The Korean Communications
in Statistics Vol. 12 No. 1, 2005
pp. 125-137

Discriminant Analysis of Binary Data with Multinomial Distribution by Using the Iterative Cross Entropy Minimization Estimation[1)]

Jung Jin Lee[2)]

Abstract

Many discriminant analysis models for binary data have been used in real applications, but none of the classification models dominates in all varying circumstances(Asparoukhov & Krzanowski(2001)). Lee and Hwang(2003) proposed a new classification model by using multinomial distribution with the maximum entropy estimation method. The model showed some promising results in case of small number of variables, but its performance was not satisfactory for large number of variables. This paper explores to use the iterative cross entropy minimization estimation method in replace of the maximum entropy estimation. Simulation experiments show that this method can compete with other well known existing classification models.

Keywords : Discriminant Analysis, Binary Data, Multinomial Distribution, Cross Entropy

1. 서론

판별분석은 1930년대 Fisher가 처음 제안한 이후로 많은 연구가 되어왔고, 현재도 새로운 기법들이 연구되고 있다(Lachenbruch(1981), Johnson and Wichern(1988), Duda et al(2001)). 그러나 대부분의 판별분석 기법들은 연속형 데이터를 위한 모형들이어서 범주형 데이터, 특히 이항데이터(boolean data)의 판별모형은 많이 연구되고 있지 않다. 하지만 최근 의학, 심리학 그리고 정보화 사회에서 발생되는 대용량 데이터에 이항데이터가 많아짐에 따라 연속형 데이터를 위한 판별기법이 아닌 이항데이터의 특수성을 이용한 판별분석에 관심이 높아지고 있다. 최근에 Asparoukhov & Krzanowski(2001)는 대표적으로 많이 이용되는 13가지의 판별분석모형을 의학실험에서 나타나는 이항데이터에 적용해 비교하는 논문을 발표하였다. 그 결과는 데이터 양의 많고 적음이나, 변수의 수가 많고 적음에 따라 각 모형의 장단점이 있어 어느 한가지 방법이 모든 데이터에 대해 우위에 있지는 못한 것으로 나타났다.

1) This research is supported in part by the 2004 Soong Sil University Research Fund.
2) Professor, Department of Statistics, Soong Sil University, Seoul, 156-743, Korea.
E-mail: jjlee@ssu.ac.kr

새로운 대안의 모색으로 이정진·황준(2003)은 최대 엔트로피 추정법(maximum entropy estimation method)을 이용한 다항분포(multinomial distribution) 판별모형을 제안하였다. 이 모형은 실험결과 변수의 수가 적을 때는 의미있는 분류결과를 보여주었으나, 변수의 수가 많은 경우에는 많은 변수의 비선형방정식을 풀어야 하는 현실적인 문제가 있어 활용되기가 쉽지 않다. 본 연구에서는 이 문제를 해결하기 위해 상대엔트로피 최소화 추정법(cross entropy minimization method)을 이용한 다항분포 판별모형을 제안한다. 이 모형은 표본 데이터의 정보를 저차원 프로젝션(projection), 즉 주변확률분포(marginal distribution)로 축약한 후, 각각의 주변확률에 대한 정보를 반복적으로 상대엔트로피를 최소화 하면서 다항분포 모수를 추정한다.

2절에서는 이항데이터에 대한 판별분석에 사용되는 모형에 대하여 살펴본다. 3절에서는 반복적 상대엔트로피 최소화 추정법으로 다항분포 모수를 추정하는 판별모형을 제안하고, 4절에서는 이 모형의 진단을 위하여 시뮬레이션으로 다른 모형과 비교한 결과를 살펴본다. 5절에서는 결론과 향후과제에 대한 제안을 한다.

2. 이항데이터 판별분석 모형

편의상 두 그룹 w_1과 w_2에 대한 판별분석을 살펴보자. p개의 확률변수를 $\vec{X} = (X_1, X_2, \cdots, X_p)$, 이 확률변수가 나타내는 값을 $\vec{x} = (x_1, x_2, \cdots, x_p)$, 각 그룹의 확률분포함수를 $f_i(\vec{x})$, 사전확률을 π_i, 그리고 a_{ij}를 실제 그룹이 w_j일 때 w_i로 결정하였을 때의 손실이라 하자. 일반적으로 많이 이용되는 베이지안(Bayesian)방법에 의한 판별식은 다음과 같다.

$$\text{만일 } \frac{f_1(\vec{x})}{f_2(\vec{x})} > \frac{a_{12}}{a_{21}}\frac{\pi_2}{\pi_1} \text{ 이면 } \vec{x} \text{ 를 } w_1\text{으로 분류, 아니면 } w_2\text{로 분류} \qquad (1)$$

위의 판별식은 훈련데이터(training data)로부터 확률분포함수 $f_1(\vec{x})$와 $f_2(\vec{x})$를 추정한 후 현실데이터에 적용하게 되는데, 확률분포함수를 어떻게 가정하느냐에 따라 판별식은 달라질 수 있다. 많이 이용되는 확률분포함수는 다변량정규분포(multivariate normal distribution)로서, 만일 $f_i(\vec{x})$가 평균이 다르고 공분산행렬이 같은 다변량정규분포라면 위의 판별식은 선형판별식이 되고, 공분산 행렬이 다르다면 이차판별식이 된다. 이때 다변량정규분포의 추정을 모수적 방법인 최우추정법(maximum likelihood estimation: MLE)이나 베이지안 추정법(Bayesian parameter estimation)으로 할 수도 있고, 비모수적 방법인 k-근접이웃(k-nearest neighbor) 방법과 커넬(kernel) 방법을 사용할 수도 있다. 이밖에 $\vec{x}$와 그룹변수와의 관계를 로지스틱(logistic)함수 형태로 가정하는 판별모형도 있고, 로그선형(log linear) 모형, 신경망(neural network) 모형, 회귀나무(regression tree) 모형, 선형계획(linear programming)이나 혼합정수계획(mixed integer programming)모형 등이 있다(Duda et al(2001) 참조).

