*Journal of Korean*
*Data & Information Science Society*
*2004, Vol. 15, No. 3, pp. 555~563*

# Pre-Adjustment of Incomplete Group Variable via K-Means Clustering[1)]

S.Y. Hwang[2)] · H.E. Hahn[3)]

## Abstract

In classification and discrimination, we often face with incomplete group variable arising typically from many missing values and/or incredible cases. This paper suggests the use of K-means clustering for pre-adjusting incompleteness and in turn classification based on generalized statistical distance is performed. For illustrating the proposed procedure, simulation study is conducted comparatively with CART in data mining and traditional techniques which are ignoring incompleteness of group variable. Simulation study manifests that our methodology out-performs.

***Keywords*** : CART, Classification, Incomplete variable, K-means clustering

## 1. 서론

판별분류분석은 주어진 다변량 관측치들을 이미 알려진 그룹으로 분류할 수 있는 함수를 추정한 후 새로운 관측치를 추정된 함수를 이용하여 어떤 그룹으로 분류할 것인가를 결정하는 분석 기법이다. 그러나 이러한 목적으로 자료를 수집할 때, 그룹에 대한 정보를 갖는 그룹변수의 값이 결측되거나, 결측값이 없다 할지라도 수집된 그룹변수의 신뢰도가 낮은 경우가 종종 발생한다. 이와 같이 불완전한 그룹변수를 이용하여 관측치들이 속한 그룹으로 판별분류를 하면 잘못된 판별함수를 추정하는 오류가 발생할 수 있으며, 잘못 추정된 판별함수를 통해 새로운 관측치를 분류할 때에도 실제 소속된 그룹과 다른 그룹으로 분류될 가능성이 높아진다. 이 논문에서는 불완전한 그룹변수의 판별분류를 위하여 판별분류분석의 선행분석으로 K-평균 군집분석사용을

---

1) This work was supported by 2004 Sookmyung Women's University research grant.

2) First Author : Professor, Dept. of Statistics, Sookmyung Women's University, Seoul, 140-742, Korea.
Email : shwang@sookmyung.ac.kr

3) Graduate student, Dept. of Statistics, Sookmyung Women's University

제안한다.

이 논문은 다음과 같이 구성되어 있다. 제 2장에서는 분류를 위한 기존의 잘 알려진 통계적 방법들로써 K-군집방법, 일반화된 거리에 의한 판별방법, 의사결정나무분석을 소개하고 제 3장에서는 불완전한 그룹변수로 인해 발생되는 문제를 개선할 수 있는 방법을 제안한다. 제 4장에서는 제안된 방법을 평가하기 위한 모의실험을 실시하고 결과를 해석하고 있다.

## 2. 분류를 위한 기존의 통계적 방법

### 2.1 K-군집방법

군집분석(cluster analysis)은 어떤 개체나 대상들 사이의 상사성(similarity) 혹은 비상사성(dissimilarity)을 근거로 하여 유사한 개체들로 이루어진 몇 개의 군집으로 집단화 하는 다변량 기법이다. 즉, 같은 군집에 속한 개체 사이에는 어떤 종류의 밀접한 상사성이, 그리고 다른 군집에 속한 개체 사이에는 상대적 비상사성이 존재한다는 것을 원칙으로 한다(cf. 성웅현(2000)). 군집분석의 목적은 군집의 개수, 내용, 구조 등이 사전에 정의 되지 않은 상황 하에서 '자연스러운' 군집을 찾고 이를 통해 전체 다변량 구조를 파악하고 군집의 형성과정과 그 특성, 그리고 식별된 군집간의 관계 등을 체계적으로 연구, 분석하는 것이다. 개체 사이의 상사성 혹은 비상사성은 거리(distance)를 통해 측정할 수 있으며, 다음과 같은 유클리드(Euclid) 거리를 많이 사용한다.

