판별 분석을 위한 Algorithm의 개발

김 종 우
함 형 범*
김 재 현**

〈목 차〉

Ⅰ. 서 론

판별 분석(Discriminant Analysis)이란 두개 이상의 집단중 어느 모집단에서 추출된 것인지 확실히 알고있는 다변량(Multivariate) 확률 분포(Random Distribution)의 추정값들을 이용하여 새로운 개체가 주어졌을때 이 개체가 어느 모집단에서 추출된 것인지를 결정하게 된다. 이때 잘못 분류될 확률(Probability of Misclassification)이 최소가 되도록 주어진 확률 표본(Random Sample)의 추정값들을 이용하여 어떤 기준을 세우는데 사용되는 분석방법이다.

역사적으로는 1936년 Fisher가 제안한 선형 판별 함수(Linear Discriminant Function)가 판별 분석의 첫 시도라 하겠다. 그러나, 각 모집단의 자료가 질적인(Qualitative) 성분으로 이루어져 있는 경우 이 집단에 근거하여 어떤 모집단으로 할당하는 경우 Fisher의 선형 판별 함수는 의미가 없으므로 Hill(1966), Cochran과 Hopkins(1961), Gilbert(1968), Moore(1973), Dillon 과 Goldstein(1978) 등이 자료가 질적인 성분으로 구성된 경우에 대하여 연구하였다.

이러한 질적인 자료들을 판별 분석하기 위한 컴퓨터 프로그램들이 존재하고 있는데 현

* 서경대학교 응용통계학과 전임강사
** 동국대학교 통계학과 박사과정 수료

재는 각 모형에 대해 개별적으로 판별 분석에 이용하고 있다. 그러므로 현재는 각 모형에 대해 개별적으로 판별 분석에 이용하고 있다. 따라서 이런 컴퓨터 프로그램을 모아 여러 모형에 대해 동시에 판별 분석을 할 수 있는 프로그램으로 만들 필요성이 증가되고 있다. 하지만 기존의 통계 프로그램인 SPSS 나 SAS는 질적인 자료들의 이산 판별 분석에는 큰 도움을 주지 못하고 있는 실정이다.

따라서, 본 논문은 자료의 모든 변수가 이분적(Dichotomous)이며 두개의 상호배반 모집단으로 이루어진 자료를 바탕으로 대상물을 어느 모집단중 하나의 모집단으로 분류(Classification)하는 이산 판별 모형에 따른 판별 분석을 하는 Algorithm의 개발에 목적을 둔다.

Ⅱ. 이산 판별 모형(Discrete Discriminant Model)

X_1, X_2, …, X_j는 이산 확률 변수(Discrete Random Variable) 이고 각각의 변수가 가질 수 있는 값들을 각각 S_1, S_2, …, S_p개로 이루어진 유한개의 다른 값들이라고 가정하자. 또한 표본 공간(Sample Space) Ω는 $S = \prod_{j=1}^{p} S_j$인 반응으로 구성되어 있거나 •다항 분포(Multinomial distribution)를 하는 확률 벡터(Random Vector) $X=(X_1, \cdots, X_p)$에 의해 생성되어 진다고 가정하자. 상호 배반 모집단 π_1과 π_2가 사전 확률(Prior Probability) δ_1과 δ_2로 결합되어 있다면 비조건 밀도 함수(Unconditional Density Function)는

$$\begin{aligned} g(x) &= \delta_1 f_1(x) + \delta_2 f_2(x) \\ &= g_1(x) + g_2(x) \end{aligned} \tag{2.1}$$

이다. 여기서 $f_i(x)$는 상호 배반인 두 개의 모집단 π_1과 π_2의 분류 조건부 다항 밀도 함수(Class Conditional Multinomial Mass Function)이고 δ_i는 $P(\pi = \pi_i)$, $(i = 1,2)$ 이다. 또 $g_i(X)$를 모집단 i에서의 판별점수(Discriminant Score)라 한다. 모집단이 두 개일 경우 분류 규칙은 표본 공간의 순서 분할 (Ordered Partition) $D=\langle D_1, D_2\rangle$로 규정되어진다. 여기서, 분류 규칙은 $X \in D_i$ $(i=1,2)$이면 x를 π_1에 할당한다.

$X = x \in D_i$일때 잘못 분류될 조건부 확률(Conditional Probability of Misclassification)은 아래와 같다.

$$t(D|X = x) = g_i(x)/g_j(x), \; i \neq j \tag{2.2}$$

반응 벡터 x가 주어졌을때 (2.2)식에 기대값을 취하면 비조건부 오차율 (Unconditional Error Rate)

$$t(D) = E\{t(D|x)\} = \sum_{D_2} g_1(x) + \sum_{D_1} g_2(x) \quad (2.3)$$

를 구할 수 있다.

