응용통계연구 제10권 2호
1997년 9월 pp. 261-274

Calibration 모형을 이용한 판별분석[1)]

이 석 훈[2)], 박 래 현[2)], 복 혜 영[3)]

요 약

기존에 제안되어온 판별분석 기법이 대상으로 하는 대부분의 자료는 각 개체가 어느 한 특정한 집단에 전적으로 소속되어 있는 것으로 국한되어 왔다. 그러나 오늘날 (0-1)의 이치논리가 퍼지(Fuzzy) 개념과 다치논리로 확장되는 현상은 어느 한 개체를 꼭 한 개의 집단에만 국한 시키는 관점 역시 변화를 요구하고 있다고 본다. 이에 본 논문에서는 한 개체가 어떤 소속확률을 갖고 여러개의 집단에 소속되어 있는 상황을 고려하여 이러한 개체들로 구성된 학습표본으로 부터 판별분석 규칙을 개발하는 것을 목표로 하였다. 방법론으로는 개체들의 특성벡터와 소속상태의 관계를 역추정(calibration) 모형으로 표현하고 판별대상개체의 특성벡터가 주어졌을 때 소속상태를 추정하도록 하며 이때 추정은 베이지안 방법, Metropolis 알고리즘 등을 사용하였다. 또한 제안된 판별규칙의 평가를 위한 기준을 제안하고 두 개의 자료를 기존의 다른 규칙들과 함께 분석하여 결과를 비교하였다.

1. 서 론

우리가 관심을 갖고 있는 대상들을 어떤 관점에서 분류한다는 것은 인간의 판단과 인식 체계에서 나타나는 가장 근본적이며 자연적인 현상 중의 하나이다. 이러한 분류에 관한 연구는 탐사적 자료분석을 필요로 하는 모든 과학 분야에서 일반적으로 또는 개별적으로 수행되고 있다.

대표적으로는 문자인식, 음성인식, 영상인식 등을 주로 다루고 있는 패턴인식(pattern recongnition) 분야를 비롯하여 환자의 상태를 분류하는 질병분류, 생물학 계통의 품종분류, 사회과학 분야의 구성원의 집단 분류 등을 꼽을 수 있다. 이러한 분류의 영역에서 특별히 판별분석은 p개의 변수(특성변수)로 표현되는 개체들의 특성인 특성벡터(feature vector) $\underline{y}=(y_1, y_2, \cdots, y_p)'$와 그 개체들이 미리 알려져 있는 (k+1)개의 집단 $G_1, G_2, \cdots, G_{k+1}$에 소속된 상태와의 관계를 발견하여 특성변수들 사이의 구조적 관계를 파악하거나 또는 소속불명의 개체가 이들 집단에 소속된 소속상태를 예측하는데 필요로 되는 모형의 개발, 구현, 평가에 관한 일련의 과정이라고 할 수 있다. 이에 대한 연구 역시 분류를 요구하는 학문 분야가 대단히 광범위하기 때문에 여러 분야에서 많은 연구가 오랫동안 수행되어 왔는데 이들 대부분의 연구가 다루고 있는 주제(issue)는 일반적인 선형(비선형) 결정함수를 사용하는 방법, 우도함수를

1) 이 연구는 96년도 한국과학재단 연구비지원에 의한 결과임. (과제번호 : 961-0105-024-1)
2) (305-764) 대전광역시 유성구 궁동 220 충남대학교 자연과학대학 통계학과, 교수.
3) (305-764) 대전광역시 유성구 궁동 220 충남대학교 자연과학대학 통계학과, 연구조교.

사용하는 방법, 비모수적 접근법 그리고 일반화된 거리함수의 정의등 집단의 모형에 관한 것이 주를 이루고 있는 것을 볼 수 있다. (McLACHLAN(1992))

본 연구에서는 기존의 연구 흐름과는 관점을 달리하여 각 개체가 집단에 소속되는 상태를 주목하고 기존의 대부분의 연구에서는 어떤 임의의 개체가 (k+1)개의 집단중에서 어느 하나의 집단에만 속하는 현상만을 연구대상으로 하였던 것을 확장하여 한 개체가 두 개이상의 집단에 부분적으로 소속되어 있는 상황을 고려하고자 한다. 이러한 상황은 경영학의 시장 조사론이나 Ripley(1994)가 언급한 전문가에 의하여 소속집단이 판단되는 경우에 쉽게 발견된다. 그럼에도 불구하고 이러한 상황이 통계학의 문제 영역에서 자주 제기되지 않았던 데에는 여러가지 이유가 있겠으나 그중 하나는 아직도 많은 학문 영역에서 종합적 인식 방법보다는 분석적 방법을 우선하고 있는 파라다임을 들 수 있다고 본다. 한편 최근에 퍼지 개념이 일부에서 적극적으로 도입되고 있는 현상이나 또는 기존의 분석적 방법에 대한 비판의 시각은 이치(bivalent) 논리를 근본으로하는 파라다임에 대한 반론을 제기하고 있는데 이와같은 분위기 역시 본 연구의 주제에 대한 필요성을 시사하고 있다. (Kosko(1993))

이러한 관점에서 개체 i 가 집단 l 에 소속될 확률을 p_{il} 이라고 하고 개체 i 의 각 집단에 속할 소속상태를 소속벡터 $\underline{p_i}=(p_{i1}, p_{i2}, \cdots, p_{ik})'$로 나타내면 본 연구는 소속벡터 $\underline{p}$ 와 특성벡터 $\underline{y}$의 관계를 모형화하는 단계와 이를 바탕으로 소속 불명인 분류대상 개체의 소속벡터를 추정하는 방법을 개발하는 단계, 그리고 제안된 모형과 추정방법의 특성을 고려하는 것을 주된 내용으로 한다.

