The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 909-917

Discriminant Analysis of Binary Data by Using the Maximum Entropy Distribution[1)]

Jung Jin Lee[2)], Joon Hwang[3)]

Abstract

Although many classification models have been used to classify binary data, none of the classification models dominates all varying circumstances depending on the number of variables and the size of data(Asparoukhov and Krzanowski (2001)). This paper proposes a classification model which uses information on marginal distributions of sub-variables and its maximum entropy distribution. Classification experiments by using simulation are discussed.

Keywords : Classification Analysis, Binary Data, Maximum Entropy Distribution

1. 서론

판별분석은 1930년대 Fisher가 처음 제안안 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 이항데이터의 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.

본 연구에서는 이항데이터의 특수성을 최대한 이용할 수 있는 판별모형을 제안한다. 이항데이터를 잘 설명할 수 있는 원시적인 모형은 이항변수의 모든 가능한 값에 대한 경우를 포함하는 다항분포(multinomial distribution) 모형이다. 하지만 변수의 수가 많아지면 데이터의 수가 충분치 못해 다항분포의 모수를 예측하기가 어려워진다. 본 논문에서는 이 문제점을 해결하기 위해 주어진 이항데이터의 정보를 저차원 프로젝션(projection)을 이용하여 최대한으로 축약한 후 이를 최대 엔

---

1) This research is supported in part by the 2002 Soong Sil University Research Fund.
2) Professor, Department of Statistics, SoongSil University, Seoul, 156-743, Korea.
E-mail: jjlee@stat.soongsil.ac.kr
3) Department of Statistics, Soong Sil University, Seoul, 156-743, Korea.

트로피(entropy) 방법으로 다항분포 모수를 추정하는 방법을 제안한다.

2절에서는 판별분석연구의 최근 동향을 살펴보고, 본 논문에서 제안하는 모형과 비교하기 위한 다섯 가지 판별모형에 대해 간단한 요약을 한다. 3절에서는 최대 엔트로피 분포를 이용한 이항데이터의 판별모형을 제안하고, 4절에서는 이 모형의 진단을 위하여 모의실험으로 다른 모형과 비교한 결과를 살펴본다. 5절에서는 결론과 향후과제에 대한 제안을 한다.

2. 판별분석연구의 최근 동향

편의상 두 그룹 w_1과 w_2에 대한 판별분석을 살펴보자. n개의 확률변수를 $\vec{x} = (x_1, x_2, \cdots, x_n)$, 각 그룹의 확률밀도함수를 $f_i(\vec{x})$, 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.

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

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

최근에 Asparoukhov & Krzanowski(2001)는 많이 이용되는 13가지의 판별분석모형을 다섯 종류의 의학 및 심리학 실험에서 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. 본 절에서는 위의 논문에서 이항데이터의 판별분석에 어느 정도의 효과를 갖는 다섯 가지 모형, 즉, 선형판별(linear discriminant)모형, 이차판별(quadratic discriminant)모형, k-근접이웃(k-nearest neighbor)모형, 로지스틱(logistic)모형, 신경망네트워크(neural network)모형들을 간단히 요약하여 보았다. 이 모형들은 3절에서 제안하는 최대 엔트로피분포 모형과 비교된다.

2.1 선형판별(linear discriminant)모형

두 그룹의 분포 $f_i(\vec{x})$가 평균이 다르고($\vec{\mu}_1 \neq \vec{\mu}_2$) 공분산행렬이 같은($\Sigma_1 = \Sigma_2 = \Sigma$) 다변량정규분포 $N(\vec{\mu}_i, \Sigma)$라면 [식 1]은 다음과 같은 선형판별식이 된다.

$$L(\vec{x}) = [\vec{x} - \frac{1}{2}(\vec{\mu}_1 + \vec{\mu}_2)]^T \Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_2)$$

이 선형판별함수는 오분류확률(misclassification probability)을 최소로 하는데 두 그룹이 멀리 떨어질수록 오분류확률은 작아진다.