표본을 근거로 하여 분류 규칙을 설정하는데에는 두 개의 표본 추출 방법이 있다. 하나는 혼합 모집단에서 N개의 대상물를 추출하는 경우와 다른 하나는 두 개의 모집단 π_1과 π_2로 부터 각각 정해진 사전 확률을 가지고 표본의 크기 n_1과 n_2로 부터 독립적으로 표본 추출하는 경우가 있다. 본 연구에서는 후자의 경우에만 국한하기로 한다.

1. 완전 다항 모형(The Full Multinomial Model)

이 모형은 비조건부 오차율 (2.3)식을 최소화 하여 만든 모형이다.

즉 t는 모든 분류 규칙의 영역인 D에서 함수이고 $t(D) = t^* = \inf\ D' \in D(D')$이면 적절한 분류 규칙이 되고 최적 분할 D가 $g_1(x) > g_2(x)$이면 $x \in D_1$ 혹은 $x \in D_2$로 임의 할당하게 된다. 그러므로

$$t = t(D) = \sum \min(g_1(x), g_2(x)) \quad (2.4)$$

이고 여기서 합계는 표본 공간 Ω에 속하는 모든 반응 벡터 x에 관해서 합친 것이다.

X_j (j = 1,2, …, P)는 0 또는 1을 가지는 이분 변수(Dichotomous Variable)라 가정하고 P차원으로부터 유도된 다항 분포는 각 모집단 π_i에서 2^P-1 개의 모수를 가지는 S $= 2^P$개의 반응 벡터 x로 이루어져 있을때 불편 추정량(Unbiased Estimator)

$$\hat{p}(x) = x(\pi_i) = n_i(x)/n_i, \quad (i=1,2) \quad (2.5)$$

을 가지는 $p(x = x|\pi_i)$를 추정하는데 근거한다. 여기서 $n_i(x)$는 모집단 π_i에서 n_i개를 표본 추출하였을때 얻어지는 각 형태의 반응 벡터 x를 가지는 각각의 갯수를 나타낸다.

〈분류 규칙〉

$\hat{\delta}_1(n_1(x)/n_1) > (1-\hat{\delta}_1)(n_2(x)/n_2)$이면 x를 π_1에 분류

$\hat{\delta}_1 (n_1(x)/n_1) < (1-\hat{\delta}_1) \quad (n_2(x)/n_2)$ 이면 x를 π_2에 분류

$\hat{\delta}_1 (n_1(x)/n_1) = (1-\hat{\delta}_1) \quad (n_2(x)/n_2)$ 이면 x를 π_1 혹은 π_2에 분류 (2.6)

여기서 $n=n_1+n_2$이고 $\delta_1 = n_1/n$이다.

2. 차수(r)가 1인 최소 근접법 모형(The Nearest Neighbor r=1)

표본에 근거한 완전 다항 모형의 표본 추출 오차(Sampling Error)를 줄이는 방법으로 Hill에 의해 최초로 제안되었고 반응 벡터(Response Patterns) x를 분류하기 위하여 표본을 기초로 하여 우도비 절차(Likelihood Ratio Procedure)를 사용했을때 반응 벡터 x의 p개 성분(Component) 중에서 r개 이상 차이가 나지않는 모든 반응벡터를 분류하기 위해 사용하는 모형이다.

모든 반응 벡터 x에 대해서

$$T_j = \{y_j \mid (x-y_j) \quad (x-y_j)' \leq r\} \tag{2.7}$$

이고 여기서 T_j는 반응 벡터 Y_j의 집합이며 이 집합의 각 원소는 x로 부터 r개 이상의 차이가 나지 않는다.

〈분류 규칙〉

$\hat{\delta}_1 \sum_{Tj} (n_1(y_j)/n_1) > (1-\hat{\delta}_1) \sum_{Tj} (n_2(y_j)/n_2)$ 이면 x를 π_1에 분류

$\hat{\delta}_1 \sum_{Tj} (n_1(y_j)/n_1) > (1-\hat{\delta}_1) \sum_{Tj} (n_2(y_j)/n_2)$ 이면 x를 π_2에 분류

$\hat{\delta}_1 \sum_{Tj} (n_1(y_j)/n_1) = (1-\hat{\delta}_1) \sum_{Tj} (n_2(y_j)/n_2)$ 이면 x를 π_1 혹은 π_2에 분류 (2.8)

3. Bahadur 모형 (The Bahadur Model)

반응 벡터가 0,1로 구성되어 있고 i번째 모집단으로부터 관찰되어진 반응 벡터 x의 확률을 $\pi_i(x)$라 하자. Bahadur(1961)는 다항 확률 $\pi_i(x)$를 재모수화(Reparame tization)하여 다음과 같음을 증명하였다.