위에서 설명한 대부분의 판별분석 기법들은 몇 가지를 제외하고는 연속형 데이터를 위한 모형들이고 범주형 데이터, 특히 이항데이터의 특수성을 고려한 방법은 아니다. 최근에 Asparoukhov & Krzanowski(2001)는 많이 이용되는 13가지의 판별분석모형을 다섯 종류의 의학 및 심리학 실

험에서 나타나는 이항데이터에 적용해 비교하는 논문을 발표하였다. 그 결과는 변수의 개수와 데이터의 수에 따라 각 모형의 장단점이 있어 어느 한가지 방법이 모든 이항데이터의 판별에 대해 우위에 있지는 못한 것으로 나타났다. 하지만 이항데이터의 판별분석에는 대개 다섯 가지 모형, 즉, 선형판별(linear discriminant)모형, 이차판별(quadratic discriminant)모형, k-근접이웃(k-nearest neighbor)모형, 로지스틱(logistic)모형, 신경망(neural network)모형 등이 어느 정도 상대적으로 효과가 있는 것으로 나타났다. 이 모형들은 3절에서 제안하는 반복적 상대엔트로피 최소화 추정법에 의한 다항분포 판별모형과 비교된다.

3. 반복적 상대엔트로피의 최소화 추정법을 이용한 다항분포의 추정

이항데이터의 다양한 형태를 잘 나타낼 수 있는 원시적인 모형은 이항변수의 모든 값이 나타날 확률을 고려하는 다항분포 모형이다. [식 1]에서 $f_1(\vec{x})$와 $f_2(\vec{x})$를 서로 다른 모수를 갖는 다항분포라 가정하면 이 판별식은 분류오류가 최소화되는 모형이 될 수 있다. 하지만 변수의 수가 증가하면 다항분포의 추정되어야 하는 모수가 지수적으로 증가하게 되어 추정에 필요한 데이터의 수가 충분치 않아 최우추정법 등 많이 이용되는 추정방법으로는 모든 모수의 추정이 어렵고 수치해석적 문제가 발생하게 된다. 예를 들어 변수가 10개일 때 추정하여야 할 다항분포 모수의 수는 $2^{10}-1 = 1023$ 개가 된다. 하지만 우리가 현실적으로 얻을 수 있는 훈련 데이터의 수는 잘 해야 몇 백 개정도 이어서 이를 이용해서 최우추정법 등으로 만족할만한 다항분포의 모수를 추정하는 것은 쉽지 않다.

Kantor · Lee(1998) 와 이정진 · 황준(2003)은 이 모형의 위와 같은 문제점을 해결하기 위해 표본 이항데이터의 주변확률분포와 최대 엔트로피 추정법을 이용한 다항분포 모수의 추정을 제안하였다. p개의 이항변수 $\vec{X} = (X_1, X_2, \cdots, X_p)$에 대한 표본공간은 모든 불린(boolean) 원소들의 집합 $S = \{\vec{x}:(0,0,\cdots,0),(0,0,\cdots,1),\cdots,(1,1,\cdots,1)\}$이고, 이 표본공간에 정의된 다항분포를 $f(\vec{x})$라 하자. 그리고 전체 이항변수 $X_1, X_2, \cdots, X_p$ 중에서 판별능력이 큰 변수들의 부분집합이 N개 있고 이를 $S_1, S_2, \cdots, S_N$ 으로 표시하자. 변수집합 S_j 에 대한 주변확률분포를 $f(\vec{x}_{S_j})$, 표본 데이터에서 추정한 이 주변확률분포를 $\hat{f}(\vec{x}_{S_j})$라 할 때, 최대 엔트로피 추정법을 이용한 다항분포 $f(\vec{x})$의 추정은 다음의 비선형계획 문제의 해를 구함으로서 추정할 수 있다.

$$\begin{aligned} &\text{Maximize} \quad -\sum_{x\in S} f(\vec{x}) \ln f(\vec{x}) \\ &\text{Subject to} \\ &\qquad \sum_{x\in S} f(\vec{x}) = 1 \\ &\qquad f(\vec{x}_{S_j}) = \hat{f}(\vec{x}_{S_j}), \quad j=1,2,\cdots,N \end{aligned} \tag{2}$$

판별능력이 큰 변수들의 집합 S_j의 결정은 두 그룹의 표본 주변확률분포에 대한 카이제곱 동질

성검정, 가우시안(Gaussian) 측도 등을 이용하거나[이정진 · 김수관(2002)], 모든 가능한 2차원 또는 3차원 변수를 다 이용하는 방법도 있을 수 있다. 만일 집합 S_j, $j=1,\cdots,p$, 가 j번째 이항변수 X_j 하나만을 포함하고 있고(즉, $S_j=\{X_j\}$) 표본에서 추정한 이 변수의 주변확률분포가 $\hat{f}(x_j)$라면, [식 2]에 의해 추정된 다항분포 $f(\vec{x})$는 각각의 주변확률분포의 곱으로 표시된다.

$$f(\vec{x}) = \hat{f}(x_1)\hat{f}(x_2)\cdots\hat{f}(x_p) \quad , \vec{x} \in S \tag{3}$$

하지만 집합 S_j 가 여러 개의 변수를 포함하고 있다면 추정하여야 할 다항분포 모수의 수는 2^p-1 개가 되어 변수의 수가 증가하면 해를 구하기가 쉽지 않다. 비선형계획 문제인 [식 2]는 현재 많이 이용되는 GENO 등 소프트웨어를 사용하였을 때 대략 7개의 이항변수, 즉 $2^7 = 128$ 개의 미지 모수가 있을 경우 해를 구할 수 있는 정도인데, 이마저도 초기치 문제 또는 지역해 문제가 있어 만족할 만한 해를 구하기 어렵다. 이정진 · 황준(2003)은 적절한 변환을 이용하여 비선형계획 문제대신 N 개의 비선형 연립방정식의 해를 구하여 각각의 모수를 추정하는 방법을 고안하였는데 이 방법도 역시 변수의 수와 집합 S_j 의 수 N 이 증가하게 되면 (현실적으로 10개 이상) 해를 구하는데 수치해석적 문제가 발생한다.