$$d_{ij} = d(X_i, X_j) = \{ (X_i - X_j)^T (X_i - X_j) \}^{1/2} \tag{2.1}$$

하지만 변수의 측정척도에 민감한 단점 때문에 이를 보완하기 위해 각 변수를 표준편차로 나눈 표준화 변수를 사용하는 경우도 많다. 또한 변수들 사이의 상관관계가 존재할 때 측정척도의 불변성과 상관관계를 고려한 (통계적) 거리를 마할라노비스(Mahalanobis) 거리라 부른다. 여기서 $S$ 는 표본 공분산 행렬을 나타낸다.

$$d_{ij} = d(X_i, X_j) = \{ (X_i - X_j)^T S^{-1} (X_i - X_j) \}^{1/2} \tag{2.2}$$

기본적인 K-평균(K-means) 군집방법은 개체간의 거리를 (2.1)의 유클리드(Euclid) 거리로 정의한 후 군집의 평균을 계산하여 군집을 형성해가는 방법이다(cf. Johnson and Wichern(2002)). K-평균 군집방법은 알고리즘이 간단하여 특히 큰 자료의 개체 군집화에 효율적인 것으로 알려져 있으나 군집들이 대체로 비슷한 크기로 형성되는 현상이 자주 발견되고 이러한 현상으로 인하여 작은 집단에 속한 개체들이 큰 집단의 일부로 간주되어 잘못 분류되는 경우가 종종 발생되는 단점이 있는 것으로 알려져 있다. 이를 보완하기 위한 변형으로서 허명회(2000)는 이중 K-평균(double K-means) 군집방법을 제안하였다. 또한 K-모드(K-modes) 알고리즘은 범주형 자료를 대상으로 하는 군집분석 방법으로 K-평균 알고리즘의 형식을 유지하면서 범주형 자료에 적합하도록 제안된 방법이다(김보화와 김규성(2002)).

### 2.2 일반화 거리에 의한 분류방법

여러 개의 판별변수들의 정보를 이용하여 관측치들을 분류하는 경우 어떤 기준에 의하여 임의의 관측치들의 소속집단을 추정 혹은 예측할 때 좋은 분류기준이 되기 위해서는 가능한 잘못 분류될 가능성이나 확률이 될 수 있는 대로 작아야 한다. 분류기준을 평가하는 기준으로 정분류율을 사용할 수 있는데, 정분류율이란 전제 관측치들 중 실제의 그룹으로 적합하게 분류된 개체의 비율을 의미한다. 즉, 정분류율이 높을수록 좋은 분류기준이라고 할 수 있다. 판별벡터변수에 근거를 둔 방법은 모부분집단들이 각각 다변량정규분포를 따른다는 가정 하에서 판별함수를 구하며 주어진 다변량관측치들을 두 개 이상의 집단으로 분류하는데 사용하는데 그 함수 자체의 해석이나 유의성은 고려되지 않으며, 모분포의 공분산 행렬은 같거나 다를 수도 있다.

특정한 개체 $X^T=(X_1, X_2, \cdots, X_p)$ 로부터 부분집단 $G_i$ 의 중심 $M_i$ 까지의 거리의 제곱과 부분집단 $G_i$ 의 공분산행렬 $S_i$ 를 사용한 일반화 거리(generalized distance)는 다음과 같이 정의된다.

$$D^2(i)=(X-M_i)^T S_i^{-1}(X-M_i)+\log|S_i| - \log(\pi_i) \tag{2.3}$$

여기서 $\pi_i$ 는 부분집단 $G_i$ 의 사전확률이다. 이 때 개체 $X^T$ 는 (2.3)의 거리를 최소화하는 부분집단 $G_i$ 로 분류한다. 이 방법은 부분집단들의 분산구조가 서로 다를 경우 그 사실을 거리 계산에 고려하고 있으며 부분집단의 모공분산 행렬이 같은 경우에는 (2.3)에서 $\log|S_i|$항을 제거하고 합동분산공분산 행렬로 대체하여 거리를 정의한다(김기영과 전명식(1996, 1997) 참조).

