응용통계연구 (2009)
22(4), 745–758

군집분석 비교 및 한우 관능평가데이터 군집화

김재희[1] · 고윤실[2]

[1]덕성여자대학교 정보 통계학과, [2]덕성여자대학교 정보 통계학과

(2009년 4월 접수, 2009년 7월 채택)

요 약

자발적인 군집을 유도하는 다변량 통계기법으로 널리 사용되는 군집분석은 데이터에 기반한 탐색적 방법으로 쓰이며 군집원칙에 따라 여러 가지 방법이 제안되어 왔다. 또한 군집화된 결과에 대하여 유효성을 측정하는 측도도 다양한 방법이 개발되었다. 본 연구에서는 계층적 군집분석 방법으로 최장연결법과 Ward의 방법, 비계층적 군집분석 방법으로 K-평균법 그리고 확률분포정보를 활용한 모형기반 군집분석방법을 이용하여 모의실험으로 군집분석을 실시하고 군집유효성 측도로는 연결성, Dunn 지수, 실루엣을 구하여 각 군집방법에 대해 유효성을 비교한다. 또한, 한우 관능평가 데이터에 군집분석을 적용하여 최적의 군집 상황을 구하고자 한다.

주요용어: 모형기반 군집분석, 연결성, 실루엣, 한우 관능평가데이터, Average Distance(AD), Average Proportion(APN), Dunn 지수, K-평균법, Ward 방법.

1. 서론

군집분석은 데이터에 기반한 탐색적 방법으로 매우 다양한 방법이 제안되어 왔다. 직관적인 방법뿐만 아니라 최근에는 확률분포를 고려한 모형도 제안되었으며 유전자 발현 데이터를 포함한 여러 분야의 데이터에 군집분석이 활용되고 있다. 그러나 어느 한개의 군집분석 방법이 우수하다고 볼 수 없으며 데이터의 상황에 맞게 적절한 방법을 통해 적절한 군집 개수와 군집을 도출해 내야한다.

본 연구에서는 네 가지 널리 쓰이는 군집분석 방법과 군집유효성기법을 적용하여 비교하고 특성을 살펴보고자 한다. 또한 실제 데이터 분석으로 한우 관능평가 데이터에 각 군집분석을 적용해보고 최적의 군집 상황을 구하고 각 군집의 특성을 파악하고자한다.

2. 네 가지 군집화 방법

대량의 데이터가 있을 때 개체 간 유사성이 높은 것들로 군집화 하여 군집간의 특성을 비교하고자 할 때 군집화의 방법은 지금까지 다양한 방법들이 제안되었다. 그 중 계층적 군집분석 방법으로 군집간 거리를 이용한 최장연결법, 군집내 제곱합을 최소로 하는 Ward 방법과 군집 개수를 정한 후 비계층적 방법으로 널리 적용되고 있는 K-평균법 그리고 확률분포에 대한 정보가 있는 경우 확률모형을 고려한 모형기반 군집 분석 방법을 설명하고 모의실험을 통해 특징을 파악하고자 한다.

p-차원의 관측벡터 n개를 $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$이라 하자.

본 연구는 2009년 덕성여자대학교 연구비 지원을 받았습니다.

[1]교신저자: (132-714) 서울시 도봉구 쌍문동 419, 덕성여자대학교 정보 통계학과, 교수.
E-mail: jaehee@duksung.ac.kr

최장연결법에서는 두 군집 CL_1과 CL_2의 거리는 군집에 속한 개체 간의 최장 거리

$$d\{CL_1, CL_2\} = \max\{d(\mathbf{x}_1, \mathbf{x}_2) | \mathbf{x}_1 \in CL_1, \mathbf{x}_2 \in CL_2\} \tag{2.1}$$

로 정의한다. 여기서 개체 간의 거리를 구하는 방법으로는 유클리드 거리, 표준화 거리(통계적 거리), 마할라노비스 거리 등을 이용하여 구할 수 있으며 거리 함수 선택이 결과에 영향을 미칠 수 있다.

Ward (1963)는 objective function인 편차제곱합 ESS(error sum of squares)을 고려하여 군집의 개수를 선택하게 된다. 모든 개체가 각각 군집이라고 보고 한 개씩 군집의 수를 줄여가면서 편차제곱합을 계산해 나가는데, 군집의 수가 1이 될 때까지 그 과정을 체계적으로 반복하는 것이며 군집내 제곱합 증분과 군집간 제곱합을 고려하여 편차제곱합이 최소가 되는 군집의 개수를 선택하게 된다.

K-평균법 (Hartigan과 Wong, 1979)은 모든 개체를 K개의 군집으로 나누는 방법으로 군집간 개체의 재배치가 가능한 군집방법이다. p-차원 관측벡터 $\mathbf{x}_1, \ldots, \mathbf{x}_n$에 대해

$$W_n = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \le J \le K} \|\mathbf{x}_i - \mathbf{a}_J\| \tag{2.2}$$

를 만족하도록 군집중심인 $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k$를 구한다. 각 $\mathbf{x}_1, \ldots, \mathbf{x}_n$을 $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k$중 가장 가까운 군집으로 분류한다. 여기서 $\|\cdot\|$는 유클리드 거리를 나타낸다. 이러한 과정의 반복 계산을 통해 군집내 제곱합을 최소화하는 방향으로 최종 군집이 선택된다. Pollard (1981, 1982)는 표본의 크기가 충분히 크면 K-평균법에 의해 구한 K개 군집 평균은 참값으로 almost sure convergence 함을 보였고 그러한 경우 optimal split이 된다고 할 수 있다.