2절에서는 본 연구에서 고려하는 자료의 판별분석을 위한 역추정(Calibration) 모형을 제안하고 그 특성을 토의하며 3절에서는 제안된 모형을 이용하여 자료를 분석할 때 요구되는 모수의 베이지안적 추론과정을 논하고, 이를 바탕으로 Metropolis 알고리즘 구현을 통한 각 집단의 소속확률의 신뢰구간을 구하는 방법 그리고 소속상태의 일반화 최우추정치를 구하는 방법을 개발하여 새로운 개체의 소속상태인 소속벡터의 추정을 논의한다. 4절에서는 제안된 판별규칙의 평가를 위한 기준을 제안하고 5절에서는 두가지의 서로 다른 형태의 자료에 대하여 역추정 모형을 사용한 결과와 로지스틱(Logistic) 모형으로 분석한 결과 그리고 일반화된 거리함수를 사용한 결과를 비교 토의 하고 6절에서는 결론을 맺는다.

2. 모형의 제안

판별분석을 위한 기존의 대부분의 모형이 소속상태를 특성벡터의 함수로 보는 로지스틱 판별분석 모형을 제외하고는 모두 다 개체가 어느 특정한 하나의 집단에만 속하는 현상 ((0-1) 현상)을 다루고 있어서 모수적이든 비모수적이든 각 집단들을 표현하기 위한 확률모형과 오류율(error rate)의 정의를 통한 결정론적 규칙에 관하여 주로 논의가 집중되어 왔다. 로지스틱 모형도 개체의 소속상태를 (0-1) 현상으로 제한하여 연구되었지만 이 모형은 Ripley(1994)에 의해서 논의된 바와 같이 신경회로망 모형의 단순 경우(trivial case)로 생각할 수 있어서 본 연구에서 고려하고 있는 소속상태를 갖는 자료의 분석을 위하여도 사용될 수는 있다. 그러나 모든 신경회로망 모형이 갖고 있는 약점인 일반화(generalization)가 어렵다는 점과 최우추정량의 근사적 특성을 제외하고는 추론을 위한 특별한 근거를 제시하지 못한다는 점을 한계로 갖고 있

다. 이절에서는 이러한 관점에서 지적된 기존 모형들의 한계를 극복하기 위한 모형을 제안하고 그 특성을 조사하고자 한다.

임의의 한 개체의 특성벡터 $\underline{y}=(y_1, y_2, \cdots, y_p)'$와 그 개체가 k+1개 집단에 소속된 상태를 나타내는 소속벡터를 $\underline{p}=(p_1, p_2, \cdots, p_k)'$라고 하자. 여기서 이 개체가 집단 k+1에 속할 확률 p_{k+1}은 $1-\sum_{l=1}^{k} p_l$이 되므로 소속벡터를 k차원 벡터로 나타낸다.

본연구에서 고찰하려는 모형은 임의의 한 개체의 특성벡터 $\underline{y}$와 각 집단의 중심과의 통계적 거리(statistical distance) $\underline{Dy}$를 이 개체의 소속벡터 $\underline{p}$의 선형함수로 정의한다. 이를 수리적으로 표현하면,

$$\underline{Dy}=\underline{\alpha}+\underline{\beta_1}p_1+\cdots+\underline{\beta_k}p_k+\underline{\varepsilon} \tag{2.1}$$

이고 여기서 $\underline{Dy}, \underline{\alpha}, \underline{\beta_l}, l=1, \cdots, k$ 은 모두 k+1차원 벡터이고, $\underline{\varepsilon}$은 평균이 $\underline{0}$이고 분산공분산 행렬이 Σ인 k+1차원 다변량 정규분포로 가정한다. 대부분의 판별분석 규칙이 특성변수들의 결합(선형 또는 비선형)을 이용하고 있거나 또는 특성변수들의 결합과 소속상태의 함수관계를 사용하고 있다는 점을 고려해 볼 때 본 연구에서 제안하고 있는 모형의 특징은 개체의 특성치와 각 집단의 중심(centroid)과의 통계적 거리를 개체의 정보로 본 것과 이 정보와 소속상태의 관계를 모형화하였다는 데에서 찾을 수 있다.

3. 추정

크기가 n인 학습표본 T에서 특성벡터에 관한 집단 l의 중심 $\underline{\mu_l}$의 추정량 $\underline{c_l}$과 개체 i의 특성벡터 $\underline{y_i}$와의 일반화 거리의 관찰값 $\underline{Dy_i}$는 다음과 같이 정의한다.

$$T=\left\{(\underline{y_i}, \underline{p_i}), i=1, \cdots, n\ ;\ \underline{p_i}=(p_{i1}, p_{i2}, \cdots, p_{ik})',\ 0\le p_{il}\le 1,\ \sum_{l=1}^{k} p_{il}\le 1\right\}$$

$$\underline{c_l}=(c_{l1}, c_{l2}, \cdots, c_{lp})'=\frac{\sum_{i=1}^{n} p_{il}\underline{y_i}}{\sum_{i=1}^{n} p_{il}}\ ,\qquad \underline{Dy_i}=(Dy_{i1}, \cdots, Dy_{ip})'$$

여기서 $Dy_{il}=(\underline{y_i}-\underline{c_l})'S_l^{-1}(\underline{y_i}-\underline{c_l})$이고 이때 S_l은 집단 l의 표본 분산공분산 행렬로서 (r, q)항 $S_l(r, q)$는 다음과 같다.

$$S_l(r, q)=\frac{\sum_{i=1}^{n}(y_{ip}-c_{lp})(y_{iq}-c_{lq})p_{il}^2}{\sum_{i=1}^{n} p_{il}^2}$$

학습표본 T를 바탕으로 2절에서 제안한 모형(2.1)을 이용하여 판별분석의 규칙을 수립하는 것은 모형에 포함된 모수를 추정하는 것과 소속불명의 개체가 특성벡터 $\underline{y_0}$로 관찰되었을 때

이 개체의 소속상태 $\underline{p_0}=(p_{01}, p_{02}, \cdots, p_{0k})'$을 추정하는 것을 의미한다. 따라서 이는 단순한 선형모형과 관련된 추정문제가 아니라, 다변량 역추정문제가 되고, 이에 맞는 추정방법이 모색되어야 한다. 다변량 역추정 모형의 추정에 관하여는 여러분야에서 많은 연구가 수행되었는데 본 연구와 관련되어서는 특별히 Brown(1982), Oman과 Wax(1984), Brown과 Sundberg(1987), Oman(1988), 박래현과 이석훈(1990), 이석훈 등(1990)의 연구 등을 들 수 있다.