2.2 이차판별(quadratic discriminant)모형

두 그룹의 분포 $f_i(\vec{x})$가 평균이 다르고($\vec{\mu}_1 \neq \vec{\mu}_2$) 공분산행렬도 서로 다른 ($\Sigma_1 \neq \Sigma_2$) 다변량 정규분포 $N(\vec{\mu}_i, \Sigma_i)$라면 [식 1]은 다음과 같은 이차형식의 판별식이 된다.

$$Q(\vec{x}) = \frac{1}{2}\ln(|\Sigma_2|/|\Sigma_1|) - \frac{1}{2}(\vec{x} - \vec{\mu}_1)^T \Sigma_1^{-1}(\vec{x} - \vec{\mu}_1) + \frac{1}{2}(\vec{x} - \vec{\mu}_2)^T \Sigma_2^{-1}(\vec{x} - \vec{\mu}_2)$$

이차판별함수를 이용할 때 발생하는 오분류확률에 대한 식은 정확히 알려져 있지 않지만 재크나이프 방법 등을 사용하여 구하기도 한다.

2.3 로지스틱(logistic)모형

로지스틱(logistic) 판별모형은 독립변수들의 분포형태에 대한 가정을 요구하지 않는 모형으로서 특히 독립변수들이 이산형과 연속형으로 혼합되어 있는 경우에 널리 사용되는 모형이다. 그룹변수를 y라 하였을 때 $\vec{x}$가 그룹 1에 속할 확률을 $p = P(y=1|\vec{x})$라 표시하면 로지스틱 판별모형은 다음과 같다.

$$Logistic(\vec{x}) = \ln\left[\frac{P(y=1|\vec{x})}{1-P(y=1|\vec{x})}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

그룹 1에 속할 확률 p가 그룹 2에 속할 확률 $(1-p)$ 보다 크면 관측치를 그룹 1로 분류한다.

2.4 k-근접이웃(nearest neighbor)모형

k-근접이웃모형은 분류하고자 하는 그룹에 대해서는 알고 있지만 각 그룹의 확률밀도함수를 추정하기가 어려울 때 많이 이용된다. 이 모형은 굳이 표본에 대한 확률 모수들을 구하지 않고 표본의 값을 그대로 좌표에 표시하여 준거집합(reference set)에서 가장 유사하거나 거리상으로 가까운(nearest) 그룹에 속하는 것으로 분류하는 방법이다. 최근접의 측도로서 사용되는 거리들에는 유클리드(Euclidean) 거리, 절대거리, 최대거리, 민코우스키(Minkowski) 거리 등이 있다.

2.5 신경망(neural network)모형

인체의 신경망을 모델로 한 판별에 대한 수리적 모형을 인공 신경망(neural network)모형이라 한다. 가장 널리 사용되는 모형은 다층인식자(multilayer perception: MLP)신경망으로서 MLP는 입력층(input layer), 은닉층(hidden layer), 그리고 출력층(output layer)이 서로 네트워크 형식으로 엮어진 모형이다. 이러한 입력층과 은닉층은 선형 또는 비선형의 결합함수(combination function)와 활성함수(activation function)로 연결되어 있다. 출력층은 목표변수에 대응하는 마디들을 갖는데 이를 이용하여 자료를 분류한다.

3. 최대 엔트로피 분포를 이용한 이항데이터의 판별모형

이항데이터의 다양한 형태를 잘 나타낼 수 있는 원시적인 모형은 이항변수의 모든 값이 나타날 확률을 고려하는 다항분포(multinomial distribution) 모형이다. [식 1]에서 $f_1(\vec{x})$와 $f_2(\vec{x})$를 서로 다른 모수를 갖는 다항분포라 가정하면 이항데이터의 판별분석은 분류오류가 최소화되는 모형이 될 수 있다. 하지만 변수의 수가 증가하면 추정되어야 하는 모수가 지수적으로 증가하게 됨으로, 모든 모수의 추정이 어렵고 추정에 필요한 데이터의 수가 충분치 못해서 수치해석적 문제가 발생하게 된다. 예를 들어 변수가 10개 일 때 다항분포를 가정하면 예측하여야 하는 모수의 수는 $2^{10}-1$ = 1024 나 된다. 하지만 우리가 현실적으로 얻을 수 있는 훈련 데이터의 수는 잘 해야 몇 백 개정도 이어서 이를 이용해서 전체 다항분포의 모수를 예측하는 것은 불가능하다.