앞에서 열거한 세 가지의 군집화 방법(최장연결법, Ward의 방법, K-평균법)은 직관적으로 합리적인 방법이긴 하나 확률분포적 모형 없이 군집화가 이루어지는 방법이므로 확률분포에 관한 정보가 있을 경우에는 이에 대한 정보를 이용하는 것이 바람직하다. 그러나 모형에 근거를 둔 군집화 방법은 여러 가지 장점을 갖고 있고 다양한 가능성이 제안되어 왔다. Scott과 Symons (1971)가 제안하고 Banfield와 Raftery (1993), Fraley와 Raftery (2002) 등이 발전시킨 방법으로 모형기반 군집화 방법을 설명하고자 한다.

모집단이 G개의 군집으로 구성되어 있으며 k번째 군집에 속한 p-차원의 관측벡터 $\mathbf{x}$의 밀도함수는 $f_k(\mathbf{x}, \boldsymbol{\theta})$라고 가정하고 $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)'$라 한다. 여기서 $\mathbf{x}_i$가 k번째 군집에 속하였으면 $\gamma_i = k$이다.

데이터가 속한 군집은 내재하는 확률분포로부터 형성되었다고 가정하고 다음의 혼합 모형(mixture model)의 가능도함수

$$L_{mix}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(\mathbf{x}_i | \boldsymbol{\theta}_k) \tag{2.3}$$

를 최대화하도록 모수를 추정하게 된다. 여기서 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_G)'$는 모수벡터이고 τ_k는 관측벡터가 k번째 군집에 속할 확률이며 $\tau_k \ge 0$, $\sum_{k=1}^{G} \tau_k = 1$이다.

관측벡터가 다변량 정규분포 $f_k(\mathbf{x}, \boldsymbol{\theta})$를 따른다고 가정하는 Gaussian 혼합모형을 고려하자. 평균벡터 $\boldsymbol{\mu}_k$와 공분산행렬 $\boldsymbol{\Sigma}_k$를 갖는 k번째 다변량 정규밀도함수일 때 가능도함수는

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \text{const} \prod_{k=1}^{G} \prod_{i \in E_k} |\boldsymbol{\Sigma}_k|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\}, \tag{2.4}$$

여기서 $E_k = \{i; \gamma_i = k\}$이다. 각 군집을 형성하는 기하학적인 특징(shape, volume, orientation)은 공분산행렬 $\boldsymbol{\Sigma}_k$에 의해 결정되는데 Banfield와 Raftery (1993)는 고유값 분해에 의하여 공분산행렬이 대

표성을 갖는 다음과 같은 일반적인 구조(general framework)를 제안했다.

$$\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k', \tag{2.5}$$

여기서 $\mathbf{D}_k$는 고유벡터의 직교행렬이고 $\mathbf{A}_k$는 각 원소가 $\boldsymbol{\Sigma}_k$의 고유값을 비례적으로 취하는 대각행렬이며 λ_k는 스칼라이다. $\mathbf{D}_k$는 군집의 orientation을 결정하고 $\mathbf{A}_k$는 군집의 shape을 결정하며 λ_k는 군집의 volume을 결정한다. Yeung 등 (2001)은 공분산행렬의 형태에 따라 5가지 경우에 대한 군집의 orientation과 shape을 정리하였다.

예상되는 군집 개수 G가 정해지면, 가능한 군집 개수 $1 \leq k \leq G$에 대해 $\tau_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$가 EM 알고리듬에 의해 추정된다. EM 알고리듬은 E(expectation) 단계와 M(maximization) 단계로 이루어지며, E 단계에서는 주어진 조건하에서 관측벡터가 각 군집에 속할 확률을 구하고 M 단계에서는 주어진 상황에서 모수가 추정된다. 각 개체가 최대 확률로 해당 그룹에 할당될 때 EM 알고리듬 결과로 수렴하게 된다. 모형 선택시 BIC(Bayesian Information Criterion)를 계산하여 BIC 값이 최대가 되는 군집 개수를 최종 모형으로 선택할 수 있다. 이와 같은 계산과정은 R 프로그램에서 MCLUST (Fraley와 Raftery, 1998) 패키지로 제공되고 있으며 본 논문에서도 이용하고자한다.

3. 군집 유효성

3.1. 내부 유효성 측도(Internal validity measure)

원래의 데이터로서 군집화된 결과만을 가지고 데이터 본래의 정보를 사용하여 군집화를 얼마나 잘 실행하였는지 평가하려는 것으로 그 측도로는 Connectivity (Handl 등, 2005), Dunn Index (Dunn, 1974), Silhouette Width (Rousseeuw, 1987) 등이 있다. 이 절에서는 이 세 가지 측도에 대해 설명하고 특징을 파악하고자한다.

n개의 개체가 G개의 집단으로 군집화된 경우의 연결성 측도(connectivity)의 정의는

$$\text{Conn}(\mathcal{C}) = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{i,nn_{i(j)}} \tag{3.1}$$

이다. 여기서 $\mathcal{C} = C_1, C_2, \ldots, C_G$, p는 개체로부터 측정되는 변수의 수이며 $nn_{i(j)}$는 개체 i로부터 j번째 가까이 위치한 개체를 의미한다. $x_{i,nn_{i(j)}}$는 i와 j가 같은 군집에 있으면 0이고 다른 군집에 있으면 1이 된다. 즉 연결성 측도는 어떤 개체가 그 개체와 가까운 거리에 있는 개체들과 얼마나 같은 군집에 배치되어 있는지를 알 수 있게 해주며 작은 값을 가질수록 군집화가 잘 되었다고 판단한다.

Dunn 지수(Dunn Index)는 같은 군집에 속해 있는 두 개체간의 가장 큰 거리에 대한 서로 다른 군집에 속해 있는 두 개체간의 가장 작은 거리의 비(ratio)를 나타내며 다음과 같이 계산할 수 있다.

$$D(\mathcal{C}) = \frac{\underset{(i \in C_k, j \in C_l) \in \mathcal{C},\ C_k \neq C_l}{\text{MIN}} \left(\underset{}{\text{MIN}}\ \text{dist}(i,j) \right)}{\underset{C_m \in \mathcal{C}}{\text{MAX}}\ \text{diam}(C_m)}, \tag{3.2}$$

여기서 C_m은 $\mathcal{C}$의 분할에서 거리가 가장 큰 두 개체가 있는 군집이며 $\text{diam}(C_m)$은 군집 C_m에서 가장 큰 두 개체의 거리를 나타낸다.