접근방식으로는 크게 베이지안 접근과 비베이지안 접근으로 구별되는데 최우추정법과 Fieller의 방법을 대표로 하는 비베이지안 방법은 $\underline{p_0}$의 구간추정시 부등식을 풀어야 하는 관계로 $\underline{p_0}$ 의 각 성분 p_{0l}의 구간추정을 힘들게 할 뿐만아니라 쓸모있는 폐쇄형 신뢰구간이 어떤 조건하에서만 구해져서 정확한 신뢰수준을 줄 수 없게 된다. 이와같은 문제점을 해결하기 위해 우리는 $\underline{p_0}$ 또는 p_{0l}을 직접적으로 다루는 베이지안 방법을 사용하였다.

3.1 베이지안 추정

학습표본 T와 소속불명의 개체 $\underline{y_0}$을 2절의 모형을 사용하여 다시 표현하면 다음과 같이 된다.

$$\begin{cases} D=ZF+E \\ \underline{d_0}=F'\underline{p_0^*}+\underline{\varepsilon_0} \end{cases} \tag{3.1}$$

여기서,

$$D=\begin{pmatrix} Dy_{11} & \cdots & Dy_{1k+1} \\ \vdots & & \vdots \\ Dy_{n1} & \cdots & Dy_{nk+1} \end{pmatrix}, \quad Z=\begin{pmatrix} 1 & p_{11} & \cdots & p_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & p_{n1} & \cdots & p_{nk} \end{pmatrix}, \quad F=\begin{pmatrix} \alpha_1 & \cdots & \alpha_{k+1} \\ \beta_{11} & \cdots & \beta_{1k+1} \\ \vdots & & \vdots \\ \beta_{k1} & \cdots & \beta_{kk+1} \end{pmatrix}, \quad E=\begin{pmatrix} \underline{\varepsilon_1}' \\ \vdots \\ \underline{\varepsilon_n}' \end{pmatrix},$$

$$\underline{d_0}=\begin{pmatrix} Dy_{01} \\ \vdots \\ Dy_{0k+1} \end{pmatrix}, \quad \underline{p_0}=(p_{01}, \cdots, p_{0k})', \quad \underline{p_0^*}=(1 \quad \underline{p_0}')', \quad \underline{\varepsilon_0}=(\varepsilon_{01}, \cdots, \varepsilon_{0k+1})'$$

위 모형 (3.1)에서 데이터 D 와 Z 는 이미 알려진 정보, 즉 학습표본(training sample)이고 이를 근거로 D 와 Z 의 회귀 방정식을 통하여 F 와 Σ의 추정치를 결정하여 판별규칙(rule)을 제공하게 된다. 새로운 개체에 대한 판별분석 문제는 역추정 모형을 통하여 $\hat{F}$ 과 $\hat{\Sigma}$ 이 주어지게 되고 이때 새로운 데이터 $\underline{d_0}$가 관찰되었을 때 미지의 소속벡터 $\underline{p_0}$ 를 추론하는 문제로 요약될 수 있다. 모형 (3.1)에 최우추정법(maximum likelihood estimation)을 사용하여 관련된 모수들의 최우추정량을 구하면 다음과 같이 구해진다.

$$\hat{F}=\begin{pmatrix} \widehat{\alpha}' \\ \widehat{B} \end{pmatrix}=(Z'Z)^{-1}ZD$$

$$\hat{\Sigma} = S = \frac{(D - Z\hat{F})'(D - Z\hat{F})}{n-k-1}$$

이와 같이 구해진 최우추정량 $\hat{F}$ 과 $\hat{\Sigma}$ 을 사용하여 $\underline{p}_0$에 대한 추론을 하게 되는데, 본 연구에서는 기본적으로 일변량 역추정에서의 베이지안 추론을 다룬 Hoadley(1970) 방법을 다변량의 경우로 확장한 Brown(1982) 방법을 사용한다.

모수 F와 분산공분산 행렬 Σ는 개체 $\underline{d}_0$에 대응하는 소속상태 $\underline{p}_0$와 서로 독립이며 비정보적인 사전분포를 따르고 $\underline{p}_0$의 사전분포 $f(\underline{p}_0)$는 p_{0l} 이 구간 [0,1]에 있도록 하기 위해 베타분포의 확장 형태인 Dirichlet 분포라고 가정하여 얻은 소속상태 $\underline{p}_0$의 사후분포는 다음과 같다.

$$\begin{aligned} f(\underline{p}_0 \mid \underline{d}_0, D, Z) &\propto f(\underline{p}_0) L(\underline{d}_0 \mid \underline{p}_0, D, Z) \\ &\propto p_{01}^{\tau_1 - 1} \cdots p_{0k}^{\tau_k - 1} (1 - p_{01} - \cdots - p_{0k})^{\tau_{k+1} - 1} \\ &\times \frac{\{\sigma^2(\underline{p}_0)\}^{\frac{\upsilon}{2}}}{\{\sigma^2(\underline{p}_0) + (\underline{d}_0 - \hat{\alpha} - \hat{B}'\underline{p}_0)' S^{-1} (\underline{d}_0 - \hat{\alpha} - \hat{B}'\underline{p}_0)\}^{\frac{\upsilon + k + 1}{2}}} \end{aligned} \tag{3.2}$$

