연구보고

집락분석과 판별분석의 활용성연구+

채성산
대전대학교 이과대학 통계학과
황정연
한국전자통신연구소 소프트웨어 종합검증실

Applicability of Cluster Analysis and Discriminant Analysis

Seong-San Chae
Dept. of Statistics, Taejon University
Jung-Yeon Hwang
Electronics and Telecommunications Research Institute

Abstract

Cluster analysis is a primitive technique in which no assumptions are made concerning the data structure. And the number of groups is known a priori discriminant analysis provides an information how well N individuals are classified into their own groups. In this study, clustering, which is any partition of a collection of data points, generated by the application of eight hierarchical clustering methods was re-classified by discriminant analysis. Then correct classification ratios were obtained for the application of discriminant analysis through each clustering method and the direct application of discriminant analysis. By comparing the correct classification ratios, the applicability of cluster analysis and discriminant analysis considered.

+ 이 논문은 1993년도 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음.

1. 서론

p-차원 공간($p \geq 2$)에서 측정된 다변량자료에서, 연구의 대상이 되는 N개의 개체를 성질이 다른 것은 이질적인 집단으로, 성질이 같은 것은 유사한 집단으로 분류하는 구조적 단순화는 통계적 자료분석의 과정에 선행되어야 할 과제이다. 이때, 어떻게 유사한 자료들의 집합을 규정할 것인가, 혹은 몇개의 서로 다른 집단으로 구분할 것인가에 관한 연구[Milligan and Cooper, 1985: Krazanowski and Lai, 1988: 김응환, 1993]는 집락분석(군집분석, Cluster Analysis, Taxonomy, Pattern Recognition, Clumping, Typology)의 영역에 속하며, 집락분석의 연구결과로 형성된 가정된 집단, 즉 이미 분류된 집단에 새로운 개체가 어느 집단에 속할 것인가, 혹은 분류가 잘 되었는가하는 연구[안윤기와 이성석, 1992: 김성주와 정갑도, 1993]는 판별분석(Discriminant Analysis, Classification, Identification)의 영역이라 할 것이며, Shumway(1982), Grometstein and Schoendorf (1982)등은 판별분석의 현실적인 응용에 대한 예를 제시하였다.

집락분석과 판별분석에 관한 연구와 더불어, Gnanadesikan and Kettenring등(1989)은 통계적 자료분석의 선행단계에서 이용되는 집락분석과 다변량 자료분석시 널리 이용되고 있는 판별분석방법에 관련된 중요한 연구과제 및 아직도 해결하여야 할 여러가지 문제점을 취급하고 있다.

본 연구에서는 채 와 Warde(1991)의 "같은 자료에 여러가지 다른 집락방법을 적용하였을 때, 그 결과들을 비교 검토하여 집락의 수를 결정할 수 있다"는 주장을 근거로, 일반적으로 이용되고 있는 판별분석의 선행단계에서 집락분석을 적용하였고, 집락분석의 결과로서 형성된 집락에 판별분석을 실시하여, 두 분석방법의 활용성을 살펴보았다. 먼저, 본 연구에서 고려한 집락분석방법과 판별분석법에 관하여 살펴보고, 모의 실험에 관한 계획을, 제 4장에서는 도출된 결과를 분석하였고, 제 5장에서는 결과를 요약하고 있다.

2. 집락분석방법과 판별분석법

본 연구의 목적을 위하여 g개의 부모집단에 대하여 추출한 $N = n_1 + n_2 + \cdots + n_g$개의 p 차원 연습표본을 X라 하면,

$$X = \begin{Bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_g \end{Bmatrix} = \begin{Bmatrix} x_{111} & x_{112} & x_{113} & x_{114} & \cdots & x_{11p} \\ x_{121} & x_{122} & x_{123} & x_{124} & \cdots & x_{12p} \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ x_{1n1} & x_{1n2} & x_{1n3} & x_{1n4} & \cdots & x_{1np} \\ x_{211} & x_{212} & x_{213} & x_{214} & \cdots & x_{21p} \\ x_{221} & x_{222} & x_{223} & x_{224} & \cdots & x_{22p} \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ x_{2n1} & x_{2n2} & x_{2n3} & x_{2n4} & \cdots & x_{2np} \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ x_{g11} & x_{g12} & x_{g13} & x_{g14} & \cdots & x_{g1p} \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ x_{gn1} & x_{gn2} & x_{gn3} & x_{gn4} & \cdots & x_{gnp} \end{Bmatrix}$$

를 나타내며, $\overline{X}=(\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \cdots, \bar{x}_p)$는 p개 변수의 전체표본평균벡터, $S_g = \frac{1}{n_g - 1} \sum_{j}^{n_g} (x_{gj} - \bar{x}_g)(x_{gj} - \bar{x}_g)'$는 g 집단에 대한 표본 분산-공분산 행렬을 나타내며, $\bar{x}_g = \frac{1}{n_g} \sum_{j}^{n_g} x_{gj}$, 는 g 집단의 $p * 1$ 표본평균벡터이다. 이때, 합동 표본 분산-공분산 행렬은, $S_p = \{\sum_{h}^{g}(n_h - 1) S_h\} / \sum_{h}^{g}(n_h - 1)$, 이다.