같은 군집에 속해 있는 두 개체간의 거리가 작을수록, 다른 군집에 속해 있는 두 개체간의 거리가 클수록 Dunn 지수는 커지므로 이 수치가 클수록 군집화가 잘 되었다고 판단할 수 있다.

Kaufman과 Rousseeuw (1990)는 Silhouette Width를 제안하고 silhouette value를 근거로 최소가 되는 군집 개수를 선택하도록 제시하였다.

개체의 silhouette value는

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \tag{3.3}$$

로 정의되며 여기서 a_i는 개체 i가 속한 있는 군집의 모든 개체들과 개체 i와의 평균거리이고 b_i는 개체 i가 속하지 않은 군집의 모든 개체들과 개체 i와의 평균거리이다. silhouette width는 각 개체의 silhouette value의 평균을 나타내며 silhouette value는 개체가 적절한 군집에 배치되었는지 측정하는 신뢰도의 개념이라고 말할 수 있다. 이 측정값들은 각 개체가 적절한 군집에 배치될수록 1에 가깝고 그렇지 않을수록 -1에 가깝다. 모든 개체에 대하여 silhouette value를 구하므로 그 평균인 silhouette width뿐만 아니라, silhouette value의 중위수, 표준편차 등 더 많은 정보를 얻을 수 있으므로 연결성 측도나 Dunn 지수보다 자세한 정보를 얻을 수 있다.

3.2. 안정성 측도(Stability measure)

원래의 데이터로 군집화된 결과와 측정변수 한 개를 제거한 후 군집화된 결과를 비교하여 군집의 안정성을 평가하려는 것이며 이 측도는 특히 데이터 간에 상호관련성이 상당히 높을 때 유용하게 쓰인다. 내부 유효성 측도의 특별한 경우라 볼 수 있으며 군집화된 결과의 일관성을 평가한다. 측도의 종류로는 APN, AD, ADM, FOM (Datta와 Datta, 2003; Yeung 등, 2001) 등이 있으며 여기서는 APN과 AD를 다루고자 한다.

원래의 데이터로 군집화된 결과와 측정변수 한 개를 제거한 후 군집화된 결과를 비교하여 같은 군집에 배치되지 않은 개체들의 평균비율(average proportion)을 측정한 통계량인 APN 측도는 다음과 같이 정의된다.

$$\text{APN} = \frac{1}{np}\sum_{i=1}^{n}\sum_{l=1}^{p}\left(1 - \frac{n(C^{i,l} \cap C^{i,0})}{n(C^{i,0})}\right), \tag{3.4}$$

여기서 $C^{i,l}$은 l번째 측정변수를 제거한 후 군집화된 결과에서 개체 i를 포함하는 군집이고 $C^{i,0}$는 원래의 데이터로 군집화된 결과에서 개체 i를 포함하는 군집이다. APN은 0부터 무한히 큰 수까지 나타날 수 있으며 그 수치가 작을수록 변수제거 전후의 군집간 개체 이동성이 작다는 의미이다. 따라서 APN은 0에 가까울수록 군집화 결과가 상당히 일관성이 있다고 판단한다.

원래의 데이터로 군집화된 결과에서 개체 i와 같은 군집에 위치해 있는 모든 개체들과 측정변수 한 개를 제거하고 군집화된 결과에서 개체 i와 같은 군집에 위치해 있는 모든 개체들과의 평균거리를 측정하여 나타내는 AD(Average Distance) 측도는

$$\text{AD} = \frac{1}{np}\sum_{i=1}^{n}\sum_{l=1}^{p}\frac{1}{n(C^{i,l} \cap C^{i,0})}\left[\sum_{i \in C^{i,0}, j \in C^{i,l}} \text{dist}(i,j)\right] \tag{3.5}$$

으로 정의된다.

AD는 0부터 무한히 큰 수까지 나타날 수 있으며 그 수치가 작을수록 변수제거 전후의 개체 i를 포함하는 두 군집간 평균거리가 작음을 의미하므로 AD값이 0에 가까울수록 군집화의 결과가 일관성이 있다고 판단할 수 있다.

군집화의 결과에 대한 안정성을 평가하기 위한 두 측도, APN과 AD는 원래의 데이터로 군집화된 결과와 변수를 한 개씩 제거해 나가면서 군집화된 결과들을 비교하여 그 수치로서 군집의 안정성을 평가한다

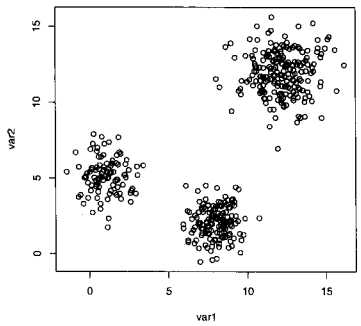

그림 **4.1.** 발생 데이터

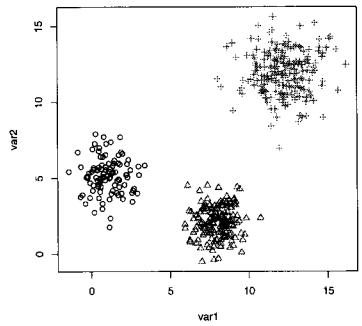

그림 **4.2.** 모형기반군집방법에 의한 군집화

는 점에서 동일하다고 볼 수 있다. 그러나 APN은 군집의 유사성을 군집간 개체의 이동성으로서 측정하고 AD는 군집의 유사성을 거리로서 측정한다는 면에서 두 측도는 차이가 있다.