여기서,

$$\sigma^2(\underline{p}_0) = 1 + \frac{1}{n} + (\underline{p}_0 - \overline{\underline{p}})'(X'X)^{-1}(\underline{p}_0 - \overline{\underline{p}}), \qquad X = \begin{pmatrix} p_{11} - \overline{p_1} & \cdots & p_{1k} - \overline{p_k} \\ \vdots & & \vdots \\ p_{n1} - \overline{p_1} & \cdots & p_{nk} - \overline{p_k} \end{pmatrix},$$

$$\overline{p_j} = \sum_{i=1}^{n} \frac{p_{ij}}{n}, \ (j = 1, \cdots, k), \qquad \overline{\underline{p}} = (\overline{p_1}, \overline{p_2}, \cdots, \overline{p_k})', \qquad \upsilon = n - 1$$

이다. 위에서 $L(\underline{d}_0 \mid \underline{p}_0, D, Z)$ 는 $\underline{p}_0$ 가 주어졌을 때 $\underline{d}_0$ 의 예측분포(predictive distribution)이다.

사전분포에 포함되어 있는 모수 τ_l $(l = 1, 2, \cdots, k+1)$에 대하여는 Dirichlet 분포의 특성으로 부터 두 가지 입장이 있는데, 하나는 분류대상 개체가 어떤 특정집단에 부분적으로 속한 것으로 보기 보다는 모든 집단에 속할 가능성이 똑같은 것으로 보는 입장으로 모든 τ 값을 $\frac{1}{k+1}$로 놓고, 다른 하나는 분류대상 개체가 여러 집단에 부분적으로 속한 것으로 보는 입장으로 연구자의 사전정보가 특별히 어느 특정한 집단을 강하게 생각하고 있으면 그 집단에 대응되는 τ 값을 1이상으로 놓고 다른 τ 값을 1이하로 놓는다. 본논문에서는 기존의 (0-1)현상의 자료들의 분석과 비교하기 위하여 τ 값을 모두 $\frac{1}{k+1}$로 놓는 입장을 취하였다.

식 (3.2)에 의해서 유도된 $\underline{p}_0$의 사후분포로부터 여러 가지 특성을 구하는 것은 상당히 복잡한 적분을 요구하고 특히 어떤 특정한 집단에 속하는 확률 p_{0l}에 대한 구간추정은 주변사후분포를 필요로 하는데 이것 역시 쉽게 다룰 수가 없다. 이와같은 어려움을 극복하기 위하여 본

논문에서는 Metropolis 알고리즘을 사용하였다.

3.2 소속상태에 관한 구간추정

소속상태에 관한 구간추정의 논의는 소속확률의 비에 대한 대수값의 구간추정을 연구한 Dawid(1976), Critchley 등(1984), Hirst 등(1990)에서 찾아 볼 수 있는데 이들의 연구 역시 모두 (0-1) 현상의 자료를 근거로 수행되었기 때문에 본연구에서 고려하고 있는 자료까지 포함하는 학습표본의 경우에는 적용하는 데에 한계가 있고, 또한 대표본에 근거한 근사방법이므로 소표본일 경우의 적용에 한계가 있다. 반면에 본연구에서 제안한 모형에서는 3.1절에서 제시된 소속상태 $\underline{p_0}$ 에 대한 사후분포로 부터 자연스럽게 집단 l 에 소속될 확률 p_{0l} 에 대한 신뢰구간을 - 엄밀한 의미에서 HPD (Highest Posterior Density) 영역 - 구할 수 있게 되고 소표본일 경우도 적용에 문제가 없다. 실제적인 계산에서 사후분포의 복잡한 함수 형태로 부터 주변분포를 얻기가 어려운 면은 이석훈과 이원돈(1989)과 Lee(1995) 등이 토의한 바 있고 Tanner(1993)가 다른 여러 가지 기법들과 함께 요약 정리한 Metropolis 등(1953)의 알고리즘을 변형하여 사용함으로써 쉽게 극복하였다. 먼저 사후분포 (3.2)를 다음과 같이 나타내자.
단 $h(\underline{d_0}, \underline{p_0} \mid D, Z) = f(\underline{p_0}) L(\underline{d_0} \mid \underline{p_0}, D, Z)$ 이다.

$$f(\underline{p_0} \mid \underline{d_0}, D, Z) = \frac{h(\underline{d_0}, \underline{p_0} \mid D, Z)}{\int h(\underline{d_0}, \underline{p_0} \mid D, Z) d\underline{p_0}} \tag{3.3}$$

본 연구에서 Metropolis 알고리즘을 이용한 모의표본추출 과정은 다음과 같다.

[단계 1] $p_{01}, \cdots, p_{0k}$의 초기치를 $p_{01}^{(0)}, \cdots, p_{0k}^{(0)}$라 할 때 $h(\underline{p_0^{(0)}})$를 구한다.

[단계 2] $l = 1, 2, \cdots, k$에 대하여 [단계 2.1]과 [단계 2.2]를 순서대로 수행한다.

[단계 2.1] (1) $p_{0l}^{(0)}$ 의 로짓 변환을 $z = \ln(p_{0l}^{(0)}/(1-p_{0l}^{(0)}))$라고 할때 정규분포 $N(z\ ,\ 0.3)$에서 난수 x 를 추출한다.

(2) (1)에서 구한 x 를 다시 역로짓 변환시킨 값을 x^* 라 하고 x^* 가 $\sum_{m=1}^{l-1} p_{0m}^{(1)} + x^* + \sum_{m=l+1}^{k} p_{0m}^{(0)} \le 1$ 을 만족하는지를 조사한다.
만족하면 (3)으로 가고, 만족하지 않으면 (1)로 돌아간다.

(3) $E(x^*) = h(p_{01}^{(1)}, \cdots, p_{0l-1}^{(1)}, x^*, p_{0l+1}^{(0)}, \cdots, p_{0k}^{(0)})$ 를 계산한다.

