분포 모의 생성 자료의 내적 측도(본래의 집단 수 : 4)

		c = 2	c = 3	c = 4	c = 5	c = 6
L=1	Silhouette(GS)	0.8295	0.5433	0.4462	0.4056	0.3839
	Dunn(D)	5.7310	1.3957	1.2699	1.1782	1.1314
	Davies-Bouldin(DB)	0.5961	2.8993	3.2489	3.2237	3.1096
L=2	Silhouette(GS)	0.7868	0.5897	0.4904	0.4396	0.4147
	Dunn(D)	4.6031	1.4529	1.3437	1.1794	1.1434
	Davies-Bouldin(DB)	0.7402	2.3000	2.7168	2.8341	2.7681
L=3	Silhouette(GS)	0.7321	0.6430	0.5503	0.4892	0.4502
	Dunn(D)	3.6954	1.4855	1.4837	1.1415	1.0730
	Davies-Bouldin(DB)	0.9074	1.6020	2.0556	2.2915	2.3581
L=4	Silhouette(GS)	0.6794	0.6815	0.5969	0.5211	0.4780
	Dunn(D)	3.0861	1.5436	1.7288	1.0345	1.0054
	Davies-Bouldin(DB)	1.0775	1.1632	1.6631	2.0040	2.1448
L=5	Silhouette(GS)	0.6241	0.7011	0.6307	0.5587	0.5094
	Dunn(D)	2.6640	1.6047	2.0240	1.0113	0.9766
	Davies-Bouldin(DB)	1.3131	0.9650	1.4025	1.7490	1.9418
L=6	Silhouette(GS)	0.5498	0.7052	0.6617	0.5870	0.5271
	Dunn(D)	2.2351	1.6420	2.5422	1.0326	0.9969
	Davies-Bouldin(DB)	1.9330	1.0575	1.2672	1.6508	1.8802
L=7	Silhouette(GS)	0.5234	0.6973	0.7120	0.6168	0.5457
	Dunn(D)	1.9941	1.6531	3.5201	1.0930	1.0347
	Davies-Bouldin(DB)	1.7674	1.1432	1.0558	1.5388	1.8217
L=8	Silhouette(GS)	0.5208	0.6998	0.7206	0.6325	0.5551
	Dunn(D)	1.9399	1.6755	3.9284	1.1280	1.0512
	Davies-Bouldin(DB)	1.6747	1.1190	1.0306	1.4807	1.7949
L=9	Silhouette(GS)	0.5343	0.7153	0.7075	0.6242	0.5515
	Dunn(D)	2.0497	1.6937	3.6352	1.0682	1.0425
	Davies-Bouldin(DB)	1.6721	1.0343	1.0829	1.5017	1.8057
L=10	Silhouette(GS)	0.5551	0.7325	0.6936	0.6110	0.5489
	Dunn(D)	2.2446	1.6982	3.2816	1.0060	1.0048
	Davies-Bouldin(DB)	1.4951	0.9524	1.1395	1.5658	1.7953

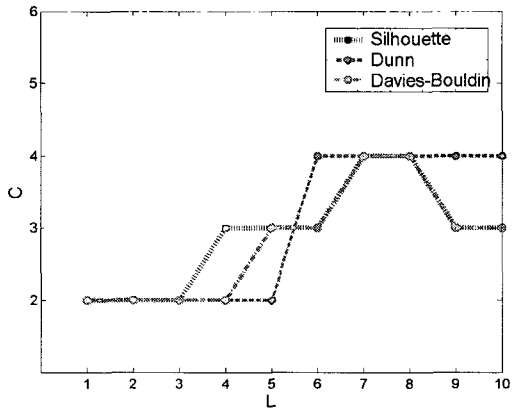

그림 4. 모의 생성 자료의 내적 측도(본래의 집단 수 : 4일 때)

달하는 것은 Dunn 지수였다. 본래의 집단 수가 4일 때 자료 형태 변화에 따른 내적 측도 중 본래의 집단 수에 제일 먼저 도달하는 것 역시 Dunn 지수였고 그 이후에도 변함없는 결과를 보여준 반면, 다른 지수(Silhouette 지수와 Davies-Bouldin 지수)들은 본래의 집단 수에 늦게 도달했을 뿐만 아니라, 그 이후에도 심한 변동이 있었다(그림 4, 표 2). 즉, 겹치는 부분이 완전히 없어지고 확연하게 네 집단으로 구분됨을 얼핏 봐도 알 수 있는 자료 구조에서 최적의 군집 수를 3으로 예측하는 오류를 범한 것을 볼 수 있다.