즉, 확률 변수(Random Variable) x_i는 베르누이 분포(Bernoulli Distribution)를하고 i 번째 (i=1,2) 모집단에서 x_j의 (j=1,2, …, P) 기대값은

$P_{ij}=E_i(x_i)=P_i\{x_i=1\}$ 이어서 이를 표준화하면

$$z_{ij} = (x_j - p_{ij}) / \sqrt{p_i(1-p_{ij})} \tag{2.9}$$

이고 대응되는 상관 계수는 다음과 같이 정의된다.

$$\rho_i(1,2,3,\cdots,p) = E(z_{i1}\ z_{i2}\ \cdots\ z_{ip}) \tag{2.10}$$

그때 $\pi_i(x)$는 다음과 같음을 보였다.

$$\pi_i(x) = \prod_{j=1}^{p} p_{ij}^{xj}(1-p_{ij})^{1-xj}\left[1+\sum_{j<k}\rho_i(jk)z_{ij}\ z_{i1}+\sum_{j<k<l}(\rho_i(jk)z_{ij}z_{ik}z_{il}+\cdots+\right.$$
$$\left.\rho_i(1,2\ \cdots,\ p)z_{i11}\ z_{i2}\cdots z_{ip}\right] \tag{2.11}$$

1) 일차 모형(The First Order Model)

추정해야할 모수(Parameter)을 줄이는 수단으로 변수들이 모두 독립이라고 가정하면 각 모집단에서 추정해야 할 모수의 갯수가 2^p-1에서 P개로 감소하게 된다. (2.8)식에서 상관 관계(Correlation)이 존재하는 모두 항을 제거하였을때 이것을 일차 모형(The First Order Model)이라 한다. 그때 Pij에 대한 불편 추정량(Unbiased Estimator)는

$$\hat{p}_{ij} = \sum_{S_j}(n_i(x)/n_i) \tag{2.12}$$

이고, 여기서 S_j는 $X_j=1$인 모든 반응 벡터 x의 집합이다.

〈분류 규칙〉

$\hat{\delta}_1(\hat{\pi}_1(x:[1])) > (1-\hat{\delta}_1)(\pi_2(x:[1]))$이면 x를 π_1에 분류

$\hat{\delta}_1(\hat{\pi}_1(x:[1])) < (1-\hat{\delta}_1)(\pi_2(x:[1]))$이면 x를 π_2에 분류

$\hat{\delta}_1(\hat{\pi}_1(x:[1])) = (1-\hat{\delta}_1)(\pi_2(x:[1]))$이면 x를 π_1 혹은 π_2에 분류 (2.13)

여기서 $\hat{\pi}_1(x:[1]) = \prod_{j=1}^{p} p_{ij}^{x_j}\ (1-p_{ij})^{1-xj}$이다.

2) 이차 모형(The Second Order Model)

$\pi_i(x)$에서 이차 이상의 상관관계가 존재하는 모든 항을 제거하면 P_{ij}와 $\rho_i(jk)$항만 존재한다. 이러한 모형을 이차모형(The Second Order Model)이라 한다. 이차 모형의

표본 상관 계수는

$$\hat{\rho}_i(jk) = \left[\sum_{S_{jk}} n_i(x)/n_i - \hat{P}_{ij}\ \hat{P}_{ik} \right] / \sqrt{p_{ij}(1-p_{ij})\,p_{ik}(1-p_{ik})} \qquad (2.14)$$

은 모집단 상관 계수를 추정하기 위해서 사용된다.여기서 S_{jk}는 $X_j=1$이고 $X_k=1$인 반응 벡터 x의 집합이다. 이차 모형의 식

$\pi_i(x : [2])$은

$$\hat{\pi}_1(x : [2]) = \prod_{j=1}^{p} \hat{p}_{ij}^{x_j} (1-p_{ij})^{1-x_j} \left[1 + \sum \rho_i(jk)\ z_{ij}\ z_{ik} \right] \qquad (2.15)$$

에 의해 추정된다.