본 논문에서는 이렇게 변수가 많을 경우 [식 2]의 변형으로서 다음과 같은 반복적 상대엔트로피(Iterative Cross Entropy: ICE) 최소화 추정법을 통해 다항분포 모수 추정에 대한 수치해석적 문제를 해결하고자 한다. 이 추정법은 다항분포의 초기 추정을 균등분포(uniform distribution) (또는 [식 3]과 같은 독립적인 이항분포의 곱)로 가정한 후, 이 균등분포와 상대엔트로피가 가장 가까우면서 표본에서 추정한 주변확률분포를 만족하는 분포를 반복적으로 찾아 나가는 알고리즘이다.

(반복적 상대엔트로피 최소화 추정법)

단계 0: 모든 불린원소의 확률이 균등분포로 가정하자. 즉,

$$f^{(0)}(\vec{x}) = \frac{1}{2^p}, \quad \vec{x} \in S \tag{4}$$

단계 k: $k=mN+j$, $j=1,2,\cdots,N$, $m=0,1,2,\cdots$ 에 대하여 다음의 비선형계획 문제의 해 $f^{(k)}(\vec{x})$를 구한다.

$$\begin{aligned} &\text{minimize} \quad \sum_{x\in S} f^{(k)}(\vec{x}) \ln \frac{f^{(k)}(\vec{x})}{f^{(k-1)}(\vec{x})} \\ &\text{subject to} \\ &\qquad \sum_{x\in S} f^{(k)}(\vec{x}) = 1 \\ &\qquad f^{(k)}(\vec{x}_{S_j}) = \hat{f}(\vec{x}_{S_j}) \end{aligned} \tag{5}$$

이와 같은 방법을 $f^{(k)}(\vec{x})$ 가 미리 정한 수렴조건, 예를 들면 $\| f^{(k-1)}(\vec{x}) - f^{(k)}(\vec{x}) \| < \varepsilon$, 이 만족될 때까지 반복한다.

이 ICE 최소화 추정법은 교차표(contingency table)에서 주어진 주변확률을 만족시키는 분포를 추정하는데 많이 이용되어 왔지만(Ireland and Kullback(1968)) 범주형 자료에 대한 판별분석에 사용된 사례는 저자가 조사한 바로는 아직 없다. 이 추정법의 가장 큰 장점은 [식 5]의 비선형계획문제의 해를 쉽게 구할 수 있고, k 가 충분히 클 경우 주변확률을 만족하는 분포에 수렴함을 보일 수 있다는 것이다(Ireland and Kullback(1968), Cramer(2000)). 예를 들어 모든 가능한 두 개의 확률변수 X_i와 X_j의 주변확률분포를 이용하여 ICE 최소화 추정법으로 다항분포를 추정하여 보자. 집합 $S_{ij=(t,u)}$ 를 $X_i=t,\ X_j=u,$ (여기서 $t=0,1;\ u=0,1$)를 갖는 불린원소들의 집합이라 하고, 표본에서 추정한 2차원 주변확률 $P(X_i=t, X_j=u)$의 값이 $c_{\ldots t\ldots u\ldots}$ 라 하자. 라그란지안 승수법(Lagrangian multiplier method)을 이용한 [식 5]의 해는 다음과 같음을 보일 수 있다.

각각의 $\vec{x} \in S,\ t=0,1;\ u=0,1$ 에 대해

$$f^{(k)}(\vec{x}) = c_{\ldots t\ldots u\ldots} \frac{f^{(k-1)}(\vec{x})}{\sum_{\vec{x}\in S_{i,j=(t,u)}} f^{(k-1)}(\vec{x})}, \quad \text{if} \ \sum_{\vec{x}\in S_{i,j=(t,u)}} f^{(k-1)}(\vec{x}) \neq 0 \qquad (6)$$

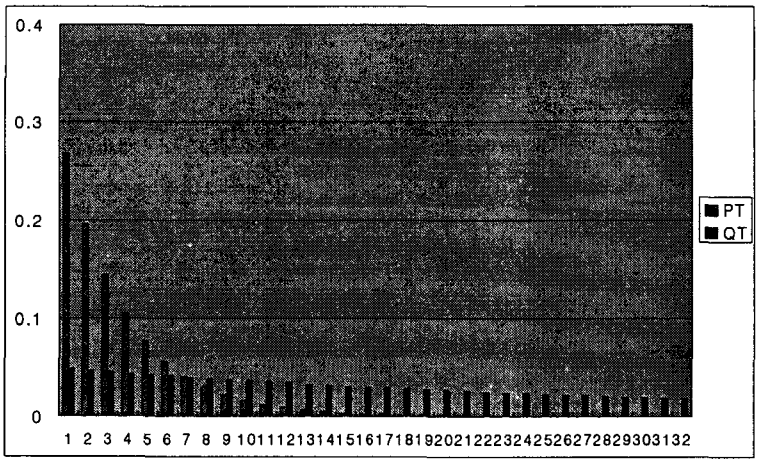
$$= c_{\ldots t\ldots u\ldots} \frac{4}{2^p}, \qquad \text{else}$$

같은 주변확률을 같는 다항분포는 한 가지 이상의 해가 있을 수 있다. 즉, [식 2]에 의한 다항분포 추정결과와 [식 5]에 의한 ICE 최소화 추정법에 의한 다항분포 추정결과는 목적함수가 서로 다르기 때문에 그 해도 서로 다를 수 있을 것이다. 두 가지 추정법의 해에 대한 비교는 수학적으로 쉽지 않다. 하지만 ICE 최소화 추정법은 반드시 해가 존재하며, 위의 알고리즘은 그 해에 수렴하고, 비선형계획문제 [식 5]는 쉽게 해를 구할 수 있는 장점이 있어 변수의 수가 많은 경우에도 이용될 수 있다.