[단계 2.2] (1) $E(x^*)$와 $E(p_{0l}^{(0)})$을 비교한다.
$E(x^*) > E(p_{0l}^{(0)})$ 이면 (3)으로 가고, 그렇지 않으면 (2)로 간다.

(2) 일양분포 $U(0,1)$에서 난수 y 를 추출하여 y 값과 $R = E(x^*)/E(p_{0l}^{(0)})$을 비교한다.
$y \le R$이면 (3)으로 가고 그렇지 않으면 [단계 2.1]로 돌아간다.

(3) $p_{0l}^{(1)}$을 x^*로 놓는다.

[단계 3] 초기값 $\underline{p_0^{(0)}}$ 로 부터 첫 번째 이동한 값인 소속벡터 $\underline{p_0^{(1)}}$ 를 구한 것과 같이 다시 $\underline{p_0^{(1)}}$ 으로 부터 $\underline{p_0^{(2)}}$를 구하는 방식으로 원하는 표본 크기가 달성될 때까지 반복 수행한다.

[단계 2.1]에서 로짓변환을 사용한 것은 일반적인 일변량분포를 사용할 경우 p_{0l}이 0이나 1에 가까운 값인 경우 조건에 맞는 난수발생에 상당한 시간이 소요된다. 따라서 가급적 현재의 위치에서 가능한 가까운 값으로 이동하도록 하는 로짓변환을 사용하였고 그 결과 주어진 조건을 만족시키면서 표본추출을 빨리 수행할 수 있었다. 또한 표준편차를 0.5이상으로 하는 것이 무난하리라 생각하여 분산을 0.3으로 정하였다. 한편 표본의 크기에 관하여는 1000회 반복 후부터 100회 추가 때마다 추출된 모의자료로부터 가장 관심있는 집단의 소속확률 p_{0l}에 대한 최단신뢰구간을 구하여 구간의 거리가 줄어드는 정도가 $\varepsilon(2\times10^{-3})$ 이하일 때 반복을 마치고 그때의 신뢰구간을 구간추정치로 받아들이기로 하였다. 이때 최단신뢰구간을 구하는 방법은 p_{0l}의 주변사후분포가 좌우대칭이 아니므로 모의표본을 크기순으로 정렬하여 표본의 95%를 포함하는 최단구간을 구하는 방법을 사용하였다.

3.3 일반화 최우추정치

한편 소속상태 $\underline{p_0}$에 대한 일반화 최우추정치는 현학적(heuristic)인 방법으로 구하였는데 첫 단계로 $\sum_{l=1}^{k} p_{0l} \leq 1$을 만족하는 $\underline{p_0}$의 가능한 범위의 값들에 일정 간격의 크기를 갖는 성긴 격자점들을 잡아 각 격자점에서 식 (3.3)의 $h(\underline{d_0}, \underline{p_0} \mid D, Z)$의 값들을 계산하여 그 값이 최대인 격자점의 근접 영역을 찾아 옮겨 간다. 둘째 단계로 옮겨진 근접 영역에서 첫 단계와 마찬가지로 일정한 간격의 크기를 갖는 세밀한 격자점을 잡아 다시 $h(\underline{d_0}, \underline{p_0} \mid D, Z)$을 계산해 그 값을 최대로 하는 격자점의 근접 영역을 다시 잡아서 같은 작업을 반복한다. 이때 격자 간격이 우리가 목표하는 값 (10^{-3}) 보다 작으면 그때의 격자점을 일반화 최우추정치, $\underline{\hat{p}_0}$ 로 택하였다.

4. 모형의 평가

일반적인 오류율을 이용한 판별규칙의 평가는 대부분이 (0-1) 현상의 자료분석에 관한 것이기 때문에 본 연구에서 고려하는 형태의 자료를 분석하는 판별규칙도 평가할 수 있는 기준을 제시할 필요가 있다. 특별히 이 기준은 (0-1) 현상의 자료분석에서 사용된 일반적인 오류율을 포함하는 것이 되도록 하는 것이 바람직할 것이다. 이에 본 연구에서는 Brown(1982)이 역추정 모형의 평가에 사용했던 기준을 변형하여 판별분석 모형의 일반적인 평가기준을 다음과 같이 제안하고자 한다.

$$\text{부정확성} = \frac{\sum_{i=1}^{m}\sum_{l=1}^{k+1}(p_{il}-\widehat{p_{il}})^2}{\sum_{i=1}^{m}\sum_{l=1}^{k+1}(p_{il}-\frac{1}{k+1})^2} \times \frac{k}{2(k+1)}$$

여기서 m은 평가에 사용되는 소속상태가 $\underline{p_i}=(p_{i1}, p_{i2}, \cdots, p_{ik})$, $p_{ik+1}=1-\sum_{l=1}^{k} p_{il}$ 로 알려진 개체의 수이고 $\widehat{p_{il}}$은 개체 i가 모형에 의하여 추정된 집단 l 에 소속될 확률이다. 부정확성의 직관적인 의미를 살펴보면 첫째항의 분자는 추정치와 실제 값과의 차이의 제곱이고 분모는 개체가 각 집단에 동등하게 소속되어 있다고 볼 때, 즉 판별규칙을 고려하지 않고 임의로 할당할 때 소속확률 $\frac{1}{k+1}$과 실제 값과의 차이의 제곱을 합한 값으로 하였다. 둘째항은 부정확성이 (0-1) 현상의 자료에서 사용될 때 일반적으로 사용되는 기준인 오류율을 나타내기 위한 정규화 상수로서 부정확성이 (0-1) 현상의 자료분석에서는 0과 1사이의 일반적인 오류율을 나타내게 된다.

부정확성의 일반적인 특징은 그 값이 $\frac{k}{2(k+1)}$이하에서만 주어진 판별규칙이 최소한도의 의미를 갖게 되어 부정확성이 $\frac{k}{2(k+1)}$보다 크게되면 분류를 포기하는 것, 즉 주어진 판별규칙은 사용하지 않는 것이 좋다는 평가를 내리게 된다.