본 논문에서는 기본적으로 다항분포 모형을 이용하면서 이 모형의 위와 같은 문제점을 해결하기 위해 다음과 같은 최대 엔트로피(entropy) 분포를 이용한 판별모형을 제안한다.

단계 1: (변수의 축약)

훈련을 위한 두 그룹의 이항데이터에서 모든 가능한 저차원 프로젝션(projection)에 대한 분포를 조사하여 이 중에서 판별능력을 최대로 갖는 변수들의 집합을 추출한다. 이 때 판별능력의 기준은 두 분포함수의 동질성 검정에 대한 카이제곱 값이나 가우시안(Gaussian) 측도 등을 이용할 수 있다. 판별능력이 큰 변수 집합 $\vec{x}_s$($\vec{x}$의 부분집합)가 N개 있을 때 이들의 프로젝션분포를 $f_1^{(j)}(\vec{x}_s)$와 $f_2^{(j)}(\vec{x}_s)$, $j=1,2,\cdots,N$ 이라 하자.

단계 2: (최대 엔트로피 분포의 추정)

단계 1에서 구한 판별능력이 큰 저차원 프로젝션분포를 이용하여 두 모집단분포 $f_1(\vec{x})$와 $f_2(\vec{x})$를 최대 엔트로피 이론(maximum entropy principle)을 이용하여 추정한다. 즉 $k=1,2$에 대하여

$$\text{Maximize} \int f_k(x) \ \ln f_k(x)\,dx$$

$$\text{Subject to projection } f_k^{(j)}(x) \quad j=1,2,\cdots,N$$

예를 들어 집합 $S=\{\vec{x}:(0,0,\cdots,0),(0,0,\cdots,1),\cdots,(1,1,\cdots,1)\}$를 n개의 이항변수가 가질 수 있

는 모든 불린(Boolean) 원소들의 집합이라 하고, 그룹 1과 그룹 2에서 각각의 원소들이 갖는 확률을 $p_{\vec{x}}^{(1)}$과 $p_{\vec{x}}^{(2)}$라 하자. 그리고 $S_{ij=(r,s)}$ 를 $x_i=r$, $x_j=s$, (여기서 $r=0,1$; $s=0,1$)를 갖는 원소들의 집합이라 하고 $c_{\ldots r\ldots s\ldots}$를 변수 $x_i=r$ 와 $x_j=s$ 인 2차원 주변확률이라 하자. 모든 가능한 2차원 주변확률을 이용하여 최대 엔트로피 분포 $p_{\vec{x}}^{(1)}$를 추정하는 식은 다음과 같은 비선형 최적화 문제로 표현할 수 있다.

$$\text{Maximize} \quad -\sum_{\vec{x}\in S} p_{\vec{x}}^{(1)} \ln p_{\vec{x}}^{(1)}$$

subject to

$$\sum_{\vec{x}\in S} p_{\vec{x}}^{(1)} = 1 \tag{2}$$

$$\sum_{\vec{x}\in S_{i,j=(r,s)}} p_{\vec{x}}^{(1)} = c_{\ldots r\ldots s\ldots}, \quad i=1,2,\cdots,n-1; \quad j=i+1,\cdots,n; \quad r=0,1; \quad s=0,1$$

같은 방법으로 그룹 2에 대한 최대 엔트로피 분포 $p_{\vec{x}}^{(2)}$를 찾을 수 있고, 3차원 주변확률분포나 기타 불린원소에 대한 확률정보를 이용한 최대 엔트로피 분포의 추정도 유사하다.

단계 3: (판별식)

만일 판별이 필요한 새로운 데이터, $\vec{x_o}$ 가 있으면 두 그룹의 오분류 손실이 같고, 사전확률 π_1과 π_2가 같을 때 추정된 최대 엔트로피 분포를 이용하여 확률이 높은 그룹으로 판별한다. 즉

$$p_{\vec{x_o}}^{(1)} \geq p_{\vec{x_0}}^{(2)} \quad \text{이면 } \vec{x_o} \text{ 를 그룹 1로 분류, 아니면 그룹 2로 분류} \tag{3}$$

최대 엔트로피 분포를 이용한 판별모형은 [식 2]와 같은 비선형 최적화 문제를 풀어야 하기 때문에 변수의 수가 많아지면 현실적으로 해를 구하기가 쉽지 않을 수 있다. 이러한 경우에 만일 n개의 확률변수 $\vec{x} = (x_1, x_2, \cdots, x_n)$ 가 서로 독립인 여러 개의 부변수(sub variable)로 나눌 수 있다면 확률밀도함수 $f_i(\vec{x})$의 추정은 근사적으로 각각의 부변수에 대한 확률밀도함수를 추정한 후 이들의 곱으로 표시할 수 있다.