4. 모의실험

군집의 개수를 알고 있는 몇 가지 군집 형태를 갖도록 데이터를 발생시킨 후 네 가지 군집방법을 적용하고 그 결과에 대해 내부 유효성 측도를 계산하고 비교한다. 모의실험은 R 프로그램을 이용하며 clValid (Brock 등, 2008) 패키지에 내장된 함수를 이용하여 군집의 유효성 측도를 계산한다.

4.1. 세 개의 군집

세 개의 군집에서 각각 2-변량 정규분포를 따르는 관측벡터 $n_1 = 100$, $n_2 = 150$, $n_3 = 200$개를 그림 4.1과 같이 발생시킨다. $\mathbf{X}_{i1} \overset{iid}{\sim} N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $i = 1, \ldots, n_1$, $\mathbf{X}_{i2} \overset{iid}{\sim} N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, $i = 1, \ldots, n_2$, $\mathbf{X}_{i3} \overset{iid}{\sim} N_2(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$, $i = 1, \ldots, n_3$ 여기서

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1 \\ 5 \end{pmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 8 \\ 2 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{\mu}_3 = \begin{pmatrix} 12 \\ 12 \end{pmatrix}, \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

표 4.1을 보면 연결성 측도는 네 가지의 군집화 방법에서 2개 또는 3개로 군집화 했을 때 0으로 최적값을 나타냈고 Dunn 지수와 실루엣은 네 가지의 군집화 방법에서 3개로 군집화 했을 때 최적값을 나타냄을 알 수 있다. 그림 4.2를 보면 모형기반 군집방법에 의하여 3개로 군집화한 결과 원래의 군집대로 잘 군집화 되었다.

4.2. 공분산이 존재하는 두 개의 군집-이변량

두 개의 군집에서 각각 2-변량 정규분포를 따르는 관측벡터 $n_1 = 200$, $n_2 = 300$개를 그림 4.3과 같이 발생시킨다.

$$\mathbf{X}_{i1} \overset{iid}{\sim} N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \quad i = 1, 2, \ldots, n_1, \qquad \mathbf{X}_{i2} \overset{iid}{\sim} N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \quad i = 1, 2, \ldots, n_2,$$

여기서

표 **4.1.** 세 집단 이변량 데이터의 내부유효성 측정값

군집방법	측정방법	군집의 개수				
		2	3	4	5	6
complete	connectivity	**0.0000**	**0.0000**	21.3361	35.2821	44.0226
	Dunn	0.2922	**0.3482**	0.0334	0.0407	0.0440
	silhouette	0.6832	0.7625	0.5358	0.5174	0.5101
Ward	connectivity	**0.0000**	**0.0000**	25.4286	37.5147	58.1000
	Dunn	0.2922	**0.3482**	0.0384	0.0455	0.0298
	silhouette	0.6832	0.7625	0.5608	0.5621	0.4050
K-means	connectivity	**0.0000**	**0.0000**	28.1964	39.4758	63.3190
	Dunn	0.2922	**0.3482**	0.0193	0.0309	0.0231
	silhouette	0.6832	0.7625	0.5648	0.5743	0.4282
model	connectivity	**0.0000**	**0.0000**	8.5925	51.1937	53.2802
	Dunn	0.2922	**0.3482**	0.0454	0.0155	0.0257
	silhouette	0.6832	0.7625	0.5835	0.5653	0.5478

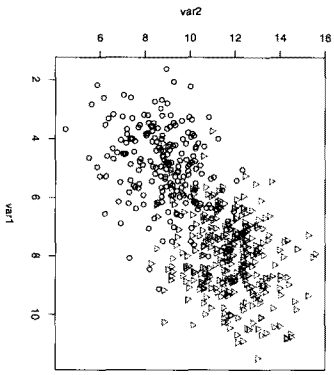

그림 **4.3.** 공분산존재 발생데이터

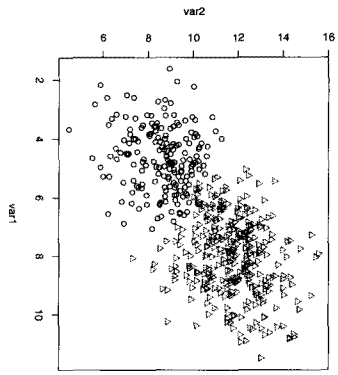

그림 **4.4.** Ward 방법에 의한 군집화

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5 \\ 9 \end{pmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 8 \\ 12 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix}.$$

군집의 개수에 따른 군집방법별 내부 유효성 측정값을 나타낸 표 4.2를 보면 연결성 측도는 Ward 방법을 이용하여 2개로 군집화한 경우 최적값을 나타내고 Dunn 지수는 Ward 방법을 이용하여 2개, 5개와 6개로 군집화한 경우 동일하게 최적값을 보이며 실루엣은 모형기반 군집 방법을 이용하여 2개로 군집화한 경우에 최적값을 나타냄을 알 수 있다. 한편 모형기반 군집방법에 의한 2개의 군집화 결과에 대한 실루엣 값 0.5130는 Ward 방법에 의한 2개의 군집화 결과에 대한 실루엣 값 0.5046와 비슷하다. 따라서 3가지 내부 유효성 측도 모두 Ward 방법에 의한 2개의 군집화(그림4.4)를 추천한다.