3.2.2 외적 측도

초반의 겹치는 부분이 상당히 넓은 상태에서는 두 집단의 공분산행렬이 같기 때문인지, 군집화가 잘 이뤄지지 않았다. 본래의 집단 수가 3일 때, (예상했듯이) 두 집단일 때 가장 군집화가 잘 된 것으로 평가했다가 겹치는 정도가 줄어들면서(평균좌표가 멀어지면서) 곧 본래의 집단 수일 때의 군집화 척도가 가장 높아졌다. 본래의 집단 수가 4일 때 역시, (예상했듯이) 겹치는 정도가 아주 높은 초반에는 군집 수 2일 때의 군집화가 가장 잘 된 것으로 평가했고, 점차 겹치는 정도가 줄어들면서(평균좌표가 멀어지면서) 군집 수 3일 때의 군집화 척도가 가장 높은 현상을 잠깐 보이다가 곧 본래의 집단 수일 때의 군집화 척도가 가장 높아졌다. 중간에 군집 수 3일 때의 최적 군집화 평가는 2개의 고정 집단들로부터 멀어지는 두 집단의 생성 위치(분포 위치)에 차이가 생기기 때문일 것으로 추론해 봄직하다. 게다가 본래의 집단 수 3일 때와는 달리, 본래의 집단 수 4보다 큰 군집 수 5일 때를 가장 잘 된 군집으로 평가한 경우가 간간이 나타나는 양상을 보였다. 실험에 이용한 외적 측도 중 가장 뛰어난 지수를 가리는 기준을 본래의 집단 수일 때를 최고 성능 점수를 계산해낸 횟수로 정하면, Jaccard 지수 : 7회, Goodman-Kruskal 지수 : 5회, Rand 수정 지수 : 2회 순이었다.

표 2. 2차원 다변량정규분포 모의 생성 자료 실험 결과

		본래의 집단 수:2		본래의 집단 수:3		본래의 집단 수:4	
		L*	#_L@	L*	#_L@	L*	#_L@
내적 측도	Silhouette(GS)	1	10	6	5	7	2
	Dunn(D)	1	10	5	6	7	4
	Davies-Bouldin(DB)	1	10	6	5	7	2
외적 측도	Jaccard(J)	1	10	2	9	2	7
	Goodman-Kruskal(GK)	1	10	1	9	3	5
	Adjusted Rand(ARI)	1	10	1	9	2	2

L* : 본래의 집단 수에 최초 도달한 단계(L)
#_L@ : 본래의 집단 수와 일치하는 단계(L)의 총횟수

3.2.3 3차원 다변량정규분포 모의 생성 자료

2차원 자료일 때와 거의 똑같은 결과임을 확인하였다.

5. 유전자 발현 자료에의 군집 타당성분석 기법 적용

5.1 유전자 발현 자료

본 논문에서 사용하는 유전자 발현 자료는 원래 똑같은 원 자료(raw data)가 다른 전처리 과정을 거쳐 유전자(gene)의 개수가 두 가지 버전(version)인 백혈병 자료이다.

5.1.1 첫 번째 백혈병 자료

첫 번째 자료는 다소 간단한 전처리 과정을 거쳐 유전자의 수가 3051개의 발현 정도로 표시되는 총 38개의 샘플로 구성돼 있다(Fort, Lambert-Lacroix, 2005). 이 중 11 샘플은 급성 림프모구 백혈병 - acute myeloid leukaemia (AML)이고, 27 샘플은 급성 골수성 백혈병 - acute lymphoblastic leukaemia (ALL)로, 본래의 집단 수가 2인 유전자 발현 자료이다.

표 3. 첫 번째 백혈병 자료(본래 집단의 수 : 2)의 타당성 지수

타당성 지수		c=2	c=3	c=4	c=5	c=6
내적 측도	GSu 지수	0.1458	0.1501	0.1237	0.1386	0.1905
	Dunn 지수	1.1560	1.1043	0.9959	0.9235	0.9179
	D-B 지수	2.4490	2.3146	2.1703	2.1075	2.0051
외적 측도	Jaccard 지수	0.6384	0.4867	0.4114	0.3563	0.3141
	G-K 지수	0.4200	0.4131	0.3866	0.4161	0.3968
	AR 지수	0.3881	0.3059	0.2650	0.2639	0.2348

표 4. 두 번째 백혈병 자료(본래 집단의 수 : 3)의 타당성 지수

타당성 지수		c=2	c=3	c=4	c=5	c=6
내적 측도	GSu 지수	0.2086	0.2407	0.1908	0.1680	0.1591
	Dunn 지수	1.1782	1.2931	1.1003	1.0611	1.0319
	D-B 지수	2.1091	1.9068	2.0097	2.0225	2.0054
외적 측도	Jaccard 지수	0.6190	0.8632	0.7030	0.5580	0.4426
	G-K 지수	0.7040	0.9068	0.8119	0.7110	0.6393
	AR 지수	0.5761	0.8739	0.7457	0.6087	0.4932

각각의 k=2,…,6에 대하여 최적의 군집 수 예측을 위한 군집 분석을 1000번씩 반복 수행한 결과(내적 측도)의 각 거리법에 대한 값과 각각의 k=2,…,6에 대하여 예측된 군집 수에서의 군집 타당성분석을 위한 군집 분석을 1000번씩 반복 수행한 결과(외적 측도)의 각 거리법에 대한 값을 얻어 평균을 계산했다.

내적 측도 : Silhouette 지수와 Davies-Bouldin 지수는 최적의 군집 수를 6으로 예측했으나, Dunn 지수만은 2개로 정확하게 예측했다(표 3).