### 2.3 의사결정나무

의사결정나무(decision tree)는 의사결정규칙(decision rule)을 나무구조(dendrogram)로 도표화하여 분류와 예측을 수행하는 분석방법이다. 이 방법은 분류 또는 예측의 과정이 나무구조에 의한 추론규칙에 의해서 표현되기 때문에 다른 방법들(판별분석, 회귀분석 등)에 비해서 연구자가 그 과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다. 또한, 연구자는 나무구조로부터 어떤 변수가 그룹의 분류에 영향을 많이 주는지 쉽게 파악할 수 있다. 데이터마이닝(data mining)에서의 의사결정나무는 탐색(exploration)과 모형화(modeling)라는 두 가지 특성을 모두 가지고 있다고 할 수 있다. 즉, 의사결정나무는 판별분석 또는 회귀분석 등과 같은 모수적 모형을 분석하기 위해서 사전에 이상치를 탐색하거나 분석에 필요한 변수 또는 모형에 포함되어야 할 교호효과를 찾아내기 위해서 사용될 수 있고, 그 자체가 분류 또는 예측모형으로 사용될 수도 있다. 의사결정나무분석을 위해서 CHAID, CART, C4.5와 같은 다양한 알고리즘이 제안되어 있으며, 일반적으로 의사결정나무분석은 다음과 같은 단계를 거치게 된다.

(1) 의사결정나무의 형성 : 분석의 목적과 자료구조에 따라서 적절한 분리기준(split criterion)과 정지규칙(stopping rule)을 지정하여 의사결정나무를 얻는다.
(2) 가지치기 : 분류오류(classification error)를 크게 할 위험(risk)이 높거나 부적절

한 추론규칙(induction rule)을 가지고 있는 가지(branch)를 제거한다.

⑶ 타당성 평가 : 이익도표(gains chart)나 위험도표(risk chart) 또는 검증용 자료(test data)에 의한 교차타당성(cross validation) 등을 이용하여 의사결정나무를 평가 한다.

⑷ 해석 및 예측 : 의사결정나무를 해석하고 예측모형을 설정한다.

이와 관련된 더 자세한 이론은 강현철 외 4인(2000)을 참고하길 바란다.

## 3. 제안된 불완전 그룹변수의 판별 알고리즘

### 3.1 결측치가 있는 그룹변수

몇 개의 집단에 속하는 다변량 관측치로부터 각 집단의 차이를 분류하고자 할 때, 그룹변수에 결측치(missing value)가 있는 경우가 생기게 된다. 설문조사시 응답자가 무응답하는 경우나 정보수집자의 실수로 인하여 그룹변수에 대한 정보가 누락되는 경우를 예로 들 수 있는데 이런 경우 그룹변수가 관측된 개체들의 정보만을 이용하여 판별분류를 하는 것은 정보의 손실이 크고 잘못된 판별함수를 추정할 수 있기 때문에 바람직하지 못하다. 따라서 결측된 그룹변수를 적절한 방법을 통해 보완하여 분석하는 방법을 제안하고자 한다.

그룹변수의 결측치를 보완하기 위해서 2절에서 소개한 K-평균 군집방법을 사용할 수 있다. 즉, 그룹변수에 대한 정보가 없을 때 K-평균 군집분석을 통하여 개체들을 판별변수의 정보만으로 K개의 그룹으로 분류할 수 있으며, 개체들이 분류된 그룹에 대한 정보로 결측된 그룹변수의 값을 대체할 수 있다. 이러한 과정으로 그룹변수의 값을 보완한 후 2절에서 소개한 일반화 거리에 의한 판별분석을 실시하면 모든 개체의 정보를 이용할 수 있으므로 보다 정확한 판별함수를 찾을 수 있다.