이러한 자료에 대하여 Lance and Williams(1966, 1967)의 공식,

$$d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \pi | d_{ik} - d_{jk} |, \qquad (1)$$

에 의해서 4개의 모수, $(\alpha_i, \alpha_j, \beta, \pi)$, 로 정의되는 집락분석방법만을 고려하였다. 위에서 d_{ij}, d_{ik}, d_{jk}는 집락 i, j, k들 중 임의의 두 집락간의 거리를, $d_{(ij)k}$는 집락 i와 j를 결합하여 형성되는 새로운 집락과 집락 k간의 거리를 나타내며, i, j, k는 개체일 수도 있다. 여기서 d_{ij}, d_{ik}, d_{jk}는 준거리(semi-metric) 개념을 만족하는 비유사측도(dissimilarity measure)로서 Squared Euclidean distance를 사용하였는데, 이는 다음을 만족한다.

1) $d_{ij} \geq 0$, and $d_{ij} = 0$ iif $x_i = x_j$,

2) $d_{ij} = d_{ji}$.

이때, 3) $d_{ij} + d_{jk} \geq d_{ik}$을 만족하면, d_{ij}, d_{ik}, d_{jk}는 거리(metric) 개념을 만족한다고 한다. 실질적으로 거리(metric)를 만족하는 다른 비유사측도를 사용하였을 경우에도 계층적 집락분석의 결과는 거의 동일하다고 알려져 있다. 한편,

$$\alpha_i = \alpha_j = \alpha,$$
$$\alpha_i + \alpha_j + \beta = 1$$

의 두가지 제한 조건하에서, $d_{ij} < d_{jk} < d_{ik}$의 일반적 가정을 설정하면, DuBien과 Warde(1987)에 의한 다음의 공식을 얻게 된다.

$$d_{(ij)k} = \frac{1-\beta-2\pi}{2} d_{ik} + \frac{1-\beta+2\pi}{2} d_{jk} + \beta d_{ij}. \qquad (2)$$

이때, 두개의 모수, (β, π),로 정의되는 평면상에는 무수한 집락분석방법이 존재하게 되는데, 본 연구를 위하여 선택된 집락분석방법은 다음과 같다.

(1) $\beta = 0.0$, $\pi = -0.5, 0.0, 0.5$;

(2) $\beta = -0.25$, $\pi = -0.25, 0.0, 0.5$;

(3) $\beta = -0.5$, $\pi = 0.0, 0.25, 0.75$.

위의 방법중에서, (.0, −.5)는 최단연결법(Single Linkage), (.0, .0)은 평균연결법(Average Linkage), (.0, .5)는 최장연결법(Complete Linkage), 그리고 (−.25, 0)과 (−.5, .0)은 유동연결법(Flexible Strategies)으로 알려져 있다. 또한, (−0.5, .75)는 위

에서 언급된 다른 방법과 동시적용시 집락수의 예측에 상당히 좋은 결과를 주고 있다고 채와 Warde(1991)에 의하여 주장된 방법이다.

한편으로, 각 부모집단에 대한 사전확률과, 부모집단 i와 j에 속하는 개체를 오분류하였을 때 소요되는 비용이 같다는 일반적 가정하에서 판별분석방법의 적용을 요약하면 다음과 같다.

분류하여야 할 집단이 2개 ($g=1$)인 경우에, S_p를 2 집단표본의 합동분산-공분산 행렬이라고 하자. 선형판별함수(Linear Discriminant Function=LDF)를 이용한 판별분석에서는,

$$(\bar{x}_1-\bar{x}_2)' S_p^{-1} x_0 > \frac{1}{2}(\bar{x}_1-\bar{x}_2)' S_p^{-1}(\bar{x}_1+\bar{x}_2),$$