3.1 ICE 최소화 추정법의 대표본 특성

ICE 최소화 추정법을 이용한 다항분포의 추정에 대한 이론적인 연구와, 특히 이 추정방법에 대한 대표본 특성에 대한 연구는 저자가 조사한 바로는 아직까지 없다. 수학적으로 이 추정법을 이용하여 표본의 크기가 커짐에 따라 얼마나 다항분포 모수를 잘 추정하고 있는지 살펴보기 위한 이론의 전개는 쉽지 않아 본 연구에서는 다음과 같이 간단한 시뮬레이션 실험을 하여 보았다.

<그림 1>은 현실의 이항데이터 판별문제에 많이 나타날 수 있는 두 가지 서로 다른 다항분포 모집단 형태를 변수가 5개인 경우 보여주고 있다.



<그림 1> 두 가지 다항분포 모집단의 형태 (모집단 1: PT, 모집단 2: QT)

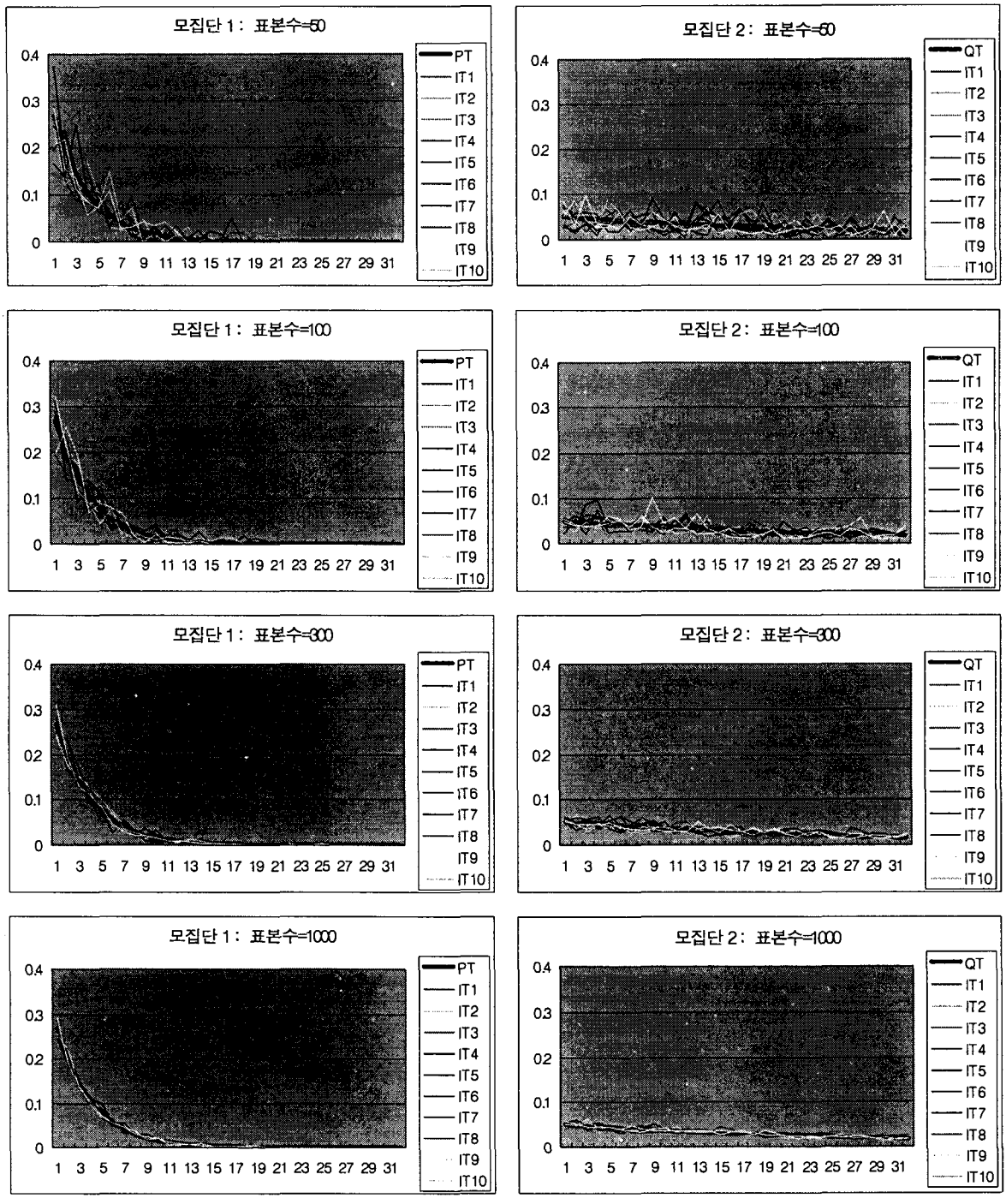
이 그림에서 X축의 숫자 1, 2, … , 32는 편의상 다음과 같은 불린원소들의 번호를 의미한다.

$$\begin{aligned} &\text{불린원소} \quad (0,0,0,0,0) \Rightarrow 1 \\ &\text{불린원소} \quad (0,0,0,0,1) \Rightarrow 2 \\ &\qquad\qquad \cdots \\ &\text{불린원소} \quad (1,1,1,1,1) \Rightarrow 32 \end{aligned} \tag{7}$$

한 다항분포(모집단 1: PT)는 각 불린원소의 확률이 한쪽 편으로 큰 형태이고, 다른 하나(모집단 2: QT)는 확률이 비슷한 경우인데 불린원소의 순서는 서로 다를 수 있으므로 이 두 가지 모집단 형태는 다양한 다항분포 형태를 나타낸다고 볼 수 있다. 이 두 가지 형태의 다항분포 모집단에서 표본을 50개, 100개, 300개, 1000개(즉 모든 불린원소의 수의 1.6배, 3.1배, 9.4배, 31.3배)를 추출한 후 모든 가능한 2차원 주변확률분포를 구하고 [식 6]을 이용한 ICE 최소화 추정법으로 다항분포를 추정하였다. 같은 실험을 10회 반복 추출하여 각각의 다항분포를 구한후 선그림으로 표시한 결과(IT_i)가 <그림 2>와 같다.