5. 실례와 비교

본연구에서 제시된 역추정 모형을 이용하여 두 가지 서로 다른 형태를 갖는 자료를 분석하였다. 이때 비교를 위하여 (0-1) 현상을 나타내는 자료의 분석에 보통 사용되는 일반화 거리 개념을 이용한 판별분석기법과 신경회로망의 가장 단순한 형태로서 로지스틱 판별분석을 함께 사용하였다.

<그림 1>로 표현된 자료 1은 Johnson 과 Wichern(1992)의 567쪽에 있는 Admission Data for Graduate School of Business로서 학습표본내의 개체가 모두 어느 한 집단에만 속해있는 (0-1) 상태의 자료가 세 집단으로 평행하게 분류되어 있다. <그림 2>로 나타난 자료 2는 Rousseeuw(1995)가 퍼지 집락분석을 위하여 만든 자료로서 세 집단이 평면상에서 하나의 이상점과 다른 하나의 중간점을 제외하고는 서로 거의 등간격으로 나누어져 있는 자료인데 이때 각 개체의 집락분석 결과인 소속확률을 소속상태 $\underline{p}$로 인정하였다. 또한 비교를 위하여 소속확률중 가장 큰값을 갖는 집단에 전적으로 속한 것으로 간주하여 (0-1) 현상의 자료로 변환한 후 분석을 시도하였다.

자료 1에서는 Johnson 과 Wichern(1992)에서 SAS로 분석한 결과와 모형 (2.1)에 의한 결과를 비교해 볼 때 <표 1.1>에서와 같이 허가(Admit) 집단의 개체 2, 24, 31을 경계선(Borderline) 집단으로 분류하여 SAS에 의한 분석에서 4개가 오분류된 데 비해 하나(개체 3)가 적은 3개를 오분류하였고, 불허가(Not Admit) 집단의 개체중 58, 59를 SAS에서와 같이 경계선

집단으로, 경계선 집단의 개체 75를(SAS에서는 개체 66을) 불허가 집단으로 오분류하였다. 그러나 소속상태를 나타내는 $\underline{p}$ 값을 SAS에서 제공하는 사후확률과 비교하여 검토하면 오분류된 6개의 개체 모두에 대하여 모형 (2.1)에 의한 결과가 실제 소속된 집단에 소속될 확률을 더 크게 보여 주고 있다. 이러한 오분류의 문제는 특별히 <표 1.2>에 나타난 신뢰구간을 보면, 개체 24와 31의 경우 허가 집단에 속할 확률 p_3의 신뢰구간이 0.5를 포함하고 있으며 개체 58, 개체 75 역시 비록 분류는 경계선 및 불허가로 추정되었으나 자신들이 실제 속해 있는 집단에 속할 확률 p_1과 p_2의 신뢰구간 역시 0.5를 포함하고 있기 때문에 이들 개체에 대하여는 보다 면밀한 추정작업이 요구됨을 정량적으로 보여준다.

자료 2에 대하여 모형 (2.1)과 로지스틱 판별분석 모형을 각각 적용하여 <표 2>의 결과를 얻었고 두 모형의 비교를 위하여 4절에서 제시한 기준을 사용할 때 모형 (2.1)을 사용한 경우 부정확성이 0.0139로서 로지스틱 모형의 0.0339보다 작은 값을 보여주었다. 이 결과는 로지스틱 모형이 소속확률을 특성변수의 비선형 함수로 정의한 점과 모형 (2.1)의 선형성을 고려할 때 모형 (2.1)이 대단히 우수한 성질을 갖는다고 말할수 있다. 한편 (0-1) 현상의 자료로 변환하여 분석한 결과도 <표 2>에 나타내었는데 로지스틱 모형이 22개 모든 개체에 대하여 0과 1로 분류한 반면 모형 (2.1)은 집단들 사이에 끼어 있는 개체 6, 13에 대하여는 그 소속정도를 확률로서 보여줌으로써 어떤 개체를 세 집단중 어느 하나로 분류한다고 하더라도 그 소속확률에 대한 정보를 추가적으로 제공하게 된다.

또한 학습표본과 시험표본(test sample)이 같기 때문에 나타나는 과대평가의 가능성을 배제하기 위한 평가를 자료 2에 대하여 다음과 같이 시도 하였다. 22개의 자료중에서 약 30%에 해당되는 6개의 개체를 임의로 추출하여 시험표본을 구성하고 나머지 16개의 개체들을 학습표본으로 하여 정해지는 판별규칙을 사용하여 이들 6개의 시험표본내의 개체들의 소속확률을 예측하고 그 실제값을 이용하여 4절에서 제기한 부정확성을 구하였다. 이 과정을 100회 실시한 결과는 모형 2.1의 평균 부정확성은 0.0088, 표준편차는 0.0057이고 로지스틱 모형의 평균 부정확성은 0.0163, 표준편차는 0.0097를 나타내어 학습표본을 모두 시험표본으로 사용한 평가결과와 같이 모형 2.1이 로지스틱 모형 보다 더 높은 예측력을 갖고 있음을 보여 준다.