이면 개체 x_0를 집단 1로 분류하고, 그렇지 않은 경우에는 개체 x_0를 집단 2로 분류한다. 한편, 집단간의 분산-공분산 행렬이 다른 경우에는 2차판별함수(Quadratic Discriminant Function=QDF)를 이용하는 것이 일반적으로 알려져 있는데,

$$M = Ln(|S_1|/|S_2|)/2 + (\bar{x}_1 S_1^{-1}\bar{x}_1' - \bar{x}_2 S_2^{-1}\bar{x}_2')/2,$$

일때,

$$M > \frac{1}{2} x_0 (S_1^{-1}-S_2^{-1}) x_0' - (\bar{x}_1 S_1^{-1} - \bar{x}_2 S_2^{-1}) x_0',$$

이면 개체 x_0를 집단 1로 분류하고, 그렇지 않은 경우에는 개체 x_0를 집단 2에 분류하게 된다. 이러한 판별분석방법은 집단의 수가 3 이상인 경우에도 확대 응용할 수 있으며, 판별이 잘되었는지에 대한 판단은 개체들이 그 각각이 속한 집단에 분류된 비율, 적정분류율(=1-오분류율(Misclassification rate))을 살펴보아야 한다.

본 연구에서는 g개의 부모집단을 갖는 연습표본 X에 대하여, 혹은 연습표본 X에 집락분석을 적용시켜 생성된 g개의 집락을 갖는 X'에 대하여, IMSL(International Mathmatical and Statistical Library)의 부프로그램, DSCRM을 이용하여 선형판별분석과 2차판별분석을 실시하였다.

3. 모의실험의 계획

생성된 자료 혹은, 연구를 위하여 취득된 자료를 연습표본 X라 하자. 이러한 X에 집락분석방법을 적용한 결과로서 새로운(혹은, 동일한) 구조를 갖는 선행자료를 X'이라 하고, 판별분석을 적용한 결과로서 새로운(혹은, 동일한) 구조를 갖는 자료를 X''이라 하자. X''은 자료의 구조가 알려진 상황에서는 X 혹은 X'에 적용한 결과라고 생각할 수 있다.

연습표본 X의 생성시, 본 연구에서 고려한 설정된 모수를 살펴보면 다음과 같다.

1. N - X에 있는 자료의 수;
2. p - 변수의 수;
3. ρ_g - 변수들간의 관계정도(정규분포),
   혹은 (λ_g, t) - 분포의 모양(비정규분포);

4. G - 부모집단(sub-population)의 수;
5. G개의 부모집단 각각에 대한 확률분포 혹은 모집단의 형태;
6. n_g - G개의 부모집단 각각에 있는 자료의 수, $g = 1, 2, \cdots, G$;
7. δ_g - G개 부모집단 평균간의 Squared Euclidean 거리.

생성된 G개 부모집단의 분포는 다변량 분포를 따른다고 가정하였고, 이러한 분포를 따르는 자료를 생성시키기 위하여 IMSL에 있는 부프로그램인 RNMVN, RNUN, RNGAM을 이용하였다.

연구의 편의상, $N=60$, $p=2$, $G=3$인 경우만을 살펴보았으며, 이변량 정규분포를 따르는 자료구조는,

$\underline{X}_i \sim \text{BVN}(\underline{\mu}_g, \Sigma_g)$, $i=1, 2, \cdots, 60$, $g=1, 2, 3$ 이고,

3개의 부모집단에 있는 자료의 수($n_1-n_2-n_3$)는 20-20-20 혹은 25-20-15 이며, 각 모집단의 평균간 거리(δ_g)는 서로 동등함을 유지하도록 하였고,

즉, $\delta_g=\delta=2, 4$, 이며,

$$\Sigma_g=\Sigma=\begin{bmatrix} 1.0 & \rho_g \\ \rho_g & 1.0 \end{bmatrix}, \rho_g = 0.0, 0.3, 0.6, \ g=1, 2, 3.$$

한편, Johnson, Tietjen, and Beckman(1980)과 Johnson(1987)에 의한 비정규분포의 자료구조를 요약하면,

U ~ Uniform(-1, 1)과

W ~ Gamma(λ_g, 1)에서 $Y=W^t$라 하면,

$X = \sigma[\,3\Gamma(\lambda_g)/\Gamma(\lambda_g+2t)\,]^{1/2}\,YU + \mu$ 인데,

$\underline{X}_i$ ~ Non-Normal ($\underline{\mu}_g$, Ψ_g), $i=1, 2, \cdots, 60$, $g=1, 2, 3$ 이고,

3개의 부모집단에 있는 자료의 수($n_1-n_2-n_3$)는 20-20-20 혹은 15-20-25 이며, 각 모집단의 평균간 거리(δ_g)는 서로 동등함을 유지하도록 하였고,