이 때, 일반화 거리에 의한 판별분석 과정에서 부분집단 $i$ 의 사전확률 $\pi_i$ 를 개체와 집단의 중심까지의 거리의 비율로 계산한다면, 개체들이 각 집단으로 보다 잘 분류될 것이다. 즉, 부분 집단 $i$ 에 속할 사전확률(비율)을

$$\pi_i = D_i / (D_i + \sum_{j \neq i} D_j) \tag{3.1}$$

와 같이 정의할 수 있으며, 이 때 개체와 그룹 중심까지의 거리 $D_i$ 는 마할라노비스 거리(2.2)를 이용하여 구한다. 결측치가 있는 그룹변수의 판별과정을 요약하면 <표 3.1>과 같다.

<표 3.1> 결측치가 있는 그룹변수의 판별방법

| 제안된 알고리즘 |
|---|
| Step 1 > K-평균 군집분석을 통하여 개체들을 그룹의 수와 같은 K개의 집단으로 군집화 한다. |
| Step 2 > 그룹변수가 결측된 개체의 그룹변수의 값을 Step 1에서 분류된 군집의 값으로 대체한다. |
| Step 3 > Step 2에서 보완한 그룹의 중심과 개체들과의 거리(2.2)를 구한다. |
| Step 4 > Step 3의 결과를 이용하여 집단에 속할 비율(3.1)을 구하고 일반화된 거리 (2.3)을 이용한 판별방법으로 개체를 분류한다. |

### 3.2 신뢰도가 낮은 그룹변수

설문조사를 통해 관심 대상의 특성을 알아보고 대상을 집단으로 분류하고자 할 때, 응답자들이 실제 소속된 집단을 밝히지 않고자 다르게 응답하거나 분석자와 다른 기준으로 응답하여 실제 집단과 다른, 즉, 신뢰도가 낮은 그룹변수의 값을 얻게 되는 경우가 종종 있다. 이런 경우 수집된 자료를 그대로 분석하기보다는 실제의 집단을 추정하여 개체를 분류하는 것이 타당할 것이다.

실제의 집단을 추정하기 위하여 K-평균 군집방법을 사용할 수 있다. 즉, K-평균 군집분석을 통하여 개체들을 판별변수의 정보만으로 K개 보다 많은 그룹으로 분류하고 분류된 개체들의 그룹변수의 값을 고려하여 개체들을 재분류한다. 예를 들면, 첫 번째 군집에 속한 개체들의 그룹변수가 1의 값을 가장 많이 갖고 있으면 군집 내의 모든 개체를 그룹 1로 분류하고 그룹변수의 값은 1로 대체한다. 이런 과정으로 개체들을 재분류한 후에 일반화 거리에 의한 판별분석을 실시한다.

판별과정은 3.1절에서와 마찬가지로 각 부분집단들의 비율 $\pi_i$ 을 개체와 집단의 중심까지의 거리의 비율로 구하여 분류한다. 즉, 개체와 그룹중심까지의 마할라노비스 거리(2.2)를 이용하여 구한 비율(3.1)을 사용하여 일반화된 거리를 구하고 판별한다.

신뢰도가 낮은 그룹변수의 판별과정을 요약하면 <표 3.2>와 같다.

<표 3.2> 신뢰도가 낮은 그룹변수의 판별방법

| 제안된 알고리즘 |
|---|
| Step 1 > K-평균 군집방법을 통해 실제 집단보다 더 많은(실제 집단의 배수) 군집으로 개체들을 분류한다. |
| Step 2 > Step 1에서 분류된 각각의 군집에 속한 개체들을 그룹변수의 응답비율이 가장 높은 그룹으로 분류하고 그룹변수의 값을 재분류된 그룹의 값으로 대체한다. |
| Step 3 > Step 2에서 재분류된 그룹의 중심과 개체들의 거리(2.2)를 구한다. |
| Step 4 > Step 3의 결과를 이용하여 집단에 속할 비율(3.1)을 구하고 일반화된 거리 (2.3)을 이용한 판별방법으로 개체를 분류한다. |