4.3. 두 개의 타원 형태 군집인 경우

큰 타원 형태의 군집 안에 작은 타원 형태의 군집이 형성되도록 그림 4.5과 같이 데이터를 발생시킨다. 군집의 개수에 따른 군집방법별 내부 유효성 측정값을 나타낸 표 4.3을 보면 연결성 측도와 Dunn 지수는 모형기반 군집방법을 이용하여 2개로 군집화한 경우 최적값을 보여주며, 실루엣은 K-평균법을 이용

표 **4.2.** 공분산이 존재하는 두 집단 이변량 데이터의 내부유효성 측정값

군집방법	측정방법	군집의 개수				
		2	3	4	5	6
complete	connectivity	31.7599	49.1698	59.3921	76.0913	83.1313
	Dunn	0.0214	0.0292	0.0305	0.0236	0.0267
	silhouette	0.3809	0.3271	0.2980	0.2673	0.2756
Ward	connectivity	**14.0635**	34.7377	45.231	62.6119	72.8897
	Dunn	**0.0320**	0.0280	0.0282	0.0320	0.0320
	silhouette	**0.5046**	0.3693	0.3708	0.3233	0.2697
K-means	connectivity	21.1321	47.7881	58.8151	94.6825	99.3794
	Dunn	0.0261	0.0103	0.0179	0.0121	0.0193
	silhouette	0.5122	0.3732	0.3858	0.3529	0.3437
model	connectivity	22.1984	54.9940	81.7361	99.1417	104.2540
	Dunn	0.0030	0.0163	0.0126	0.0039	0.0291
	silhouette	0.5130	0.3765	0.3134	0.3141	0.3170

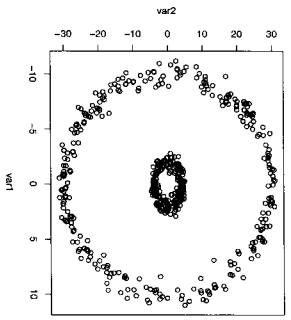

그림 **4.5.** 타원발생데이터

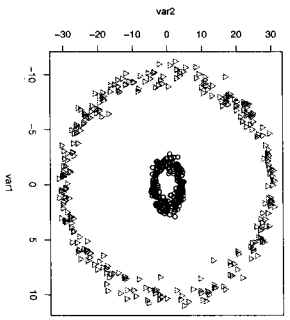

그림 **4.6.** 모형기반군집화

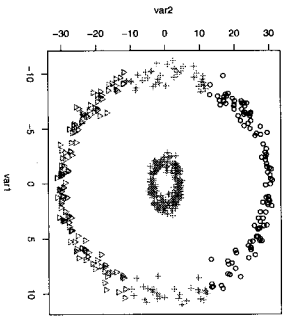

그림 **4.7.** K-평균법

표 **4.3.** 타원 형태 두 집단 데이터의 내부유효성 측정값

군집방법	측정방법	군집의 개수				
		2	3	4	5	6
complete	connectivity	5.8246	15.1603	16.7171	16.7171	23.5770
	Dunn	0.0499	0.0384	0.0441	0.0538	0.0562
	silhouette	0.4876	0.5510	0.5175	0.5195	0.4772
Ward	connectivity	6.3071	7.1417	13.4008	14.1230	14.1230
	Dunn	0.0303	0.0350	0.0381	0.0420	0.0550
	silhouette	0.5067	0.5437	0.5068	0.5049	0.5379
K-means	connectivity	6.5373	21.5698	26.4881	24.5206	22.1579
	Dunn	0.0379	0.0316	0.0237	0.0237	0.0398
	silhouette	0.5349	**0.5995**	0.5358	0.5115	0.5418
model	connectivity	**0.0000**	11.2694	18.5786	17.4079	25.3952
	Dunn	**0.0955**	0.0184	0.0083	0.0126	0.0122
	silhouette	0.1916	0.3547	0.3811	0.4398	0.3327

하여 3개로 군집화한 경우에 최적값을 나타낸다. 그림 4.6은 모형기반 군집방법에 의해 2개로 군집을 형성한 결과이며 그림 4.7은 K-평균법에 의해 3개의 군집을 형성한 결과이다.

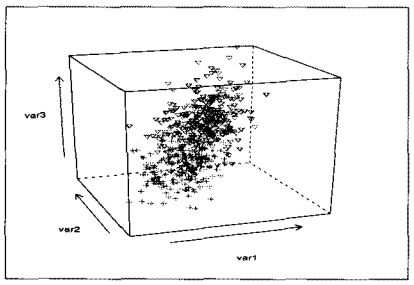

그림 **4.8.** 삼변량 두 개 군집 발생데이터

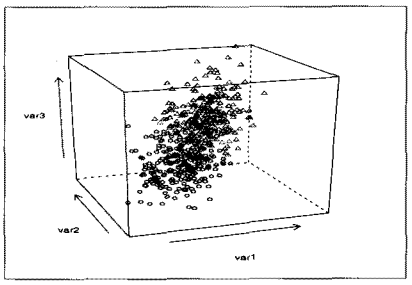

그림 **4.9.** K-means 군집화

표 **4.4.** 공분산 존재하는 두 집단 삼변량 데이터의 내부유효성 측정값

군집방법	측정방법	군집의 개수				
		2	3	4	5	6
complete	connectivity	33.5460	78.2321	109.3560	129.8830	150.1910
	Dunn	0.0334	0.0357	0.0366	0.0425	0.0447
	silhouette	0.3503	0.2699	0.2553	0.2690	0.2553
Ward	connectivity	**20.2353**	57.1421	90.8675	107.0910	127.1870
	Dunn	0.0548	0.0420	0.0310	0.0396	0.0396
	silhouette	0.4447	0.2810	0.2547	0.2623	0.2345
K-means	connectivity	39.9786	91.4115	105.9960	118.9840	141.8080
	Dunn	0.0271	0.0285	0.0245	0.0360	0.0387
	silhouette	**0.4507**	0.3305	0.2823	0.3093	0.3022
model	connectivity	39.6512	119.6650	134.6150	126.0160	150.2630
	Dunn	0.0187	0.0124	0.0138	0.0276	0.0292
	silhouette	0.4101	0.2857	0.2879	0.2740	0.2506