즉, $\delta_g=\delta=2, 4$, 이며,

$t=0.2, 0.8$, $\lambda_g=1.0, 1.5, 2.0$, $g=1, 2, 3$, 로 분포의 모양을 변화하였으며, 범위모수($\sigma=1.0$)를 고정하면,

$$\Psi_g=\Psi=\begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}.$$

이때, μ는 위치모수, λ, t는 형태모수이며, λ와 t의 변화에 따라서, $\lambda-t=1$은 Exponential Power 분포, $\lambda=1$, $t=0$은 Uniform, $\lambda=1.5$, $t=0.5$는 Normal, $\lambda=2.0$, $t=1.0$은 Laplace 분포를 따르게 된다. 여기서 각 부모집단의 첨도(Kurtosis)는 $9\Gamma(\lambda_g+4t)\Gamma(\lambda_g)/5\Gamma^2(\lambda_g+2t)$, $g=1, 2, 3$, 이다.

판별분석의 적용시, 3개($G=3$)의 부모집단을 갖도록 생성된 자료(X)에 대하여 부모집단 i와 j에 속하는 개체를 오분류하였을 때 소요되는 비용은 같다고 가정하고, 사전확률은 각 부모집단에 속하는 개체들의 수에 따라 결정되도록 하였다. 먼저, 설정된 집단이 유사한 자료들의 집합인가에 대하여 LDF와 QDF에 의한 판별분석을 실시하여 X''을 형성하

고, X와 X″을 비교하여 적정분류율(Correct classification ratio)을 계산하였다. 다음으로, 다양한 집락분석방법(예: Single Linkage = Nearest Neighbor, Average Linkage, Complete Linkage = Furtherest Neighbor, Flexible Strategies 등)을 생성된 자료 X에 적용하여 $G=3$을 갖는 새로운 선행집단(X')을 형성시킨 후, 이들을 대상으로 판별분석을 실시하여 X''을 형성하고, X'과 X''을 비교하여 적정분류율을 계산하였으며, 다음과 같이 그 결과들을 비교 검토하였다.

4. 분석 및 결과

본 연구는 Monte Carlo Simulation을 실행하여 집락방법들에 의하여 복원된 각 선행집락들을 배경으로 판별분석을 적용, 주어진 자료에 대하여 더욱 명확한 집단을 형성하는가를 살펴보고자 하였다. 이때, 각 집락방법들의 적용과 판별분석에 의한 분류의 적합분류비율에 관한 검토는 집락방법들의 비교연구 및 그들의 특성을 밝히는 것과 상통한다.

모수설정에 따른 자료구조에 대하여, 9개의 집락분석을 통하여 형성된 선행집락(X')과 또한 자료생성시 형성된 원래의 자료(X)에 선형판별분석(LDF)과 2차판별분석(QDF)을 실시하였다. 이때, 집락분석의 적용에 의하여 형성되는 선행집락의 수는 자료생성시 설정하였던 3개의 집락을 형성하도록 하였다. 연구의 목적상 집락분석방법의 적용에 의하여 형성되는 각 선행집락의 크기(집락내 개체의 수, n_g)는 3 이상인 경우와, 각 집락의 표본 분산-공분산행렬이 정칙행렬(Non-Singular matrix)인 경우로 제한하였다. 이러한 제한조건하에서 최단거리법($(0.0, -0.5)$ = Single Linkage)의 특성을 고려하면, 다른 계통적집락방법과의 동시적용은 문제점을 지닌다. 실질적으로 생성된 집락들의 평균간 거리가 근접 할때, 최단거리법의 적용은 1개 혹은 2개의 집락만을 형성하는 경우가 대부분이었으며, 변수들간의 관계정도가 클수록 이러한 경향은 크게 나타난다. 따라서, 본 연구에서는 최단거리법을 제외한 8개의 집락방법만을 동시 적용하였다.

판별분석의 적용시 전체 개체 중에서 원래의 집단(X)으로 적합하게 분류된 개체의 비율 및 8개의 집락분석방법의 적용에 의하여 생성된 선행집단(X')으로 적합하게 분류된 개체의 비율을 적정분류율이라고 할때, 설정된 모수, (ρ_g, δ, n_g), 혹은 $(\lambda_g, t, \delta, n_g)$에 대하여 각각 100번의 반복이 이루어졌고, 적정분류율의 표본평균(AVG)과 AVG의 표본표준편차(STD)가 얻어졌다.