표본의 수가 50개 또는 100개인 경우 각 불린원소의 기대 반복수는 상당히 작다. 이 경우 추정된 분포는 모집단과 유사하나 추정이 불안정함을 알수 있다. 하지만 표본의 수가 300개 또는 1000개인 경우 점차로 두 모집단의 분포에 수렴함을 알 수 있다. 즉 각각의 불린원소에 대한 기대 반복수가 대략 10배 이상인 경우 ICE 최소화 추정법으로 다항분포를 어느 정도 정확히 추정할 수 있다고 볼 수 있다. 이러한 시뮬레이션 실험결과는 표본의 모든 2차원 주변확률분포를 이용하여 ICE 최소화 추정법으로 다항분포를 추정해도 어느 정도 만족할만한 결과를 얻을 수 있다는 것을 보여주고 있다.

변수의 수가 5개 이상인 경우도 비슷한 결과를 보여 주고 있으나 지면관계상 생략하기로 한다. 4절에서는 이와 같은 ICE 최소화 추정법을 이용한 판별모형과 기존에 많이 사용되고 있는 판별모형에 대한 비교 실험을 하고자 한다.



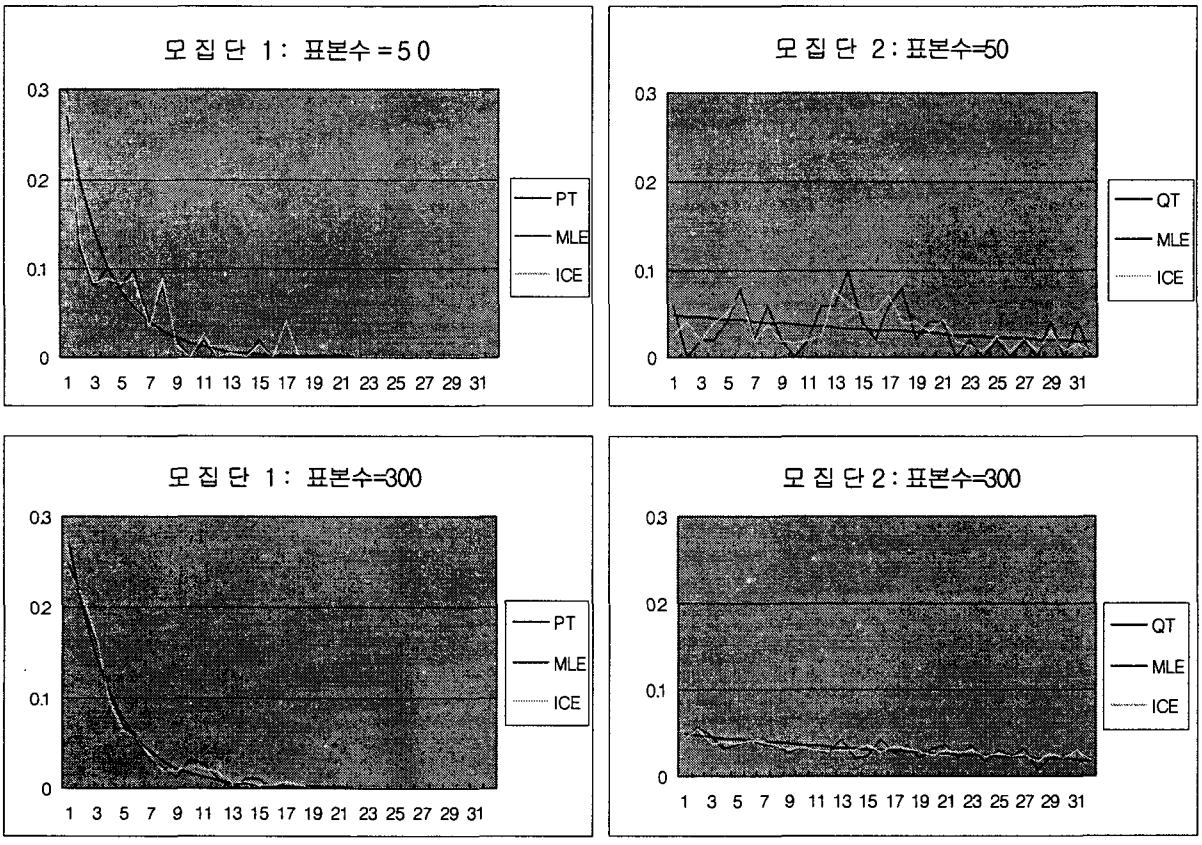
- 모집단 1 (PT)의 추정 -

- 모집단 2 (QT)의 추정 -

<그림 2> 표본수의 변화에 따른 ICE 최소화 추정법의 다항분포 추정(IT_i는 반복을 뜻함)

3.2 최우추정법과 ICE 최소화 추정법의 비교

다항분포의 모수를 최우추정법(maximum likelihood estimation: MLE)을 이용하여 추정하면 각 불린원소의 확률이 표본비율이 된다. <그림 1>의 두 가지 다항분포 모집단에 대해 최우추정법과 ICE 최소화 추정법을 이용하여 시뮬레이션 실험을 한 결과가 <그림 3>과 같다. 불린원소의 수(32)보다 표본의 수가 충분히 크면(300인 경우) 두 가지 방법이 비슷한 추정을 하게된다. 하지만 불린원소의 수보다 표본의 수가 충분치 않으면(50인 경우), 특히 각 원소의 확률이 비슷한 모집단 2의 경우(<그림 3>의 우측 상단), 불린원소에 대한 표본이 없는 경우가 많은데, 이때 최우추정법은 표본도수가 없는 불린원소를 0 으로 추정하는 반면, 상대엔트로피 최소화 추정법은 주변확률을 고려하여 0 이 아닌 확률을 추정하게 되어 상대적으로 더 모집단에 근사한 추정을 하게된다. 즉 표본 수가 충분치 않은 경우 표본이 가지고 있는 정보를 최대화하는 상대엔트로피 최소화 추정법의 장점을 잘 보여 주고 있다.