## 4. 모의실험을 통한 제안된 방법의 평가 및 해석

본 장에서는 제안한 방법을 평가하기 위하여 앞장에서 고려한 두 가지 경우 - 그룹변수가 결측된 경우와 그룹변수의 신뢰도가 낮은 경우 - 에 대해 각각 SAS/IML을 사용하여 모의실험을 실시하고 그 결과를 기존의 방법과 비교, 평가하고자 한다. 논의의 간편성을 위하여 2 그룹 경우를 고려하였다.

### 4.1 그룹변수가 결측된 경우

3.1절에서 제안한 방법을 평가하기 위하여 3변량 정규분포를 따르는 판별변수( $X$ )를 200개 생성시켰고, 분류방법에 따라 판별하여 실제값( $Y$ )와 비교하였다. $f_1(X)$와 $f_2(X)$는 각각 다음과 같이 평균이 $\mu_1$, $\mu_2$이고 분산-공분산행렬 $\Sigma$를 갖는 3변량 정규분포를 가정하였다.

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix} \tag{4.1}$$

그룹변수는 판별변수의 분포와 대응하여 각각 100개씩 1과 2의 값을 갖도록 생성하였고, 그 후에 그룹 1에서 50%, 그룹 2에서 50%씩 무작위로 추출하여 변수의 값을 결측값으로 만들었다. 이 때 처음 생성한 그룹변수는 실제 그룹을 나타내는 것이고, 결측값으로 바꾼 그룹변수는 그룹정보가 누락된 자료의 값을 나타낸다.

새로운 판별방법의 평가를 위하여 (결측자료는 제외하고 분석한)기존의 방법과 (결측값을 표3.1과 같이 대체한)제안한 방법을 동일한 자료에 대해 수행하였고, 이러한 모의실험은 100회 반복 실시하였다. 모의실험 결과로부터 실제 그룹으로 분류된 개체들의 비율의 평균, 즉, 정분류율을 구하여 비교한 결과는 <표 4.1>과 같다.

<표 4.1> 결측치가 있는 그룹변수의 판별 결과

| 판별방법 | 정분류율 | 표준편차 |
|---|---|---|
| 기존의 방법 | 0.6879 | 0.0191 |
| 제안된 방법 | 0.8005 | 0.0738 |

그룹변수가 결측된 자료를 실제 그룹으로 분류하기 위하여 판별분석을 하였을 때, 기존의 방법의 정분류율 0.6879 보다 제안한 방법의 정분류율 0.8005 이 높음을 알 수 있다. 이 수치는 기존의 방법 보다 월등할 뿐만 아니라 일반적인 판별분류 분석의 결과로서도 좋은 결과라고 할 수 있다. 또한, 두 방법의 정분류율의 표준편차를 보면 제안된 방법의 표준편차가 더 크지만 모의실험결과 100번의 실험 중 92번의 실험에서 제안된 방법의 정분류율이 높게 나타났다. 따라서 제안된 방법이 기존의 방법보다 더

좋은 판별분류 결과를 제공하고 있다고 판단된다.