〈분류 규칙〉

$\hat{\delta}_1(\hat{\pi}_1(x : [2])) > (1-\hat{\delta}_1)(\pi_2(x : [2]))$ 이면 x를 π_1에 분류

$\hat{\delta}_1(\hat{\pi}_1(x : [2])) < (1-\hat{\delta}_1)(\pi_2(x : [2]))$ 이면 x를 π_2에 분류

$\hat{\delta}_1(\hat{\pi}_1(x : [2])) = (1-\hat{\delta}_1)(\pi_2(x : [2]))$ 이면 x를 π_1 혹은 π_2에 분류 (2.16)

4. 선형 판별 함수 (The Linear Discriminant Function)

이 모형은 1936년 Fisher가 처음 제안하였다.

두 모집단 π_1과 π_2의 평균 벡터(Mean Vector)와 분산 공분산 행렬(Variance Covariance Matrix)이 각각 μ_1, μ_2, Σ_1, Σ_2이며 두 모집단의 분산 공분산이 같다.

$\Sigma_1=\Sigma_2=\Sigma$ 라고 가정하자. 그러면 함수 벡터 $X_{(px1)}$을 선형 결합시킨 관찰치 $Y_{1(kx1)} = l'_{(1xp)} X_{(px1)}$의 분산에 대해 두 모집단 π_1, π_2의 평균차 제곱의 비는

$$\frac{\text{Y 평균차의 제곱}}{\text{Y의 분산}} = \frac{(\mu_{1y}-\mu_{2y})^2}{\sigma_y^2}$$

$$= \frac{l'(\mu_1-\mu_2)\ (\mu_1-\mu_2)'l}{l'\Sigma l} \qquad (2.17)$$

로 주어지며 여기서 μ_{1y}, μ_{2y}, σ_y^2 은 Y의 평균과 분산이고 이는 $\mu_{1y}=l'\mu_1$, $\mu_{2y}=l'\mu_2$,

$\sigma_{y2}(l'x) = l'\sum l$로 정의된다. (2.17)을 최대화 시키는 선형 결합 계수 (Linear Combination Coeffcient)들의 벡터와 선형 판별 함수는 다음과 같다.

$$\begin{aligned} l' &= (\mu_1 - , \mu_2)' \sum^{-1} \\ Y_{(1\times1)} &= l'_{(1\times p)} X_{(p\times1)} \\ &= (\mu_1 - \mu_2)' \sum^{-1} x \end{aligned} \tag{2.18}$$

그러나 μ_1, μ_2, $\sum$를 알지 못하기 때문에 각 모집단에서 표본 추출된 크기 n_1, n_2개로부터 추정된 평균 벡터 $\hat{p}_1$, $\hat{p}_2$와 합공 표본 분산 공분산 행렬 (Pooled Sample Variance Covariance Matrix) S_p를 사용하여 선형 판별 함수를 구하면

$$Y = (\hat{p}_1 - \hat{p}_2)' \; S_p^{-1} x \tag{2.19}$$

이고 또한 두 모집단에서 추정된 평균의 중앙점

$$m = 1/2 (\hat{p}_1 - \hat{p}_2)' S_p^{-1} (\hat{p}_1 + \hat{p}_2) \tag{2.20}$$

이다. 선형 판별 함수와 두 모집단에서 추정된 평균의 중앙점들간의 차는

$$\hat{L}(x : [1]) = (\hat{p}_1 - \hat{p}_2)' \; S_p^{-1} x - 1/2 \, (\hat{p}_1 - \hat{p}_2)' \, S_p^{-1} (\hat{p}_1 - \hat{p}_2) \tag{2.21}$$

이며 여기서 $\hat{p}_1'_{(1\times p)} = [\hat{p}_{11}, \; \hat{p}_{12}, \; \cdots, \; \hat{p}_{1p}]$,
$\hat{p}_2'_{(1\times p)} = [\hat{p}_{21}, \; \hat{p}_{22}, \; \cdots, \; \hat{p}_{2p}]$ 이다.

〈분류 규칙〉

$\hat{L}(x : [1]) < \ln(\hat{\delta}_1 / (1 - \hat{\delta}_1))$ 이면 x를 π_1 분류

$\hat{L}(x : [1]) > \ln(\hat{\delta}_1 / (1 - \hat{\delta}_1))$ 이면 x를 π_2에 분류

$\hat{L}(x : [1]) = \ln(\hat{\delta}_1 / (1 - \hat{\delta}_1))$ 이면 x를 π_1 혹은 π_2에 분류 (2.22)

5. 각 모형의 잘못 분류될 확률(The Probability of Misclassification)

위에서 보인 판별 모형의 우도비 (Likelihood Ratio)는 다음과 같다.