- 모집단 1 (PT)의 추정 -　　- 모집단 2 (QT)의 추정 -

<그림 3> 최우추정법(MLE)과 ICE 최소화 추정법에 의한 다항분포 추정 비교

4. ICE 최소화 추정법을 이용한 판별모형의 비교실험

두 그룹의 표본데이터를 이용하여 3절에서 설명한 ICE 최소화 추정법으로 다항분포 $\hat{f}_1(\vec{x})$와 $\hat{f}_2(\vec{x})$ 를 추정하였다고 하자. 만일 판별이 필요한 새로운 데이터 $\vec{x}_o$ 가 있으면 [식 1]은 두 그룹의 오분류 손실이 같고, 사전확률 π_1과 π_2가 같다고 가정할 때 판별식은 다음과 같다.

$$\hat{f}_1(\vec{x}_0) \geq \hat{f}_2(\vec{x}_0) \text{ 이면 } \vec{x}_o \text{ 를 그룹 1로 분류, 아니면 그룹 2로 분류} \qquad (8)$$

즉, 추정된 다항분포를 이용하여 $\vec{x}_o$ 가 나타날 확률이 높은 그룹으로 판별한다. 이 절에서는 다항분포를 직접 추정하여 판별하는 ICE 최소화 추정모형, 최우추정(MLE) 모형과, 이항데이터의 판별에 어느 정도 효과적이라고 알려져 있는 판별모형 즉, 선형판별식(linear discriminant function: LDF)모형, 이차판별식(quadratic discriminant function: QDF)모형, k-근접이웃(k-nearest neighbor, k-NN)모형, 로지스틱 회귀(logistic regression: Lreg)모형에 대해 시뮬레이션 실험으로 비교하여 보았다. 시뮬레이션 실험을 위한 데이터의 생성과 구체적인 실험방법은 다음과 같다.

데이터

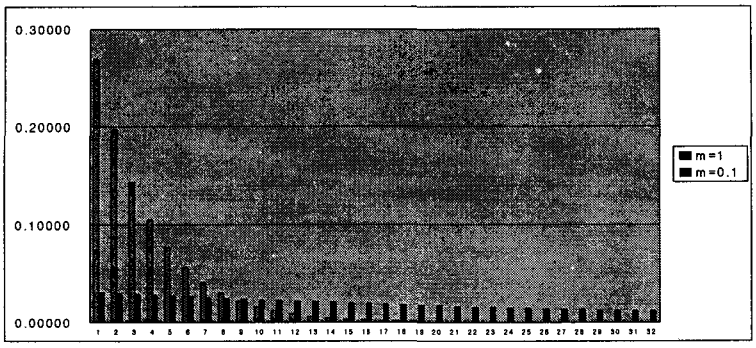
두 모집단의 판별을 위해 현실에서 많이 나타나는 데이터의 형태는 여러 가지가 있을 수 있으나 크게 <그림 4>와 같이 두 가지 분포모형으로 구분할 수 있다. 여기서 X축은 (7)식과 같은 불린원소의 표시방법이다. 첫 번째 분포모형은 두 모집단의 분포가 유사하고, 확률이 큰 불린원소들이 서로 상당히 겹치는 경우이다. 두 번째 분포모형은 두 모집단의 분포가 아주 다르고, 불린원소들이 겹치는 경우가 확률이 작은 부분에서 나타나는 경우이다.

이와 같은 다항분포 데이터는 근사적으로 지수분포를 2^p개의 등구간으로 나누어 각 구간의 확률을 구한 다음 역변환법(inverse transformation method)에 의해 원하는 표본수 만큼 생성하였다. 첫 번째 분포모형은 모집단 1(막대가 회색)을 지수분포(m=1.0), 모집단 2(막대가 검정색)는 지수분포(m=0.1)를 이용하였고, 두 번째 분포모형은 모집단 1을 지수분포(m=0.3), 모집단 2는 지수분포(m=0.3)의 대칭인 분포를 이용하여 표본을 생성하였다. 지수분포인 경우 무한대까지의 값을 가질 수 있으므로 최대값은 99.99퍼센타일(percentile)을 이용한 후 확률분포가 되도록 조정하였다.

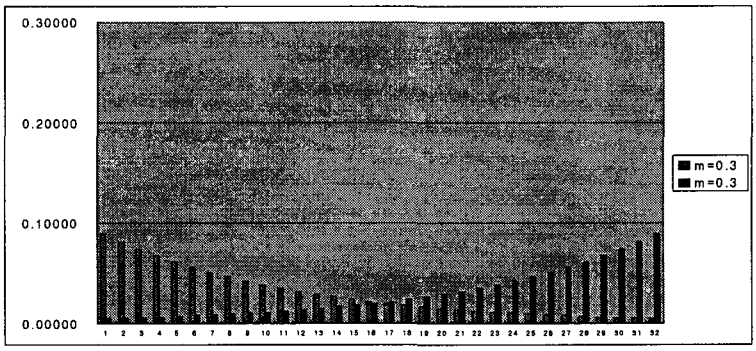
ICE 최소화 추정법을 이용하는 다항분포 모수의 추정에는 변수의 수에 대한 제약은 크게 없다. 하지만 다른 판별모형은 변수의 수에 따라 실험시간 등 여러 가지 문제가 발생되어 본 실험에서는 현실문제에서 많이 나타나는 이항변수의 수 5개와 7개에 대하여만 우선적으로 실험하였다.

실험방법

위와 같이 생성되는 이항데이터를 이용한 구체적인 시뮬레이션 실험방법은 다음과 같다.