### 4.2 그룹변수의 신뢰도가 낮은 경우

3.2절에서 제안한 방법을 평가하기 위하여 3변량 정규분포를 따르는 판별변수( $X$ )를 200개 생성시켰고, 분류방법에 따라 판별한 후 실제값( $Y$ )과 비교하였다. $f_1(X)$와 $f_2(X)$는 식 (4.1)을 가정하였다. 그룹변수는 판별변수의 분포와 대응하여 각각 100개씩 1과 2의 값을 갖도록 생성하였고, 그 중 일정 비율의 개체를 무작위로 선택하여 그룹변수에 대하여 다른 값을 부여하였다. 즉, 그룹변수의 값이 1인 경우에는 2로, 2인 경우에는 1로 바꾸었다. 이 때, 처음 생성한 그룹변수는 실제 그룹( $Y$ )을 나타내는 것이고, 무작위로 선택하여 다른 값으로 대체한 그룹변수는 신뢰도가 낮은 그룹변수를 나타낸다. 신뢰도의 수준에 따른 결과를 살펴보기 위해 불완전의 비율을 10%부터 40%까지 5%씩 증가시켜 7가지 수준의 자료를 생성하였으며, 판별분석 전의 군집화 과정에서 군집의 수는 4개, 6개, 8개, 10개의 네 가지 경우에 대하여 고려하였다. 불완전 비율은 신뢰도의 개념으로 생각할 수 있는데, 불완전한 비율이 10%~40%라는 것은 그룹변수의 신뢰도가 60%~90%라는 것과 같은 개념이다.

제안된 새로운 방법을 평가하기 위하여 위와 같은 28가지 경우에 대하여 제안한 방법<표 3.2>를 수행하였고, 같은 자료에 대해 기존의 방법을 수행하였다. 이러한 모의실험은 각각의 경우에 대하여 동일하게 100회 실시하였다. 모의실험 결과로부터 실제 그룹으로 분류된 개체들의 비율의 평균, 즉, 정분류율을 구하여 비교한 결과는 <표 4.2>와 같다.

<표 4.2> 신뢰도가 낮은 그룹변수의 판별 결과

| 신뢰도 | 기존의 방법 | 제안된 방법(K=군집 수) | | | |
|---|---|---|---|---|---|
| | | K=4 | K=6 | K=8 | K=10 |
| 90% | 0.8403 | 0.7950 | 0.8274 | 0.8410 | 0.8485 |
| 85% | 0.8199 | 0.7818 | 0.8242 | 0.8398 | 0.8456 |
| 80% | 0.8099 | 0.7793 | 0.8160 | 0.8282 | 0.8483 |
| 75% | 0.7931 | 0.7790 | 0.8097 | 0.8253 | 0.8320 |
| 70% | 0.7594 | 0.7675 | 0.7873 | 0.8051 | 0.8103 |
| 65% | 0.4872 | 0.4847 | 0.5179 | 0.5000 | 0.4970 |
| 60% | 0.4901 | 0.5006 | 0.5180 | 0.4800 | 0.5167 |

<표 4.2>의 결과를 보면, 그룹변수의 신뢰도가 70%인 경우에는 군집의 수와 상관없이 제안된 방법의 정분유율이 높으며, 75%~95% 인 경우에는 군집의 수가 많아질수록 제안된 방법의 정분류율이 높아져 K=8 이상인 경우에는 기존의 방법보다 좋은

결과가 나타남을 알 수 있다.

군집의 수가 많을수록 정분류율이 높아지는 것은 개체들을 군집화한 후에 군집에 속한 개체들의 그룹변수의 응답 비율에 따라 군집에 속한 개체를 비율이 높은 그룹으로 할당하는 과정에서 응답한 것과 다른 그룹으로 할당되는 개체의 수가 감소되기 때문이다. 예를 들면, 어떤 개체가 그룹 1로 응답하였으나 개체가 속한 군집 내에서는 그룹 2로 응답한 비율이 높으면 그룹 2로 분류되는데 군집의 수가 많을수록 응답한 그룹과 다르게 분류되는 개체수가 감소한다는 것이다.

신뢰도가 60%~65%인 경우에는 제안된 방법과 기존의 방법 모두 낮은 정분류율을 보이고 있는데, 제안된 방법은 군집화한 후에 개체를 분류하는 과정에서 신뢰성이 낮은 그룹변수를 근거로 개체를 분류하기 때문이다. 따라서, 신뢰도가 65%이하인 실제 자료에 대하여 판별분류분석을 하였을 때, 정분류율이 낮다면 분석방법의 문제가 아니라 자료의 문제라고 봐야할 것이며, 좋은 결과를 얻기 위해서는 자료수집부터 다시 이뤄져야 할 것이다.