여기서 AVG와 STD는 이미 집단이 분류되어 있다는 가정하에 적용하는 판별분석의 적정분류성과, 집단이 알려져 있지 않다고 생각하고 8개의 집락분석방법을 통하여 분류한 선행집단에 판별분석을 적용한 결과의 적정분류성에 대한 정보를 제공한다. 〈표 1〉과 〈표 2〉는 LDF 및 QDF를 적용한 결과로서 AVG만을 요약 정리하였으며, 이러한 정보에 근거하여 (β, π) 평면상에 정의된 집락분석방법과 판별분석의 활용성을 제고해 보고자 하였다.

먼저, 〈표 1〉을 이용하여 정규분포하에서 LDF 및 QDF의 결과를 설정된 모수, (ρ, δ, n_g)의 변화에 대하여 살펴보았다. 표에서 DISCRIM은 직접적인 판별분석의 적용결과

를 나타내며, 나머지는 8개의 집락분석을 선행적으로 실시한 후 판별분석을 적용한 결과이다.

모집단의 공분산이 같은 경우, LDF는 ρ와 각 집락에 속한 개체의 수(n_g)의 변화에 적정분류율의 변화는 거의 없으며, QDF는 ρ의 변화에 적정분류율의 변화는 거의 없으나, 각 집락에 속한 개체의 수(n_g)에 영향을 받고 있음을 알 수 있었다. 즉, 일반적으로 알려진 것과 같이 집단의 공분산이 같은 경우, (ρ, δ, n_g)의 설정에 관계없이 LDF의 적용이 QDF의 적용보다 선호됨을 확인 할 수 있었다. 여기서 LDF의 적용이 선호된다고 하자. 이 결과를 선행적으로 집락분석을 적용한 후 판별분석을 적용한 결과와 비교하면, 판별분석을 직접적으로 이용하는것 보다 선행적으로 집락분석을 적용한 후 생성된 선행집락에 판별분석을 이용하는 것이 더욱 타당하다고 생각되었다. 이러한 결과는, δ가 가까운 경우에 더욱 명백하게 나타났다.

각 부모집단의 공분산($\rho_1=0.0$, $\rho_2=0.3$, $\rho_3=0.6$)이 상이한 경우에도 공분산이 같은 경우와 거의 유사한 결과를 보여주고 있다. 즉, 정규분포인 경우에는 공분산에 차이가 있더라도 QDF 보다는 LDF의 이용이 선호되며, δ가 가까울수록 선행적으로 집락분석을 적용한 후 생성된 선행집락에 판별분석을 이용하는 것이 타당하다고 생각되었다. 더불어, 비정규분포를 따르는 자료에 대하여 그 결과를 살펴보면 다음과 같다.

〈표 2〉는 $\lambda=\lambda_1=\lambda_2=\lambda_3=1.5$ 이고 t(kurtosis)=.2(2.01), .8(5.10)인 경우와, $\lambda_1=1.5$, $\lambda_2=1.0$, $\lambda_3=2.0$ 이고 t(kurtosis)=.2(2.01, 2.13, 1.96), .8(5.10, 6.83, 4.24)인 경우의 모의실험결과이다.

$\lambda=\lambda_1=\lambda_2=\lambda_3=1.5$ 이고 t(kurtosis)=.2(2.01), .8(5.10)인 경우, 즉, 부모집단의 모양이 같으면 QDF 보다 LDF의 적용이 타당하다고 생각되며, 모집단 분포의 뾰쪽한 정도에 따라 LDF와 QDF의 적정분류율이 증가하는 결과가 관찰된다. 그러나, LDF 혹은 QDF의 적용문제에 있어서 정규분포인 경우와 같이 LDF의 이용이 QDF의 이용보다 선호됨을 알 수 있었다. 마찬가지로, LDF가 선호된다는 가정하에 집락분석 후 판별분석을 실시한 결과와 비교하면, ($t=0.2$, $\delta=4.0$)인 경우를 제외하고는 판별분석을 직접적으로 적용하는 것보다는 선행적으로 집락분석을 적용한 후 판별분석을 이용하는 것이 타당함을 알 수 있었다.

$\lambda_1=1.5$, $\lambda_2=1.0$, $\lambda_3=2.0$ 이고 t(kurtosis)=.2(2.01, 2.13, 1.96), .8(5.10, 6.83, 4.24)인 경우, 즉, 각 부모집단의 모양이 모두 다를 때에도 LDF가 QDF에 선호됨을 알 수 있다. 그러나, QDF의 적정분류율이 상당한 수준으로 증가되었음을 살펴볼 수도 있다. 이러한 측면에서, 본 연구의 결과 만을 가지고 LDF가 QDF에 항상 선호된다고 주장할 수는 없으며, 어떤 형태의 자료에 QDF의 선호가 나타날 수도 있을 것이다. 또한, 집락분석후 판별분석의 적용에 관한 비교에 있어서 ($t=0.2$, $\delta=4.0$)인 경우를 제외하고는 δ가 가까울수록 선행적으로 집락분석을 적용한 후 생성된 선행집락에 판별분석을 이용하는 것이 바람직함을 확인 할 수 있었다.