4. 판별모형의 비교실험

최대 엔트로피 분포를 이용한 판별모형과 2절에서 소개한 다섯 가지 판별모형에 대해 모의실험으로 비교하여 보았다. 모의실험을 위한 데이터의 생성과 구체적인 실험방법은 다음과 같다.

데이터:

n개의 이항변수가 가질 수 있는 모든 불린(Boolean) 원소들을 다음과 같은 순서로 표시하자.

$$\begin{array}{c} x_1\, x_2 \cdots x_n \\ \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \\ & \cdots & & \\ 1 & 1 & \cdots & 1 \end{pmatrix} \end{array} \tag{4}$$

각각의 불린(Boolean)원소가 가질 수 있는 다항분포 확률의 생성은 여러 가지 분포함수를 이용할 수 있는데, 두 모집단에 대해 <그림 1>과 같이 일반적으로 많이 나타날 수 있는 세 가지 패턴으로 구별하여 보았다. 첫 번째 패턴은 두 모집단의 불린원소들이 서로 많이 겹치는 경우로서 모집단 1을 지수분포(m=1.0), 모집단 2는 지수분포(m=0.1)를 이용하였다. 두 번째 패턴은 불린원소들이 겹치는 경우가 첫 번째 보다는 적지만 그래도 서로 겹치는 것이 어느 정도 있는 경우로서 모집단 1을 지수분포(m=0.3), 모집단 2는 지수분포(m=0.3)의 대칭인 분포를 이용하여 표본을 생성하였다. 세 번째 패턴은 두 모집단의 불린원소들이 서로 겹치는 확률이 매우 적은 경우로서 모집단 1을 지수분포(m=0.5), 모집단 2는 지수분포(m=0.5)의 대칭인 분포를 이용하여 표본을 생성하였다.

데이터를 생성하기 위해서는 가정된 분포를 2^n개의 구간으로 나누어 각 구간의 확률을 구한 다음 역변환법(inverse transformation method)에 의해 원하는 표본 개수만큼 이항데이터를 생성하였다. 지수분포인 경우 무한대까지의 값을 가질 수 있으므로 최대값은 99.99퍼센타일(percentile)을 이용한 후 확률분포가 되도록 조정하였다.

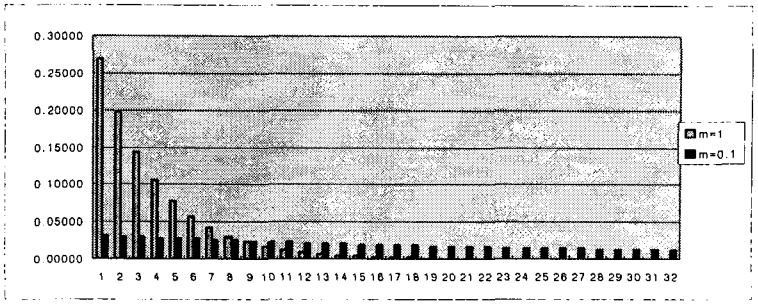
실험방법:

가상의 이항데이터를 이용한 구체적인 모의실험 방법은 다음과 같다.