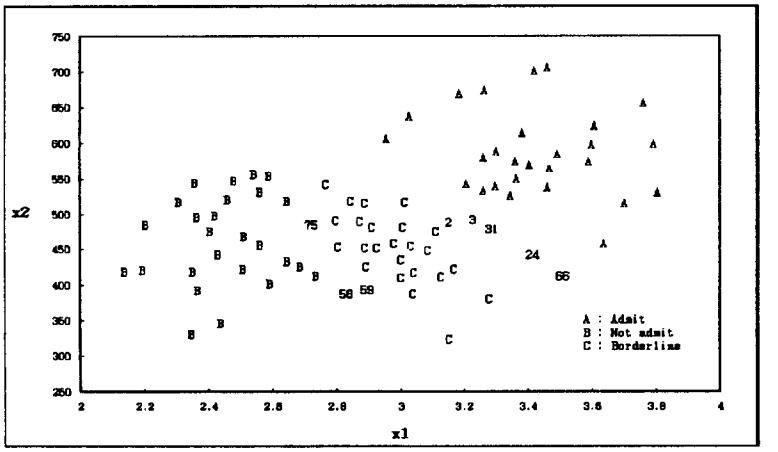
6. 결론

기존의 분류기법들이 (0-1)방식으로만 분류된 개체들을 바탕으로 개발된 것인데 반하여 본 연구에서는 어떤 소속확률을 갖고 여러개의 집단에 소속되어 있는 개체들을 바탕으로 하는 분류기법을 개발 제안하였다. 기본 모형은 개체들의 특성벡터 또는 개체들의 특성벡터가 각 집단의 중심으로 부터 떨어진 거리들을 소속확률의 선형 결합으로 하였고 미분류된 새로운 개체의 소속확률의 추정을 역추정 문제로 보았다.

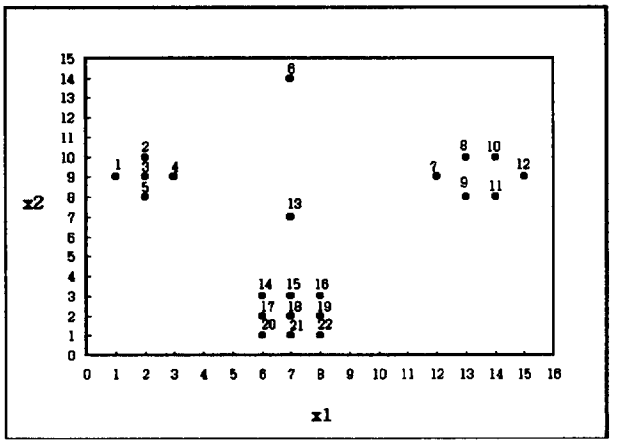
연구 내용면에서는 추정과정에서 베이지안 방법을 사용하여 사후분포의 분석과정을 논의하였고 이때 신뢰구간을 위하여 주변분포의 추정에 Metropolis 알고리즘을 사용하였다. 한편 모형의 특성 및 비교를 위하여 평가기준을 제시하고 구체적인 실례적용을 위하여 두 개의 자료를 제안된 모형 이외에도 SAS에서 사용하는 일반화된 거리개념에 의한 분류기법과 로지스틱 분석

기법을 사용하여 분석하였는데 이때 제안된 모형이 선형모형임에도 불구하고 비선형모형에 바탕을 둔 로지스틱 기법보다 제안된 평가기준에서 우수한 것으로 나타났다.

향후 연구과제로는 이 연구에서 사용된 특성벡터의 중심으로 부터의 거리개념을 기타 다른 모형들에도 적용시켜 보는 것과, 제안한 모형에 관련된 이론적 특성을 조사하는 것이라 하겠다.



<그림 1> 허가, 불허가, 경계선으로 분류된 Admission Data
for Graduate School of Business (x1=GPA, x2=GMAT)



<그림 2> 두 개의 중간 개체를 갖는 퍼지 집락분석을 위한 자료

<표 1.1> 오분류된 개체의 소속확률의 사후확률

관찰치	From	To	SAS 결과			모형 (2.1)		
			Admit	Border	Notadmit	Admit	Border	Notadmit
2	admit	border	0.120	0.878	0.002	0.351	0.648	0.001
3	admit	border	0.365	0.634	0.001	****	****	****
24	admit	border	0.477	0.523	0.000	0.451	0.548	0.001
31	admit	border	0.296	0.703	0.001	0.467	0.532	0.001
58	notadmit	border	0.000	0.755	0.245	0.002	0.526	0.472
59	notadmit	border	0.000	0.867	0.133	0.002	0.655	0.343
66	border	admit	0.534	0.466	0.000	****	****	****
75	border	notadmit	****	****	****	0.002	0.483	0.515

<표 1.2> 모형 (2.1)에 의한 소속확률의 신뢰구간

관찰치	C. I. of p_3 (admit)	C. I. of p_2 (borderline)	C. I. of p_1 (not admit)
2	(0.2746 , 0.4005)	(0.5995 , 0.7281)	(0.0000 , 0.0042)
24	(0.3453 , 0.5232)	(0.4960 , 0.6547)	(0.0000 , 0.0026)
31	(0.4076 , 0.5303)	(0.4622 , 0.5900)	(0.0000 , 0.0021)
58	(0.0002 , 0.1597)	(0.4240 , 0.5663)	(0.4100 , 0.5388)
59	(0.0006 , 0.1327)	(0.5796 , 0.7160)	(0.2775 , 0.4000)
75	(0.0009 , 0.1557)	(0.3700 , 0.5634)	(0.4203 , 0.5997)

<표 2> 모형 (2.1)과 Logistic 모형에 의한 각 개체의 소속확률
(퍼지 자료 & (0-1) 자료)