4.4. 공분산 존재하는 두 개 군집인 경우-삼변량

공분산이 존재하는 두 집단 3-변량 데이터 $n_1 = 200$, $n_2 = 300$개를 그림 4.8과 같이 발생시킨다.

$$\mathbf{X}_{i1} \overset{iid}{\sim} N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \quad i = 1, 2, \ldots, n_1, \qquad \mathbf{X}_{i2} \overset{iid}{\sim} N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \quad i = 1, 2, \ldots, n_2,$$

여기서

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 5 \\ 9 \\ 14 \end{pmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 2 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 3 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 8 \\ 12 \\ 17 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 4 \end{pmatrix}.$$

표 4.4를 보면 실루엣은 K-평균법을 이용하여 2개로 군집화한 경우에 최적값을 나타냄을 알 수 있다. 그림 4.9는 K-평균법에 의해 2개의 군집을 형성한 결과이다. 또한 각 측도가 추천하는 군집방법이 일치하지 않을 수 있으므로 군집의 특성을 파악하여 군집 방법을 선택해야한다.

모의실험결과를 전반적으로 살펴보면 그림 4.1과 같이 군집분리가 명백한 경우에는 군집방법 선택에 영향을 받지않으나 그림 4.3의 경우에는 연결성 측도와 Dunn 지수를 동시에 고려하면 Ward 방법이 선택된다. 그림 4.5와 같이 타원형 군집을 가진 경우에는 다변량 정규분포 모형 기반 군집방법이 선택되나 그림 4.8의 경우 실루엣 값을 고려하면 K-means 방법을 선택할 수도 있게 된다. 이와 같이 내부 유효

표 5.1. 한우관능평가데이터의 내부유효성 측정값

군집방법	측정방법	군집의 개수				
		2	3	4	5	6
complete	connectivity	**83.3671**	207.2944	219.6139	238.3583	351.7837
	Dunn	0.0341	0.0355	0.0373	**0.0401**	0.0365
	silhouette	0.3754	0.3103	0.3057	0.2582	0.2223
Ward	connectivity	120.2373	213.1131	286.9024	325.3052	362.3976
	Dunn	0.0301	0.0267	0.0278	0.0223	0.0248
	silhouette	0.4039	0.2546	0.2358	0.1760	0.1720
K-means	connectivity	138.0159	232.1786	348.9095	435.6401	484.9810
	Dunn	0.0167	0.0086	0.0101	0.0067	0.0057
	silhouette	**0.4219**	0.3207	0.2707	0.2546	0.2232
model	connectivity	386.6532	1339.7901	1512.6413	939.4246	1060.7524
	Dunn	0.0042	0.0038	0.0036	0.0038	0.0035
	silhouette	0.1283	−0.0302	0.0007	0.0918	0.0888

성측도 값에 따라 추천하는 군집방법이 다르므로 다각적인 정보를 활용해 군집분석 방법을 선택해야하며 어떤 한 방법이 우월하다고 할 수 없다.

5. 한우 관능평가 데이터에 대한 군집분석

5.1. 한우 데이터

축산과학원에서는 2006년 전 국민을 대상으로 추출한 표본 소비자에게 한우관능평가를 실시하였다. 수소(bull)의 8개 부위를 탕, 구이, 스테이크의 세 가지 요리 방법으로 조리한 후 소비자가 시식하여 연도, 다즙성, 향미, 전반적인 기호도를 0~100점 사이의 점수로 평가하도록 하였다. 여기서 사용되는 데이터는 구이를 시식한 총 1,701명의 소비자가 평가한 위 네 가지 관능평가 점수 데이터이다. 한우 수소 관능평가 데이터에 기반한 적합한 군집화를 시도하고 각 군집의 특성을 파악하고자 한다.

5.2. 군집 유효성

군집의 개수에 따른 군집방법별 내부 유효성 측정값을 나타낸 표 5.1을 보면 연결성 측도는 최장연결법을 이용하여 2개로 군집화한 경우 최적값을 나타내고, Dunn 지수는 최장연결법을 이용하여 5개로 군집화한 경우 최적값을 나타내며, 실루엣은 K-평균법을 이용하여 2개로 군집화한 경우에 최적값을 나타냄을 알 수 있다.

활용할 확률 정보가 없고 군집 배치에 대한 신뢰도 개념으로 실루엣 측도를 이용하여 군집방법을 선택하기로 하면 K-means 방법을 선택하게된다. 또한 군집 개수에 따라 군집방법별 안정성 측정값을 나타낸 표 5.2를 보면 APN 측도는 K-평균법을 이용하여 2개로 군집화한 경우 최적값을 나타내고 AD 측도의 경우 K-평균법을 이용하지만 6개로 군집화한 경우가 최적값을 나타낸다. 그림 5.1은 APN 측도를 고려하여 K-평균법에 의하여 2개로 군집화한 경우의 연도와 다즙성의 산점도를 나타낸다.

5.3. 군집방법과 군집개수의 선택

내부 유효성 측도인 연결성 측도는 변수의 수에 영향을 받기 때문에 관능평가 데이터와 같이 개체수(1,701개)에 비하여 측정변수(4개)가 매우 적은 데이터의 군집 유효성을 평가하는 측도로서 적합하지

표 **5.2.** 한우관능평가데이터의 안정성 측정값

군집방법	측정방법	군집의 개수				
		2	3	4	5	6
complete	APN	0.2957	0.4010	0.4907	0.4984	0.5020
	AD	50.9095	43.5403	42.0270	41.6407	39.7791
Ward	APN	0.2708	0.3438	0.4117	0.4472	0.4653
	AD	46.4462	40.7202	38.6395	37.1155	35.6040
K-means	APN	**0.1035**	0.1728	0.2994	0.3612	0.3974
	AD	41.5592	36.7300	35.7337	34.4690	**33.3699**
model	APN	0.2613	0.4973	0.3885	0.4392	0.5149
	AD	52.6385	54.0067	49.0585	47.0262	45.2524