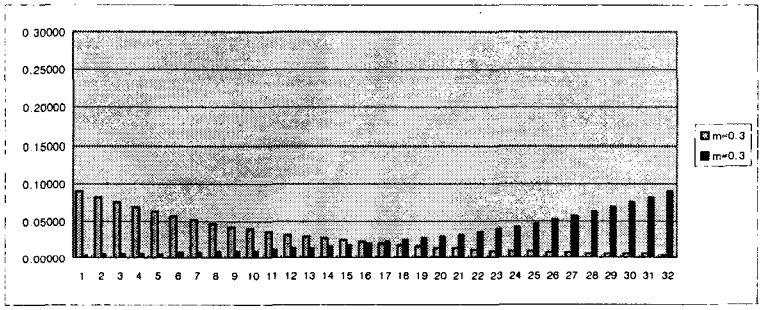
1) 두 모집단에 대한 다항분포의 확률을 얻기 위해 위에서 가정한 세 종류의 모집단에서 표본을 추출한다. 표본의 개수는 두 모집단에서 각각 50, 100, 300개를 추출한다.
2) 추출된 표본의 모든 가능한 2차원 주변확률을 구한 후 그룹 1과 그룹 2가 통계적으로 서로 다른 분포인지 카이제곱 동질성검정을 하여 판별에 의미 있는 저차원 분포를 선택한다. 본 논문에서는 실험의 편의상 가능한 2차원 주변확률을 모두 이용하였다.
3) 저차원 주변확률을 이용하여 [식 2]로 최대 엔트로피 분포를 추정한다.
4) 추정된 분포와 [식 3]을 이용하여 표본자료에 대한 판별을 실시하여 정분류율을 계산한다.
5) 같은 자료에 대해 2절에서 소개한 다섯 가지 판별방식을 적용하여 정분류율을 계산한다. 이 실험에는 SAS, S-PLUS 등을 이용하였다.
6) 위의 1)에서 5)까지의 실험을 30회씩 반복하여 정분류율의 평균과 표준편차를 계산한다.

최대 엔트로피 분포를 추정하기 위해서는 비선형 최적화 문제를 풀어야 하는데 본 연구에서는 GRG2 소프트웨어를 이용하였다. 변수의 수가 4개 이하는 현실성이 약하므로 제외하고, 변수의 개수가 8개 이상이면 다항분포의 불린원소의 수가 256개나 됨으로 프로그램의 제약으로 인하여 변수의 수를 5, 6, 7개에 대하여만 제한적으로 실험하였다. 하지만 변수의 수가 7개인 경우에도 [식

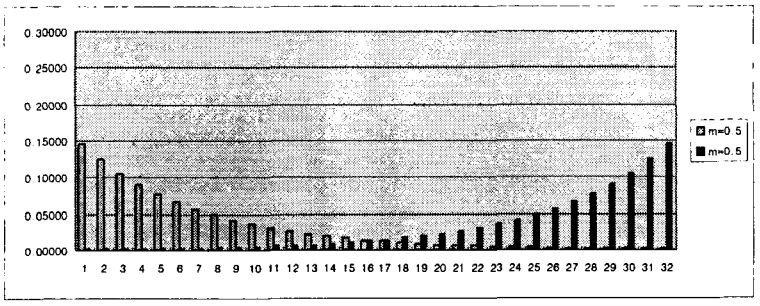
2]의 비선형 최적화 문제는 2^7=128개의 미지변수에 대한 해를 구하여야 함으로 초기치 문제나 지역 최적해(local solution) 문제 등 많은 어려움이 발생하였다.



- 두 모집단의 불린 원소들이 확률적으로 많이 겹치는 경우 -
(모집단 1은 지수분포(m=1.0), 모집단 2는 지수분포(m=0.1), n=5)



- 두 모집단의 불린 원소들이 확률적으로 중간정도로 겹치는 경우 -
(모집단 1은 지수분포(m=0.3), 모집단 2는 지수분포(m=0.3)의 대칭분포, n=5)



- 두 모집단의 불린 원소들이 확률적으로 겹치지 않는 경우 -
(모집단 1은 지수분포(m=0.5), 모집단 2는 지수분포(m=0.5)의 대칭분포, n=5)

<그림 1> 모의실험을 위한 가정된 세 종류의 모집단 패턴

실험결과:

모의실험 결과가 [표 1]에서 [표 3]까지 정리되어 있다. 세 종류의 모집단 패턴 모두에서 로지스틱(LD) 모형이 우수한 판별력을 보이는 것으로 나타났고, 최대 엔트로피 분포를 이용한 판별방식은 변수의 수나 표본의 수에 따라 서로 다른 판별력을 보이는 것으로 나타났다.