- 분포모형 1: 두 모집단의 확률이 큰 불린원소들이 많이 겹치는 경우 (p=5) -



- 분포모형 2: 불린원소들이 겹치는 경우가 확률이 작은 부분에서 나타나는 경우 (p=5) -

<그림 4> 시뮬레이션을 위해 가정된 두 종류의 모집단 분포모형

1) 가정된 세 종류의 분포모형에 대해서 각 모집단에서 표본을 추출한다. 앞 절에서 살펴보았듯이 ICE 추정법은 표본수가 300 이상이면 원래의 다항분포에 근접함으로 실험을 위한 훈련(training)데이터는 각 모집단에서 50, 100, 300개씩 각각 추출한다. 그리고 같은 수 만큼의 시험(testing)데이터를 추출한다.
2) 추출된 훈련데이터의 모든 가능한 2차원 주변확률을 구한다. 현실적으로는 이러한 주변확률 중에서 그룹 1과 그룹 2가 통계적으로 서로 다른 분포인지 카이제곱 동질성검정을 하여 판별에 의미 있는 저차원 분포를 선택할 수도 있다.
3) 훈련데이터의 모든 2차원 주변확률을 이용하여 ICE 최소화 추정법으로 다항분포를 추정한다.
4) 훈련데이터에서 추정된 분포를 이용하여 시험데이터에 [식 8]의 판별식으로 판별을 실시하여 정분류율을 계산한다.
5) 같은 훈련데이터에 대해 나머지 다섯 가지 판별방법을 적용하여 모형을 만든 후 시험데이터를 이용하여 정분류율을 계산한다.
6) 위의 1)에서 5)까지의 실험을 10회씩 반복하여 정분류율의 평균과 표준편차를 계산한다. 실험의 회수를 더 많이 할 수도 있으나 예비실험결과 실험의 회수를 증가하면 표준오차는 어느 정도 감소하나 평균정분류율은 거의 변하지 않아 전체 실험시간을 고려하여 10회로 정하였다.

실험을 위해서 ICE와 MLE는 C언어를 이용하여 직접 프로그램을 작성하였고 나머지 방법은 R 통계패키지를 이용하였다.

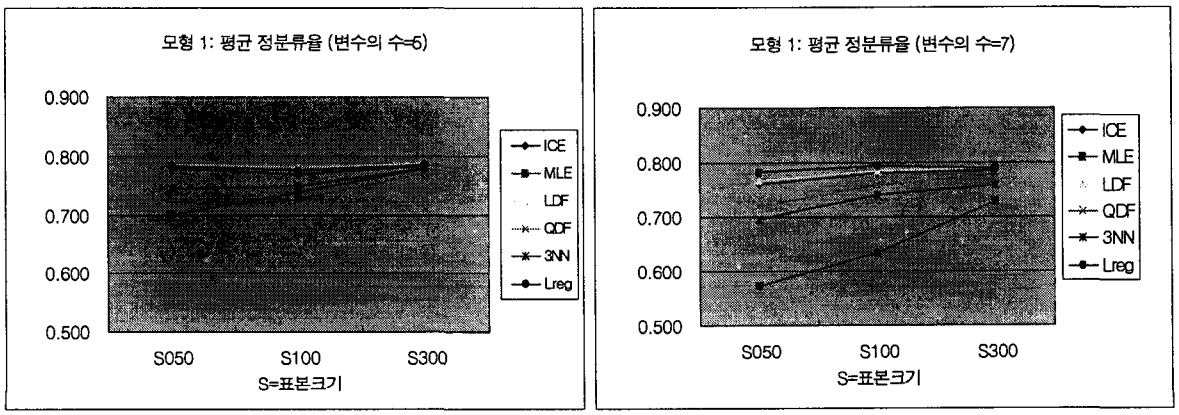
실험결과

시뮬레이션 실험결과가 <표 1>과 <표 2>에 정리되어 있다. 전체적으로 표본의 크기가 충분히 크면 MLE를 제외한 여러 가지 모형들이 비슷한 판별력을 보이고 있다. 이 사실은 실험에서 비교한 판별모형들이 이미 Asparoukhov 등 (2001)에 의해 이항데이터의 판별에 어느 정도 경쟁력 있는 모형으로 제안되었던 것을 뒷받침한다. 하지만 표본의 크기가 불린원소의 수보다 작은 경우에는 모집단 형태에 따라 서로 다른 판별력을 보여준다.

<표 1>은 두 모집단의 확률이 큰 불린원소들이 서로 많이 겹치는 경우의 실험결과로서, ICE, LDF, Lreg 모형이 표본수가 적은 경우 상대적으로 나머지 모형보다 우수한 판별력을 보인다. 세 모형은 표준오차 범위 내에서 서로 비슷한 판별력을 보이나 Lreg이 변수의 수가 7인 경우 표본의 수가 적은 경우(50 또는 100)일 때 오차범위 내이지만 상대적으로 좋은 평균정분류율을 보인다. 하지만 표본수가 증가하면 모든 분석모형의 표준오차가 줄어들고 여러 가지 판별모형이 비슷한 정분류율을 보인다.

<표 1> 분포모형 1에 대한 여섯 가지 판별방법의 평균 정분류율(표준오차)의 비교
(ICE:반복상대엔트로피 MLE:최우추정 LDF:선형 QDF:이차 3NN:근접이웃 Lreg:로지스틱)

총변수	표본수	ICE	MLE	LDF	QDF	3NN	Lreg
5	50	0.779(0.034)	0.700(0.054)	0.785(0.032)	0.708(0.045)	0.744(0.033)	0.783(0.042)
	100	0.768(0.026)	0.738(0.035)	0.768(0.028)	0.724(0.054)	0.749(0.036)	0.774(0.023)
	300	0.783(0.020)	0.779(0.022)	0.786(0.014)	0.770(0.028)	0.780(0.023)	0.786(0.019)
7	50	0.759(0.043)	0.574(0.037)	0.767(0.049)	0.720(0.047)	0.695(0.046)	0.782(0.036)
	100	0.779(0.028)	0.633(0.032)	0.785(0.029)	0.758(0.054)	0.740(0.041)	0.794(0.030)
	300	0.785(0.016)	0.729(0.014)	0.789(0.014)	0.778(0.032)	0.757(0.016)	0.789(0.013)