### 4.3 의사결정나무와의 비교

앞 절의 결과에 의하면 제안된 방법이 기존의 방법보다 더 좋은 결과를 제공하지만, 최근 많은 관심을 받고 있는 데이터 마이닝 방법과 그 결과를 비교하는 것 또한 의미가 있다고 생각된다. 분류를 목적으로 하는 데이터 마이닝 방법으로는 2.3절에서 언급한 의사결정나무(decision tree)에 의한 방법이 있으며 그룹변수가 이진(binary)인 경우에는 CART알고리즘에 의하여 의사결정나무를 구할 수 있다. 세 가지 분류방법-기존 방법, 제안된 방법, CART -의 분류결과를 비교하기 위하여 4.2절에서 생성하였던 동일한 자료에 대하여 CART알고리즘에 의한 의사결정나무 분석을 하였고, 그 결과로부터 정분류율과 정분류율의 표준편차를 구하였으며 결과는 <표 4.3>과 같다. <표 4.3>의 결과를 보면 신뢰도가 70%~90%의 모든 경우에 제안된 방법이 두 방법 - 기존의 방법과 CART - 보다 높은 정분류율을 나타내고 있고 있음을 알 수 있다.

<표 4.3> CART 알고리즘과의 판별 결과 비교

| 신뢰도 | 판별방법 | 정분류율 | 표준편차 |
|---|---|---|---|
| 90% | 기존의 방법 | 0.8061 | 0.0314 |
| | 제안된 방법 | 0.8485 | 0.0417 |
| | CART | 0.7619 | 0.0304 |
| 80% | 기존의 방법 | 0.7294 | 0.0343 |
| | 제안된 방법 | 0.8483 | 0.0402 |
| | CART | 0.7631 | 0.0287 |
| 70% | 기존의 방법 | 0.6270 | 0.0344 |
| | 제안된 방법 | 0.8103 | 0.0603 |
| | CART | 0.7636 | 0.0304 |

## 5. 결론 및 감사의 글

본 연구에서는 불완전 그룹변수를 수반하는 다변량 자료의 판별분류분석에서 불완전 그룹변수를 K-평균방법으로 사전 조정한 후 분석하는 방법에 대하여 논의하였다. 특히 문제가 되는 두가지 경우(결측치가 많은 경우와 신뢰할 수 없는 그룹변수의 경우)를 각각 고려하여 알고리즘을 제시하였으며 모의실험을 통해 제안된 방법의 실용성을 제시하였다. 본 연구의 확장을 통해 추후 좀더 정교한 알고리즘 개발이 이루어지기를 희망하며, 본 논문을 심사해 주신 심사위원님들께 감사를 드립니다.

## 참 고 문 헌

1. 강현철, 한상태, 최종후, 김은석, 김미경 (2000). *데이터마이닝 -방법론 및 활용-.* 자유아카데미
2. 김기영, 전명식 (1996). *SAS 군집분석.* 자유아카데미
3. 김기영, 전명식 (1997). *SAS 판별 및 분류분석.* 자유아카데미
4. 김보화, 김규성 (2002). K-모드 알고리즘과 ROCK 알고리즘의 개선. 「응용통계연구」, 제 15권 2호, 381-393.
5. 성웅현 (2000). *응용 다변량 분석.* 탐진
6. 허명회 (2000). 이중 K-평균 군집화. 「응용통계연구」, 제 13권 2호, 343-352.
7. R. A. Johnson, D. W. Wichern (2002). *Applied Multivariate Statistical Analysis.* Prentice Hall, N.Y.

[ 2004년 6월 접수, 2004년 8월 채택 ]