이러한 위의 결과에 대한 분석은 8개의 계통적 집락분석의 적용에 의하여 생성된 선행집락이 $g=3$으로 타당하게 분류되었음을 전제로 한 것이다.

〈 표 1 〉 정규분포하에서 X와 X'에 대한 1차 및 2차판별분석의 적정분류율(AVG)

공 분 산			같 음						다 름	
			1 차			2 차			1 차	2 차
n_g	δ	집락방법\ ρ	.0	.3	.6	.0	.3	.6	(.0, .3, .6)	
		DISCRIM	.7455	.7627	.8017	.3675	.3612	.3687	.7550	.3692
	2	(.0 , .0)	.9605	.9662	.9613	.5875	.5755	.5887	.9605	.5803
		(.0 , .5)	.9550	.9547	.9575	.5480	.5538	.5753	.9562	.5797
20		(−.25, −.25)	.9625	.9570	.9677	.5645	.5492	.5813	.9572	.5548
:		(−.25, .0)	.9655	.9642	.9657	.5488	.5390	.5452	.9653	.5350
20		(−.25, .5)	.9560	.9575	.9593	.5398	.5360	.5642	.9602	.5188
:		(−.5 , .0)	.9612	.9640	.9613	.5380	.5365	.5262	.9647	.5362
20		(−.5 , .25)	.9642	.9615	.9618	.5322	.5163	.5198	.9597	.5205
		(−.5 , .75)	.9450	.9485	.9450	.5262	.5095	.5218	.9390	.4902
		DISCRIM	.9642	.9618	.9677	.4818	.4702	.4880	.9565	.4797
	4	(.0 , .0)	.9725	.9743	.9743	.4865	.5138	.5347	.9755	.5290
		(.0 , .5)	.9668	.9725	.9740	.4918	.5080	.5188	.9735	.5152
		(−.25, −.25)	.9725	.9733	.9785	.4900	.4955	.5388	.9722	.5113
		(−.25, .0)	.9733	.9758	.9765	.4842	.4857	.5255	.9780	.5105
		(−.25, .5)	.9730	.9728	.9788	.4802	.4927	.5093	.9747	.4972
		(−.5 , .0)	.9745	.9760	.9798	.4862	.4983	.5170	.9740	.5070
		(−.5 , .25)	.9758	.9763	.9812	.4768	.5040	.5125	.9755	.4898
		(−.5 , .75)	.9727	.9748	.9768	.4873	.4928	.5208	.9712	.4963
		DISCRIM	.7545	.7715	.8095	.2908	.2963	.2937	.7692	.2923
	2	(.0 , .0)	.9580	.9592	.9622	.5615	.5918	.5810	.9632	.5858
		(.0 , .5)	.9552	.9593	.9543	.5543	.5657	.5277	.9553	.5657
25		(−.25, −.25)	.9585	.9695	.9620	.5443	.5745	.5603	.9667	.5693
:		(−.25, .0)	.9642	.9593	.9638	.5388	.5553	.5577	.9673	.5292
20		(−.25, .5)	.9585	.9627	.9533	.5202	.5340	.5513	.9538	.5225
:		(−.5 , .0)	.9608	.9638	.9612	.5230	.5238	.5485	.9640	.5193
15		(−.5 , .25)	.9610	.9607	.9570	.5165	.5195	.5433	.9582	.5120
		(−.5 , .75)	.9450	.9508	.9537	.5190	.5228	.5212	.9390	.5270
		DISCRIM	.9645	.9650	.9718	.3903	.3820	.4025	.9613	.3952
	4	(.0 , .0)	.9750	.9728	.9710	.5342	.5333	.5600	.9713	.5260
		(.0 , .5)	.9675	.9723	.9695	.5227	.5187	.5380	.9692	.5242
		(−.25, −.25)	.9733	.9753	.9768	.5228	.5245	.5662	.9723	.5255
		(−.25, .0)	.9768	.9772	.9778	.5138	.5130	.5348	.9783	.5283
		(−.25, .5)	.9743	.9767	.9770	.5085	.5065	.5378	.9770	.5218
		(−.5 , .0)	.9752	.9748	.9807	.5157	.5082	.5342	.9767	.5293
		(−.5 , .25)	.9750	.9765	.9753	.5102	.5158	.5297	.9770	.5210
		(−.5 , .75)	.9723	.9718	.9775	.5113	.5135	.5390	.9740	.5237