[표 1] 모집단 1은 지수분포(m=1.0), 모집단 2는 지수분포(m=0.1)를 이용하여 표본을 30회 생성하고 실험을 하였을 때 평균 정분류율(표준오차)
(MEP:최대엔트로피 LDF:선형 QDF:이차 3NN:근접이웃 LD:로지스틱)

총변수	표본수	MEP	LDF	QDF	3NN	LD	신경망
5	50	0.781(0.091)	0.796(0.036)	0.740(0.050)	0.728(0.073)	0.847(0.041)	0.734(0.068)
	100	0.751(0.033)	0.786(0.024)	0.734(0.049)	0.734(0.043)	0.830(0.028)	0.759(0.041)
	300	0.767(0.013)	0.785(0.016)	0.779(0.019)	0.769(0.028)	0.831(0.012)	0.778(0.025)
6	50	0.752(0.095)	0.801(0.037)	0.752(0.047)	0.708(0.071)	0.850(0.043)	0.750(0.057)
	100	0.734(0.036)	0.789(0.029)	0.738(0.043)	0.730(0.047)	0.848(0.030)	0.751(0.043)
	300	0.728(0.057)	0.792(0.016)	0.777(0.027)	0.758(0.035)	0.849(0.016)	0.770(0.027)
7	50	0.700(0.140)	0.799(0.031)	0.757(0.047)	0.712(0.078)	0.874(0.033)	0.739(0.064)
	100	0.660(0.100)	0.800(0.026)	0.749(0.030)	0.712(0.078)	0.862(0.027)	0.732(0.050)
	300	0.646(0.093)	0.779(0.061)	0.775(0.031)	0.728(0.029)	0.853(0.012)	0.767(0.032)

[표 2] 모집단 1은 지수분포(m=0.3), 모집단 2는 지수분포(m=0.3)의 대칭분포를 이용하여 표본을 30회 생성하고 실험을 하였을 때 평균 정분류율(표준오차)
(MEP:최대엔트로피 LDF:선형 QDF:이차 3NN:근접이웃 LD:로지스틱)

총변수	표본수	MEP	LDF	QDF	3NN	LD	신경망
5	50	0.785(0.094)	0.826(0.033)	0.843(0.035)	0.773(0.062)	0.892(0.035)	0.763(0.061)
	100	0.754(0.092)	0.809(0.029)	0.825(0.020)	0.787(0.038)	0.878(0.023)	0.787(0.045)
	300	0.712(0.083)	0.814(0.011)	0.814(0.010)	0.796(0.022)	0.877(0.011)	0.809(0.022)
6	50	0.770(0.133)	0.835(0.031)	0.853(0.028)	0.778(0.078)	0.898(0.034)	0.766(0.070)
	100	0.734(0.087)	0.813(0.028)	0.828(0.023)	0.773(0.054)	0.884(0.027)	0.762(0.052)
	300	0.707(0.043)	0.822(0.015)	0.823(0.014)	0.797(0.021)	0.885(0.013)	0.808(0.028)
7	50	0.681(0.120)	0.831(0.037)	0.858(0.034)	0.765(0.068)	0.903(0.027)	0.711(0.078)
	100	0.666(0.129)	0.821(0.023)	0.834(0.019)	0.773(0.037)	0.894(0.021)	0.768(0.050)
	300	0.647(0.049)	0.819(0.015)	0.823(0.011)	0.791(0.021)	0.886(0.011)	0.799(0.025)

[표 3] 모집단 1은 지수분포(m=0.5), 모집단 2는 지수분포(m=0.5)의 대칭분포를 이용하여 표본을 30회 생성하고 실험을 하였을 때 평균 정분류율(표준오차)
(MEP:최대엔트로피 LDF:선형 QDF:이차 3NN:근접이웃 LD:로지스틱)

총변수	표본수	MEP	LDF	QDF	3NN	LD	신경망
5	50	0.917(0.055)	0.929(0.026)	0.932(0.019)	0.896(0.049)	0.973(0.025)	0.903(0.046)
	100	0.896(0.032)	0.921(0.019)	0.923(0.016)	0.903(0.022)	0.972(0.009)	0.906(0.031)
	300	0.889(0.032)	0.922(0.009)	0.921(0.009)	0.914(0.016)	0.969(0.004)	0.909(0.017)
6	50	0.880(0.065)	0.921(0.029)	0.927(0.023)	0.905(0.047)	0.976(0.015)	0.876(0.012)
	100	0.853(0.069)	0.921(0.019)	0.923(0.018)	0.914(0.026)	0.970(0.016)	0.902(0.038)
	300	0.084(0.040)	0.923(0.008)	0.923(0.007)	0.909(0.020)	0.972(0.006)	0.909(0.021)
7	50	0.800(0.175)	0.929(0.022)	0.936(0.017)	0.896(0.044)	0.983(0.011)	0.889(0.050)
	100	0.803(0.054)	0.925(0.015)	0.929(0.012)	0.890(0.041)	0.975(0.012)	0.891(0.033)
	300	0.797(0.088)	0.923(0.011)	0.924(0.010)	0.907(0.021)	0.972(0.005)	0.912(0.017)