외적 측도 : 세 가지 지수 모두 군집 수 2일 때가 가장 군집화가 잘 된 것으로 평가했다(표 3).

5.1.2 두 번째 백혈병 자료

두 번째 자료는 첫 번째 자료에서와는 다른 전처리 과정을 거쳐 유전자의 수가 100개의 발현 정도로 표시되는 총 38개의 샘플로 구성돼 있다(Handl, Knowles, Kell, 2005). 이 자료는 첫 번째 자료에서의 ALL을 두 집단으로 세분화하여, 본래의 집단 수가 총 3이다. 실험 방법은 첫 번째 자료에서와 동일하다.

내적 측도 : 세 지수 모두 최적의 군집 수를 3으로 정확하게 예측했다(표 4).

외적 측도 : 세 지수 모두 군집 수 3일 때가 가장 군집화가 잘 된 것으로 평가했다(표 4).

6. 결론

모의 생성 자료 실험에서는 세 가지 내적 측도 중 Dunn 지수가 Silhouette이나 Davies-Bouldin 지수보다는 최적의 군집 수 예측력이 다소 뛰어나며 강력(robust)함을 확인할 수 있었고, 세 가지 외적 측도는 실험 전에 성능 비교 목적은 없었으나 이들 역시 성능을 비교할 만한 결과가 도출되었는데, Jaccard 지수가 Goodman-Kruskal 지수나 Rand 수정 지수보다는 조금 더 나은 면을 가지고 있다고 결론지을 수 있겠다. 추가로, 두 집단이 상당히 많이 겹쳐 있는 상황에서는 외적 측도들까지도 K = 본래의 집단 수로 군집 분석을 했을 때 이외에 더 높은 값이 나온 것은, 실험에서 사용한 K-평균 군집 알고리즘이 이런 상황에서는 군집화를 썩 잘 해내지 못 한다고 볼 수 있다.

모의 실험 자료를 생성하는 과정에서 고정된 집단으로부터 점차 멀어지는 방향으로 이동하는 그 증분을 0.6에서 더 작게 설정하면, 보다 정밀한 결과를 얻을 수 있을 것이다. 그리고, 본래의 집단 수가 4일 때, 겹치는 부분이 완전히 없어지고 확연하게 네 집단으로 구분됨을 얼핏 봐도 알 수 있는 자료 구조에서 다른 지수(Silhouette 지수와 Davies-Bouldin 지수)들은 최적의 군집 수를 3으로 예측하는 오류를 범하는 것에 대해서 이유를 밝혀보는 것도 의미가 있을 듯하다.

실제 자료인 유전자 발현 자료 중 첫 번째 백혈병 자료에서는 세 가지 내적 측도 중 Dunn 지수만이 군집 수 예측을 정확하게 해냈고, 세 가지 외적 측도 모두 본래의 집단 수 2일 때를 군집화가 가장 잘 되었다고 판단해 주었다. 두 번째 유전자 발현 자료에서는 내적 측도와 외적 측도 총 여섯 가지 모두 본래의 집단 수 3일 때가 최적으로 군집화 되었다는 결과를 내주었다.

두 번째 유전자 발현 자료는 첫 번째 자료에 비해 잡음(noise)이 많이 제거됐거나, 세 집단(AML, B-ALL, T-ALL)으로 분류되도록 특징 추출(feature selection)이 이뤄지지 않았을까 하는 추론을 조심스럽게 내리면서, 특히 차원이 높은 데이터일수록 차원 축소 내지 특징 추출 단계가 매우 중요한 의미를 가지고 있음을 되새겨 보았다.

참고문헌

[1] Handl, J., Knowles, J. and Kell, D. B. (2005). Computational cluster validation in post-genomic data

analysis, *Bioinformatics,* Vol. 21, 3201-3212.

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

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

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

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

[6] Hubert, L. and Schultz, J. (1976). Quadratic assignment as a general data-analysis strategy, *The British journal of mathematical & statistical psychology,* Vol. 29, 190 -241.

[7] Jaccard, P. (1912). The distribution of flora in the alpine zone, *New Phytologist,* Vol. 11, 37-50.

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

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

[10] Hubert, L. and Arabie, P. (1985). Comparing partitions, *Journal of Classification,* Vol. 2, 193-218.

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

[12] Bolshakova, N. and Azuaje, F. (2003). 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.

[13] Bolshakova, N. and Azuaje, F. (2003). Cluster validation techniques for genome expression data classification, *Signal Processing,* Vol. 83, 825-833.

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

Comparison of the Cluster Validation Techniques using Gene Expression Data

YunKyoung Jeong[1], *JangSun Baek*[2]

Abstract

Several clustering algorithms to analyze gene expression data and cluster validation techniques that assess the quality of their outcomes, have been suggested, but evaluations of these cluster validation techniques have seldom been implemented. In this paper we compared various cluster validity indices for simulation data and real genomic data, and found that Dunn's index is more effective and robust through small simulations and with real gene expression data.

[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