〈 표 2 〉 비정규분포하에서 X와 X′에 대한 1차 및 2차판별분석의 적정분류율(AVG)

λ			같음($\lambda=\lambda_1=\lambda_2=\lambda_3=1.5$)				다름($\lambda_1=1.5$, $\lambda_2=1.0$, $\lambda_3=2.0$)			
			1 차		2 차		1 차		2 차	
n_g	δ	집락방법\t	.2	.8	.2	.8	.2	.8	.2	.8
		DISCRIM	.7215	.7937	.3593	.4053	.7115	.7317	.5508	.5878
	2	(.0 , .0)	.9597	.9658	.5912	.6355	.9633	.9612	.8010	.8833
		(.0 , .5)	.9577	.9583	.5643	.5957	.9553	.9548	.7955	.8783
20		(−.25,−.25)	.9668	.9672	.5673	.5675	.9672	.9568	.7870	.8420
:		(−.25, .0)	.9675	.9677	.5447	.5558	.9602	.9578	.7607	.8372
20		(−.25, .5)	.9570	.9633	.5463	.5225	.9583	.9372	.7473	.7962
:		(−.5 , .0)	.9582	.9667	.5302	.5243	.9590	.9447	.7398	.7798
20		(−.5 , .25)	.9535	.9648	.5168	.5205	.9620	.9350	.7362	.7660
		(−.5 , .75)	.9400	.9495	.5167	.5120	.9450	.9138	.7398	.7332
		DISCRIM	.9770	.9552	.4483	.5547	.9690	.9003	.6710	.6620
	4	(.0 , .0)	.9667	.9718	.4632	.5875	.9697	.9625	.7085	.7973
		(.0 , .5)	.9742	.9753	.4575	.5723	.9678	.9698	.7018	.7967
		(−.25,−.25)	.9678	.9757	.4640	.5772	.9707	.9682	.7100	.7685
		(−.25, .0)	.9722	.9772	.4542	.5683	.9722	.9675	.6958	.7353
		(−.25, .5)	.9760	.9775	.4647	.5672	.9693	.9660	.6988	.7210
		(−.5 , .0)	.9742	.9810	.4603	.5687	.9698	.9695	.6998	.7243
		(−.5 , .25)	.9730	.9750	.4695	.5588	.9700	.9675	.6993	.7265
		(−.5 , .75)	.9660	.9680	.4597	.5518	.9623	.9503	.6905	.7170
		DISCRIM	.7275	.7945	.2772	.3228	.7073	.7008	.4875	.5163
	2	(.0 , .0)	.9667	.9552	.5865	.6268	.9603	.9663	.8155	.8805
		(.0 , .5)	.9608	.9608	.5935	.5818	.9593	.9530	.7882	.8537
25		(−.25,−.25)	.9660	.9632	.5510	.5710	.9672	.9590	.7893	.8322
:		(−.25, .0)	.9713	.9662	.5493	.5430	.9610	.9600	.7445	.8182
20		(−.25, .5)	.9552	.9552	.5238	.5350	.9590	.9438	.7353	.7963
:		(−.5 , .0)	.9633	.9657	.5323	.5305	.9637	.9507	.7272	.7815
15		(−.5 , .25)	.9580	.9593	.5067	.5137	.9622	.9450	.7370	.7655
		(−.5 , .75)	.9340	.9427	.5053	.5190	.9362	.9303	.7300	.7398
		DISCRIM	.9807	.9562	.3650	.4702	.9655	.8937	.5897	.5787
	4	(.0 , .0)	.9720	.9772	.5020	.6200	.9712	.9622	.7143	.8190
		(.0 , .5)	.9727	.9747	.4942	.6010	.9677	.9603	.7178	.7942
		(−.25,−.25)	.9667	.9758	.4923	.5903	.9728	.9665	.7012	.7640
		(−.25, .0)	.9747	.9803	.4925	.5847	.9688	.9703	.7123	.7133
		(−.25, .5)	.9690	.9760	.4955	.5892	.9717	.9627	.7062	.6802
		(−.5 , .0)	.9740	.9777	.4963	.5912	.9728	.9663	.7185	.6890
		(−.5 , .25)	.9757	.9775	.4927	.5833	.9707	.9672	.7045	.6892
		(−.5 , .75)	.9617	.9712	.4857	.5848	.9633	.9562	.6983	.6858

5. 결론

본 연구에서는 주어진 자료에 대하여 어떤 가정이 필요치 않는 계통적집락분석과 일반적 가정하의 판별분석에 대한 두 분석방법간의 상호보완적인 활용성을 비교검토하고, 현실적인 응용에 대해 알아 보고자 하였다. 설정된 모수, (ρ, δ, n_g), 혹은 $(\lambda, t, \delta, n_g)$에 대하여 생성된 자료$(X)$ 및 8개의 계통적 집락분석의 적용에 의하여 생성된 선행집락(X')에 판별분석을 적용한 결과를 요약하면 다음과 같다.