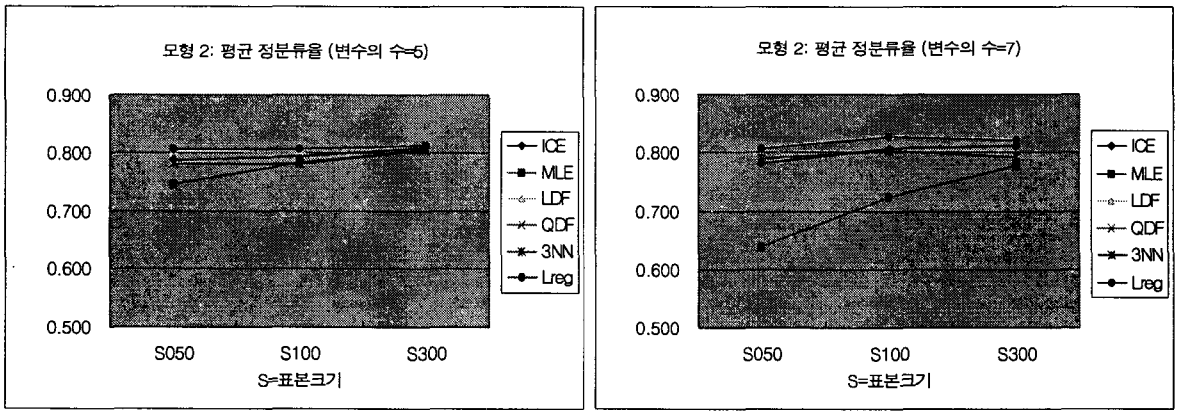


<그림 5> 분포모형 1에 대한 여섯 가지 판별방법의 평균 정분류율 비교

<표 2>는 두 모집단의 불린원소 들이 겹치는 곳이 확률이 작은 부분에서 나타나는 경우의 판별결과이다. 두 모집단의 형태가 완전히 구별되는 이 경우에는 두 모집단이 서로 다른 공분산행렬을 갖게된다. 이 경우에는 ICE, LDF, Lreg 와 더불어 QDF, 3NN도 표준오차의 범위 내에서 어느 정도 경쟁력 있는 판별력을 보이고 있다. 하지만 표본의 수가 50인 경우 MLE는 상대적으로 낮은 판별력을 보이고 있는데 이는 훈련데이터의 불린원소의 표본비율이 0 인 경우가 있어 시험데이터를 이용한 판별시 정분류율이 낮게 나타난 것이다. 여러 모형 중에서는 표준오차의 범위 내이지만 LDF와 Lreg이 상대적으로 좋은 평균정분류율을 보인다. 하지만 표본수가 증가하면 모든 분석모형의 표준오차가 줄어들고 여러 가지 판별모형이 비슷한 정분류율을 보인다.

<표 2> 분포모형 2에 대한 여섯 가지 판별방법의 평균 정분류율(표준오차)의 비교
(ICE:반복상대엔트로피 MLE:최우추정 LDF:선형 QDF:이차 3NN:근접이웃 Lreg:로지스틱)

총변수	표본수	ICE	MLE	LDF	QDF	3NN	Lreg
5	50	0.787(0.028)	0.746(0.041)	0.801(0.043)	0.798(0.038)	0.781(0.050)	0.806(0.041)
	100	0.793(0.016)	0.782(0.026)	0.805(0.021)	0.800(0.020)	0.780(0.026)	0.805(0.024)
	300	0.807(0.016)	0.805(0.018)	0.813(0.016)	0.814(0.016)	0.804(0.019)	0.812(0.015)
7	50	0.781(0.045)	0.638(0.047)	0.809(0.054)	0.796(0.048)	0.790(0.040)	0.806(0.040)
	100	0.807(0.029)	0.724(0.042)	0.830(0.031)	0.828(0.032)	0.803(0.029)	0.828(0.027)
	300	0.811(0.015)	0.778(0.024)	0.823(0.012)	0.820(0.014)	0.792(0.018)	0.823(0.013)



<그림 6> 분포모형 2에 대한 여섯 가지 판별방법의 평균 정분류율 비교

5. 결론 및 향후과제

본 연구에서는 표본 이항데이터의 주변확률을 이용하여 ICE 최소화 추정법으로 다항분포를 추정하는 방법을 연구하고, 이를 이용하여 데이터를 판별하는 모형을 제안한 후 시뮬레이션 실험을

이용하여 다른 판별모형과 비교 실험을 실시하였다. 그 결과 ICE 모형이 이항데이터의 판별에 우수하다고 알려져 있는 여러 가지 다른 판별모형과 비교하여 어느 정도 경쟁력 있는 판별력을 보임을 알 수 있었다. 하지만 비교실험의 제약상 제한된 변수의 수와 표본에 대해서만 실험을 하였기 때문에 위와 같은 결론을 일반화하기 위해서는 좀 더 다양한 환경에서의 실험과 이론적인 연구가 뒷받침되어야 한다고 할 수 있다. 무엇보다도 ICE 최소화 추정법을 이용한 다항분포 판별모형의 평가는 다양한 상황에서의 실제 현실데이터를 이용한 후 이루어질 수 있을 것이다.

참고문헌

[1] 이정진, 황준 (2003). Discriminant Analysis of Binary Data by Using the Maximum Entropy Distribution, 「한국통계학회논문집」, 제10권, 제3호, 909-917.

[2] 이정진, 김수관 (2002). Classification Analysis in Information Retrieval by Using Gauss Patterns, 「한국통계학회논문집」, 제9권, 제1호, 1-11.

[3] Asparoukhov, O.K. and Krzanowski, W.J. (2001). A comparison of discriminant procedures for binary variables. *Computational Statistics and Data Analysis*, 38, 139-160.

[4] Cramer, E. (2000). Probability Measures with Given Marginals and Conditionals: I-Projections and Conditional Iterative Proportional Fitting, *Statistics & Decisions,* Vol 18, 311-329.

[5] Duda, R.O., Hart, P.E., and Stork, D.G. (2001). *Pattern Classification*, Wiley.

[6] Johnson, R. and Wichern, D. (1988). *Applied Multivariate Statistical Analysis*, Prentice Hall.

[7] Ireland, C.T. and Kullback, S. (1968). Contingency tables with given marginals, *Biometrika,* Vol 55, 1, 179-188.

[8] Lachenbruch (1981). *Discriminant Analysis*, Prentice Hall.

[9] Kantor, P.B. and Lee, J.J. (1998). Testing the Maximum Entropy Principle for Information Retrieval. *Journal of American Society for Information Science*, Vol 49, 6, 557-566.

[2004년 10월 접수, 2005년 2월 채택]