관찰치	퍼지 자료		모형 (2.1)		Logistic모형		(0-1) 자료		모형 (2.1)		Logistic모형	
	p_{i1}	p_{i2}	p_{i1}	p_{i2}	p_{i1}	p_{i2}	p_{i1}	p_{i2}	p_{i1}	p_{i2}	p_{i1}	p_{i2}
1	0.87	0.06	0.90	0.04	0.88	0.03	1	0	1.00	0.00	1.00	0.00
2	0.88	0.05	0.89	0.07	0.88	0.05	1	0	1.00	0.00	1.00	0.00
3	0.93	0.03	0.90	0.05	0.85	0.05	1	0	1.00	0.00	1.00	0.00
4	0.86	0.06	0.86	0.05	0.80	0.04	1	0	1.00	0.00	1.00	0.00
5	0.87	0.06	0.86	0.04	0.80	0.07	1	0	1.00	0.00	1.00	0.00
6	0.42	0.35	0.38	0.28	0.59	0.38	1	0	0.70	0.30	1.00	0.00
7	0.08	0.82	0.04	0.83	0.08	0.78	0	1	0.00	1.00	0.00	1.00
8	0.06	0.87	0.07	0.83	0.06	0.86	0	1	0.00	1.00	0.00	1.00
9	0.06	0.86	0.06	0.84	0.05	0.78	0	1	0.00	1.00	0.00	1.00
10	0.06	0.87	0.05	0.85	0.04	0.90	0	1	0.00	1.00	0.00	1.00
11	0.06	0.86	0.08	0.86	0.03	0.83	0	1	0.00	1.00	0.00	1.00
12	0.07	0.84	0.08	0.91	0.02	0.90	0	1	0.00	1.00	0.00	1.00
13	0.36	0.27	0.35	0.23	0.36	0.26	1	0	0.51	0.26	1.00	0.00
14	0.12	0.08	0.16	0.08	0.17	0.08	0	0	0.14	0.00	0.00	0.00
15	0.08	0.07	0.06	0.05	0.13	0.10	0	0	0.10	0.00	0.00	0.00
16	0.10	0.10	0.09	0.13	0.10	0.16	0	0	0.00	0.12	0.00	0.00
17	0.08	0.06	0.11	0.04	0.12	0.06	0	0	0.00	0.00	0.00	0.00
18	0.04	0.04	0.05	0.05	0.10	0.08	0	0	0.00	0.00	0.00	0.00
19	0.07	0.07	0.06	0.09	0.07	0.10	0	0	0.00	0.00	0.00	0.00
20	0.10	0.08	0.09	0.04	0.09	0.04	0	0	0.00	0.00	0.00	0.00
21	0.07	0.06	0.05	0.05	0.07	0.05	0	0	0.00	0.00	0.00	0.00
22	0.09	0.09	0.05	0.08	0.05	0.07	0	0	0.00	0.00	0.00	0.00

참 고 문 헌

[1] Brown, P.J. (1982), "Multivariate calibration", *Journal of the Royal Statistical Society, B*, 44, 287-321.

[2] Brown, P.J., and Sundberg, R, (1987), "Confidence and conflict in multivariate calibration", *Journal of the Royal Statistical Society, B*, 49, 46-57.

[3] Critchley, F., and Ford, I. (1984), "Interval estimation in discrimination : the multivariate normal equal covariance case", *Biometrika* 72, 109-116.

[4] Dawid, A.P. (1976), "Properties of diagnostic data distributions", Biometrics 32, 647-658.

[5] Hirst, D.J., Ford, I., and Critchley, F. (1990), "An empirical investigation of methods for interval stimation of the log odds ratio in discriminant analysis", *Biometrika* 77, 609-615.

[6] Hoadley, B. (1970), "A Bayesian look at inverse regression", *Journal of the American Statistical Association*, 65, 356-369.

[7] Kosko, B. (1993), Fuzzy thinking, Hyperion, New York.

[8] Lee, H. S. (1995), "Markov Chain Monte Carlo methods", Ph. D. Theses, Department of Statistics, University of Missouri, Columbia, Mo.

[9] McLACHLAN, G.J. (1992), Discriminant Analysis and Statistical Pattern Recognition, John Wiley Sons, Inc., New York.

[10] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953), "Equations of state calculations by fast computing machines", *Journal of Chemical Physics* 21, 1087-1091.

[11] Johnson, R.A., Wichern, D.W. (1992), Applied Multivariate Statistical Analysis, 3rd ed., Prentice Hall, New Jersey.

[12] Ripley, B.D. (1994), "Neural networks and related methods for classification", (with discussion), *Journal of the Royal Statistical Society, B* 56, 409-437.

[13] Oman, S.D, (1988), "Confidence regions in multivariate calibration", *Annals of statistics*, 16, 174-187.

[14] Oman, S.D., and Wax, Y. (1984), "Estimating fetal age by ultrasound measurements : an example of multivariate calibration", *Biometrics*, 40, 947-960.

[15] Rousseeuw, P.J. (1995), "Fuzzy clustering at the intersection", *Technometrics*, Vol. 37, No 3, 283-286.

[16] Tanner, M.A. (1993), Tools for Statistical Inference, 2nd ed., Springer-Verlag, New York.

[17] 박래현, 이석훈 (1990), "A Bayesian analysis in multivariate bioassay and multivariate calibration", *Journal of the Korean Statistical Society*, 19, 71-79.

[18] 이원돈, 이석훈 (1989), "신경회로망 최적화 기법의 배경 및 응용", 「전기학회지」, 38권 2호, 23-30.

[19] 이석훈, 박래현, 최종석 (1990), "Bayesian control problem in multivariate mixture model", 「응용통계연구」, 3권 2호, 27-37.

Discriminant Analysis Based on a Calibration Model[4)]

Sukhoon Lee[5)], Nae-Hyun Park[5)] , Hye-Young Bok[6)]

Abstract

Most of the data sets to which the conventional discriminant rules have been applied contain only those which belong to one and only one class among the classes of interest. However the extension of the bivalence to multivalence like Fuzzy concepts strongly influences the traditional view that an object must belong to only one class.

Thus the goal of this paper is to develop new discriminant rules which can handle the data each object of which may belong to more than two classes with certain degrees of belongings.

A calibration model is used for the relationship between the feature vector of an object and the degree of belongings and a Bayesian inference is made with the Metropolis algorithm on the degree of belongings when a feature vector of an object whose membership is unknown is given.

An evalution criterion is suggested for the rules developed in this paper and comparision study is carried using two training data sets.

4) The Research was supported by the Korea Science & Engineering Foundation Grant, 1996. (961-0105-024-1)

5) Department of Statistics, College of Natural Sciences, Chungnam National University, Daejeon 305-764, Korea

6) Department of Statistics, College of Natural Sciences, Chungnam National University, Daejeon 305-764, Korea