첫째로, 정규분포하에서는 (ρ, n_g)의 설정에 관계없이 2차판별함수(QDF)에 대하여 선형판별함수(LDF)의 적용이 더욱 바람직함을 알 수 있으며,

둘째로, 비정규분포하에서는 (λ, t, n_g)의 설정에 따라 선형판별분석 혹은 2차판별분석의 적용을 선별적으로 고려할 필요성이 있으며,

셋째로, (ρ, n_g) 혹은 (λ, t, n_g)의 설정에 관계없이 집단간의 거리, δ가 가까운 경우에는, 생성된 자료(X)에 선형판별분석 혹은 2차판별분석을 직접적으로 이용하는것 보다 선행적으로 집락분석을 적용하여 생성된 선형집락에 판별분석을 적용하는 것이 판별분석에 의한 오분류율(Misclassification ratio)을 줄이는 방법임을 알 수 있다.

이러한 결과는 판별분석을 자료분석의 도구로 이용하는 사회과학 혹은 자연과학분야에서, 분석하여야 할 자료에 대하여 집단에 대한 명확성이 결여 되어 있거나, 혹은 집단에 대한 정보가 전혀 없는 경우, 선행적인 집락분석의 적용을 통하여 판별분석에 의한 오분류율을 감소시킬 수 있다.

참고문헌

[1] 김응환 (1993), 「쿨롱에너지 네트워크의 확장을 통한 집락분석」, 충남대 박사학위 논문.

[2] 김성주와 정갑도 (1993), "공분산행렬이 서로 다를 경우 그래프에 의한 판별분석," 「응용통계연구」, 제6권 2호, pp. 409-419.

[3] 안윤기와 이성석 (1992), "투사지향방법에 의한 판별분석의 모의실험분석," 「응용통계연구」, 제5권 1호, pp. 103-111.

[4] 채성산과 William D. Warde (1991), "A Method to Predict the Number of Clusters," 「통계학연구」, 제20권 2호, pp. 162-176.

[5] DuBien, J. L. (1976), *Comparative Techniques for the Evaluation of Clustering Methods*, Unpub. Ph. D. thesis, Oklahoma State University.

[6] DuBien J. L. and Warde, W. D. (1987), "A Comparison of Agglomerative Clustering Method with respect to Noise," *Communication of Statistics. Theory and Method*, Vol. 16, pp. 1433-1460.

[7] Gnanadesikan, R. and Kettenring, J. R. 외 다수 (1989), "Discriminant Analysis and Clustering," *Statistical Science*. Vol. 4, pp. 34-69.

[8] Grometstein, A. A. and Schoendorf, W. H. (1982), "Applications of Pattern Recognition in Radar," *Handbook of Statistics*, Vol. 2, pp. 575-593.

[9] Krzanowski, W. J. and Lai, Y. T. (1988), "A Criteria for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering," *Biometrics*, Vol. 44, pp. 23-34.

[10] Lance, G. N. and Williams, W. T. (1966), "A generalized Sorting Strategy for Computer Classification," *Nature*, Vol. 212, p. 218.

[11] Lance, G. N. and Williams, W. T. (1967), "A general Theory of Classificatory Sorting Strategies," *1. Hierarchical Systems*, *The Computer Journal*, Vol. 9, pp. 373-380.

[12] Mark E. Johnson, Gary L. Tietjen, and Richard J. Beckman (1980), "A New Family of Probability Distributions With Applications to Monte Carlo Studies," *Journal of the American Statistical Association*. Vol. 75, pp. 276-279.

[13] Mark E. Johnson (1987), *Multivariate Statistical Simulation*. New York: John Wiley & Sons

[14] Milligan, G. W. and Cooper, M. C. (1985), "An Examination of Procedures for Determining the Number of Clusters in a Data Set," *Psychometrika*, Vol. 50, pp. 159-179.

[15] Rand, W. M. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of American Statistical Association*, Vol. 66, pp. 846-850

[16] Shumway, R. H. (1982), Discriminant Analysis for Time Series. *Handbook of Statistics*, Vol. 2, pp. 1-46.