[표 1]은 두 모집단의 불린원소들이 서로 많이 겹치는 경우로서, 로지스틱(LD) 모형이 표본의 수나 변수의 수와 관계없이 우수한 정분류율을 보이고, 선형판별모형(LDF)이 두 번째로 좋은 정분류율을 보이고 있다. 최대 엔트로피 모형(MEP)은 변수의 수가 5개인 경우 상대적으로 어느 정도 경쟁력이 있다고 볼 수 있으나, 변수의 수가 6개나 7개인 경우 판별력이 상대적으로 더 떨어짐을 알 수 있다. 이는 비선형방정식의 해의 수렴성 문제에 원인이 있는 것으로 생각된다. [표 2]는 두 모집단의 불린원소들이 서로 겹치는 것이 어느 정도 있는 경우로서, 역시 변수의 수나 표본의 수에 관계없이 로지스틱 모형(LD)이 제일 우수한 정분류율을 보이고, 그 다음으로는 이차판별모형(QDF)이 우수한 정분류율을 보여준다. [표 3]은 두 모집단의 불린원소들이 서로 겹치는 확률이 매우 적은 경우로서, 역시 변수의 수나 표본 수에 관계없이 로지스틱 모형(LD)이 우수한 정분류율을 보인다. 그 다음으로는 이차판별모형(QDF)과 선형판별모형(LDF)이 비슷한 정분류율을 보여준다.

5. 결론 및 향후과제

본 연구에서는 이항데이터의 주변확률을 이용하여 최대 엔트로피 방법으로 모집단의 분포를 예측한 후 데이터를 판별하는 모형을 제안하고, 모의실험을 이용하여 다른 분류방법들과의 비교 실험을 실시하였다. 그 결과 세 종류의 모집단 패턴 모두에 대해 로지스틱 모형이 우수한 판별력을 보이고, 본 논문에서 제안된 최대 엔트로피 분포를 이용한 판별 모형은 특정한 변수의 수나 표본의 수에 대해서만 어느 정도의 상대적인 경쟁력이 있는 것을 확인하였다. 만족할만한 성과는 아니지만 향후 제약식에 관한 정보를 효율적으로 관리하는 모형으로 발전시킬 수 있는 가능성을 본 것으로 생각된다. 하지만 본 연구에서 최대 엔트로피 이론을 이용한 확률분포의 추정은 이항변수의 수가 n개일 때 2^n개의 변수의 수를 갖는 비선형 최적화 문제의 해를 구하여야 함으로 지역해나 초기해의 문제가 발생함을 알 수 있다. 직접적인 엔트로피 분포의 계산은 한계가 있을 수밖에 없고, 이를 극복하기 위해 라그랑지 승수법을 활용한 비선형 시스템 방정식을 통한 해법을 이용하여 구할 수 있으리라 판단된다. 변수의 선택에 있어서도 2개의 변수를 통한 주변확률을 통해서 최대 엔트로피 분포를 추정하였으나 3개의 변수를 통한 분포의 추정 또한 향후 실험되어져야 할 것이다.

참고문헌

[1] Asparoukhov, O.K and Krzanowski, W.J. (2001) A comparison of discriminant procedures for binary variables. Computational Statistics and Data Analysis, 38, 139-160.
[2] Duda, R.O., Hart, P.E., and Stork, D.G. (2001) Pattern Classification, Wiley.
[3] Johnson, R. and Wichern, D. (1988) Applied Multivariate Statistical Analysis, Prentice Hall.
[4] Lachenbruch (1981) Discriminant Analysis, Prentice Hall.

[2003년 9월 접수, 2003년 11월 채택]