완전 다항 모형

$$\hat{L}(x : [m]) = \ln\left(\frac{n_2(x)/n_2}{n_1(x)/n_1}\right) \tag{2.23}$$

차수 최소 근접 모형

$$\hat{L}(x : [m]) = \ln\left(\frac{\sum_{Tj} (n_2(y_j)/n_2)}{\sum_{Tj} (n_1(y_j)/n_1)}\right) \tag{2.24}$$

1차 모형

$$\hat{L}(x : [1]) = \ln\left(\frac{\hat{\pi}_2(x : [1])}{\hat{\pi}_1(x : [1])}\right) \tag{2.25}$$

2차 모형

$$\hat{L}(x : [2]) = \ln\left(\frac{\hat{\pi}_2(x : [2])}{\hat{\pi}_1(x : [2])}\right) \tag{2.26}$$

이며 선형 판별 함수에서 직접 추정된 우도비는 $\hat{L}(x : [1])$ 이다.

이 5가지 판별 모형의 우도비가 $\ln(\hat{\delta}_1/(1-\hat{\delta}_1))$ 보다 값이 작을때 반응 벡터(Response Vector) x를 모집단 π_1에 분류할 확률을 $B(x : [\mu])$라 하면 분류를 할 때 사용되는 베이즈 법칙 (Bayes Rule)은 다음과 같다.

$$B(x : [\mu]) = \begin{cases} \hat{L}(x : [\mu]) < \ln(\hat{\delta}_1/(1-\hat{\delta}_1)) \\ \hat{L}(x : [\mu]) > \ln(\hat{\delta}_1/(1-\hat{\delta}_1)) \end{cases} \tag{2.27}$$

분류 모형이 추정된 우도비에 근거할때 잘못 분류될 확률(Probability of Misclassification)을 실제 오차 (Apparent Error)로 정의되며 이는 아래와 같다.

$$A[\mu] = \hat{\delta}_1 \sum B(x : [\mu])\ p_1(x) + (1-\hat{\delta}_1)\sum(1-B(x : [\mu]))\ p_2(x) \tag{2.28}$$

Ⅲ. 결 론

본 연구에서는 기존의 통계 소프트웨어가 처리할 수 없었던 이산형 자료의 판별분석을 완전 다항 모형, 최소 근접 모형, 1차 모형, 2차 모형 그리고 선형 판별 함수 모형

등을 동시에 처리할 수 있는 알고리즘을 제안하였고 이를 사용해 각 모형에 따른 잘못 분류될 확률을 구함으로써 이들 모형중 어느 것이 가장 적합한지를 판단할 수 있게 하였다.

앞으로 여기에 포함되지 못한 다른 판별 모형을 추가해 통합적인 소포트웨어를 만든다면 이산 자료의 판별 분석에서 최적의 모형을 선택할 수 있으리라 보인다.

참고 문헌

1. Bahadur, R.R. (1961). "A Representation of the Joint Distribution of Response to Dichotomous Items, in Studies in Item Analysis and Prediction", H. Solomon (ed.), Stanford University Press, California
2. Cochran, W.G. and Hopkins, C.E. (1961). "Some Classification Problem with Multivariate Qualitative Data", Biometrics, Vol.17, pp.10-32.
3. Dillon, W.R. and Goldstein, M. (1978). "On the Performance of Some Multinomial Classification Rules", Journal of American Statistical Association, Vol. 78, p.362.
4. Gilbert, E.S. (1968). "On Discrimination Using Qualitative Variables", Journal of American Statistical Association, Vol. 63, pp. 1399-1418.
5. Goldstein, M. and Dillon, W.R. (1978). "Discrete Discriminate Analysis", Wiley, New York.
6. Hills, M. (1966). "Allocation Rules and Their Error Rates (with Discussion)". Journal of the Royal Statistical Society, Series B, Vol. 28, pp. 1-31.
7. Hoel, P.G. and Peterson, R.P. (1949). "A Solution to the Problem of Optimum Classification ", Annals of Mathematical Statistics, Vol. 20, pp. 433-438.
8. Moore, D.h. Ⅱ. (1973). "Evaluation of Five Discrimination Procedures for Binary Variables", Journal of American Statistical Association, Vol. 68, pp. 339-404.
10. Welch, B.L. (1939). "Note on Discriminant Functions", Biometrics, Vol. 22, p.268.