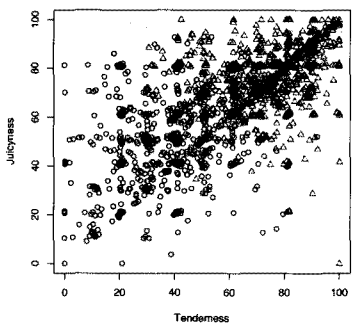

그림 **5.1.** K-평균법에 의한 군집화

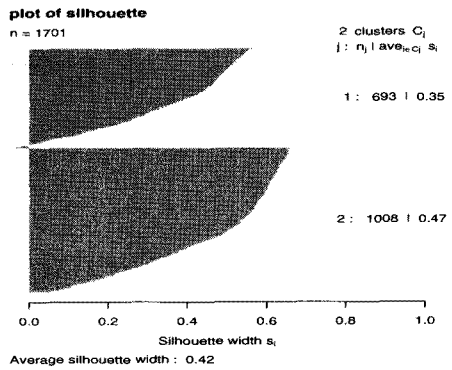

그림 **5.2.** 실루엣 측정값

않다. 또한 Dunn 지수는 같은 군집에 속해 있는 두 개체간의 가장 큰 거리에 대한 서로 다른 군집에 속해 있는 두 개체간의 가장 작은 거리의 비인 대표값의 비율로 나타나기 때문에 모든 개체간의 실제 거리를 반영하는 실루엣 측도가 Dunn 지수보다는 신뢰성이 높다고 판단된다. 따라서 3가지의 내부 유효성 측도 중 관능평가 데이터의 군집화에 적합한 측도로 실루엣이 추천하는 K-평균법에 의한 2개 군집화를 결정하였으며 또한 안정성 측도인 APN 측도도 K-평균법에 의한 2개 군집화 결정을 뒷받침해 주고 있다. 그림 5.2는 K-평균법에 의하여 2개로 군집화 하였을 때 군집1의 실루엣 값은 0.35, 군집2는 0.47임을 보여준다.

5.4. 군집의 비교

그림 5.3은 관능평가 변수인 연도, 다즙성, 향미, 전반적인 기호도의 스무딩 기법을 적용한 분포함수를 군집별로 보여주는데, 군집1은 대체로 낮은 점수이고 군집2는 높은 점수 분포임을 알 수 있다. 표 5.3에서는 군집1보다 군집2가 맛변수들의 평균이 전반적으로 높고 표준편차는 군집2보다 군집1에서 전반적으로 더 큰 것을 알 수 있다. 또한 두 군집간의 각 관능평가 변수에 대한 F-통계량과 p-값을 보여주며 각 변수의 평균은 유의한 차이가 있다. 두 군집에 대한 다변량 분산분석으로 Wilks Lambda 통계량은 0.32이고 p-값< 0.0001으로 두 군집간 관능평가변수들의 평균벡터는 통계적으로 유의한 차이가 있었다. 표 5.4와 같이 각 군집을 성별, 거주지, 연령, 수입, 시식한 한우의 부위, 고기의 만족도(여기서 만족도란 시식자가 고기에 대한 만족도를 "만족하지 못한다"(1), "만족한다"(2), "매우 만족한다"(3), "극도로

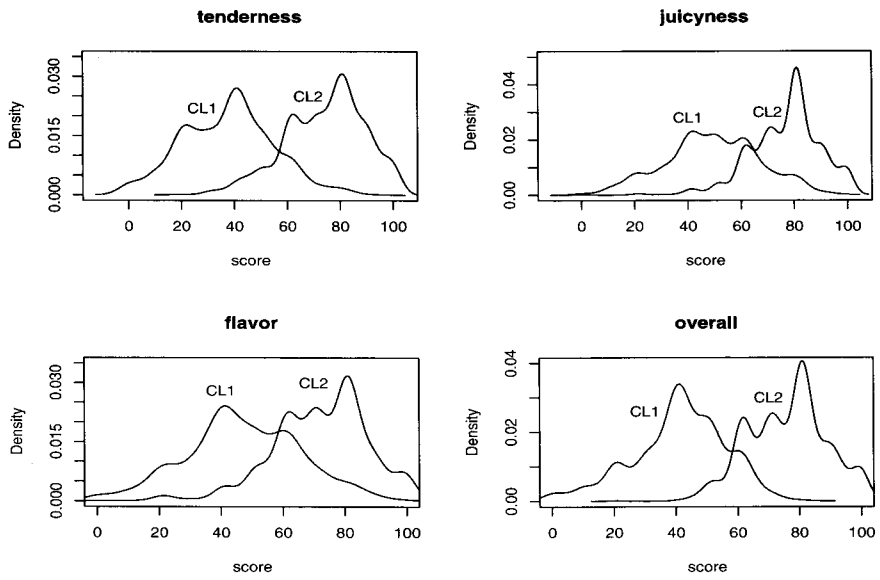

그림 **5.3.** 한우 관능평가 점수분포의 군집별 비교

표 **5.3.** 군집별 관능평가 변수들의 기초통계량(***: significant at 0.001 level)

	군집 1 (693개)		군집 2 (1,008개)		ANOVA	
	mean	sd	mean	sd	F-값	p-값
연도	37.6721	17.1613	74.4423	15.0014	2191.7	< 0.0001***
다즙성	49.1553	17.7070	76.8550	13.1837	1365.7	< 0.0001***
향미	47.2689	18.4850	72.1772	15.3166	915.8	< 0.0001***
전반적인 기호도	41.0795	15.0122	76.0682	12.6067	2703.1	< 0.0001***

표 **5.4.** 두 군집의 성별 등 카이제곱 동일성 검정 결과

	통계량	p-값
성별	0.0001	0.9927
거주지	17.5708	0.0005***
연령	10.2254	0.0368***
수입	22.7388	0.0004***
부위	111.9091	< 0.0001***
만족도	798.1111	< 0.0001***

만족한다"(4)의 네 등급으로서 평가한 결과이다)를 중심으로 카이제곱 독립성 검정을 한 결과, 성별을 제외한 나머지 특성에서 군집간 유의한 차이가 있었다.

그림 5.4부터 그림 5.9는 군집별로 성별, 거주지, 나이, 수입 등의 빈도를 막대그래프로 표현한 것이다. 그림 5.8은 군집별로 부위의 빈도를 나타내는데 맛있는 그룹인 군집2에는 등심, 보섭의 빈도가 높고, 군집1에는 우둔, 설도의 빈도가 높게 나타나 군집별로 부위의 특성을 나타낸다. 그림 5.9에서는 군집1에서는 1값의 빈도가 높고 군집2에서는 2와 3의 빈도가 높고 4의 빈도도 군집1에 비해 높게 나타나 군집1에 비해 군집2는 맛있는 그룹으로 설명될 수 있다.

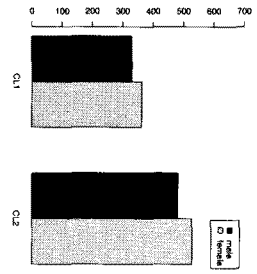

그림 5.4. 성별

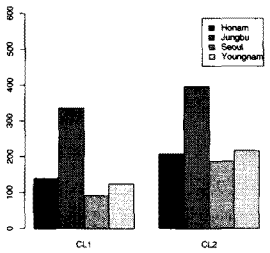

그림 5.5. 거주지

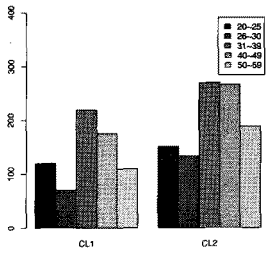

그림 5.6. 나이

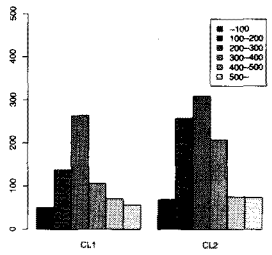

그림 5.7. 수입

그림 5.8. 부위

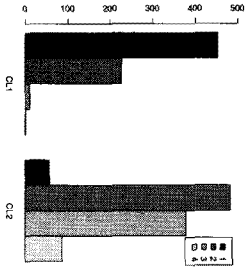

그림 5.9. 만족도

6. 결론

주어진 데이터에 대하여 최적의 군집방법과 군집의 개수를 찾는 것은 쉬운 문제가 아니다. 모의실험 결과 적절한 군집방법은 군집모양과 분포에 의존한다는 것을 알았다. 따라서 데이터의 특성을 파악하여 상황에 맞는 군집방법과 적절한 군집의 개수를 선택하는 것이 필요하며 또한 군집이 결정된 후에는 군집에 대한 분석이 뒤따라야 한다.

한우 관능평가 데이터에 대한 군집분석으로는 실루엣 측도, APN 측도를 고려하여 K-평균법에 의한 2개의 군집을 추정하였고 그 결과 관능평가점수가 전반적으로 낮은 군집과 높은 군집으로 나타났으며 두 군집은 시식자의 거주지, 연령, 수입의 분포에 있어 유의한 차이가 있었고 특히 쇠고기 부위의 특성을 반영하고 있었다. 데이터의 특성에 따른 군집방법과 군집의 유효성을 평가하는 도구로 앞으로도 다양한 측도가 개발될 것을 기대한다.

감사의 글

본 연구에 귀중한 한우 관련 데이터를 제공해 주신 국립 축산 과학원 안종남 과장님, 김동훈 과장님과 조수현 박사님께 감사드립니다.

참고문헌

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and Non-Gaussian clustering, *Biometrics*, **49**, 803–821.

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.

Dunn (1974). Well-separated clusters and optimal fuzzy partitions, *Journal of Cybernetics*, **4**, 95–104.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method-answers via model-based cluster analysis, *Computation Journal*, **41**, 578–588.

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

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). K-means clustering algorithm, *Applied Statistics*, **28**, 100–108.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.

Pollard, D. (1981). Strong consistency of K-means clustering, *Annals of Statistics*, **9**, 135–140.

Pollard, D. (1982). Central limit theorems for K-means clustering, *Annals of Statistics*, **10**, 919–926.

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

Scott, A. J. and Symons, M. (1971). Clustering methods based on likelihood ratio criteria, *Biometrics*, **27**, 387–397.

Ward, Jr., J. H. (1963). Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association*, **58**, 236–244.

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

A Comparison of Cluster Analyses and Clustering of Sensory Data on Hanwoo Bulls

Jaehee Kim[1] · Yoon Sil Ko[2]

[1]Department of Statistics, Duksung Women's University;
[2]Department of Statistics, Duksung Women's University

(Received April 2009; accepted July 2009)

Abstract

Cluster analysis is the automated search for groups of related observations in a data set. To group the observations into clusters many techniques has been proposed, and a variety measures aimed at validating the results of a cluster analysis have been suggested. In this paper, we compare complete linkage, Ward's method, K-means and model-based clustering and compute validity measures such as connectivity, Dunn Index and silhouette with simulated data from multivariate distributions. We also select a clustering algorithm and determine the number of clusters of Korean consumers based on Korean consumers' palatability scores for Hanwoo bull in BBQ cooking method.

Keywords: Average Distance(AD), Average Proportion(APN), complete linkage, connectivity, Dunn Index, K-means, model-based clustering, silhouette width, Ward's method.

This research is supported by 2009 Duksung Women's University Research Fund.
[1]Corresponding author: Professor, Department of Statistics, Duksung Women's University, 419 Ssangmun-Dong, Dobong-Gu, Seoul 132-714, Korea. E-mail: jaehee@duksung.ac.kr