응용통계연구 제 11권 1호, 1998년, pp. 177-184

종속구조를 가진 집단변수의 판별-분류에 관한 연구 *

황선영 † 나은정 ‡

요 약

일반적인 판별분류분석은 자료점들 간의 독립성 가정을 기본 전제로 한다. 본 연구에서는 자료점 간의 종속구조를 반영하는 판별분류기준을 제안하였으며 이를 위하여 조건부 자기 로지스틱(conditional autologistic)모형을 이용하였다. 또한 모의실험을 통해 제시된 기준을 기존의 방법과 비교하였다.

1. 서론

두 집단의 판별분류분석은 관측치들을 별개의 집단으로 분리하고, 새로운 관측치를 사전에 정의된 집단으로 할당하는 다변량분석기법이다. 판별분류분석의 대상이 되는 자료는 집단간의 차이를 식별하는데 사용되는 판별변수(discriminant variables)와 관측치가 속한 사전에 정의된 집단을 나타내는 집단변수(group variable)로 구성되는 다변량 자료이다. 기존의 판별분류분석에서는 주로 분석의 대상이 되는 자료가 독립적으로 얻어진다는 가정하에 논의가 전개되어왔다(Johnson & Wichern(1992), Dillon & Goldstein(1984), 김기영과 전명식(1989) 참조).

본 논문에서는 판별변수가 추출된 기저 모집단에 대한 정보를 제공하는 집단변수가 이산시계열(discrete time series)모형을 따른다고 가정하기로 한다. $\{X_t : t = 1, \cdots, n\}$를 시점 t의 관측치에 대한 집단변수라 하고 연관된 판별변수 벡터를 $y_t : p \times 1$ 벡터라고 하자. 분석대상이 되는 자료는 $\{(X_t, y_t), t = 1, \cdots, n\}$로 구성되어 있으며 이 자료를 이용하여 합리적인 판별기준을 마련하고 이 기준을 통해서 시점 $n+1$에서 y_{n+1}이 관측되는 경우, 연관된 X_{n+1}을 예측(판별)하고자 한다.

2. 집단변수 구조의 파악

집단변수 { $X_t = x_t : X_t =$ "0"(실패), 또는 "1"(성공) }가 이진확률변수로서 다음과 같은 m-차 마코프 종속구조를 가진다고 가정하자.

$$P[X_t = x_t | past] = P[X_t = x_t | X_{t-1}, \cdots, X_{t-m}] \tag{2.1}$$

*본 연구는 1996년도 숙명여자대학교 특별연구비 지원에 의해 수행되었음

† (140-742) 서울시 용산구 청파동, 숙명여자대학교 통계학과 부교수

‡ (150-010) 서울시 영등포구 여의도동, 한국쎄스소프트웨어주식회사

조건부 성공확률을 $p_{(t)}$, 조건부 실패확률을 $q_{(t)}(p_{(t)}+q_{(t)}=1)$이라 하자. 즉,

$$p_{(t)} = P[X_t = 1|X_{t-1}, \cdots, X_{t-m}]. \tag{2.2}$$

이제, 집단변수$\{X\}$에 다음과 같은 조건부 자기로지스틱(conditional autologistic) 모형을 고려하기로 한다.

$$LG(p_{(t)}) = \beta_0 + \beta_1 S_{t-1} \tag{2.3}$$

여기서 $S_{t-1} = X_{t-1} + \cdots + X_{t-m}$로서 시점 t에서 이전 m시점까지의 총수를 표시하며, 따라서 S_{t-1}은 $0, 1, \cdots, m$의 $m+1$개의 가능한 값을 가지며 $LG(\cdot)$는 표준로짓함수이다. 즉,

$$LG(p) = \ln[p/(1-p)], \quad 0 < p < 1$$

이며, 윗 식 (2.2)와 (2.3)으로부터 다음 식을 쉽게 얻을 수 있다.

$$p_{(t)} = \exp[\beta_0 + \beta_1 S_{t-1}]/[1 + \exp\{\beta_0 + \beta_1 S_{t-1}\}]. \tag{2.4}$$

조건부 자기 로지스틱모형을 이용해서 $\{X_t\}$의 종속구조를 모형화 하는 장점은 다음과 같다. 첫째, 모수의 절약측면에서 식 (2.3)은 두 개의 모수 β_0, β_1만은 가지나 일반적으로 m차 마코프 모형에서는 많은 수의 모수를 필요로 한다. 예를 들어 $m=2$인 경우 상태공간의 확장을 통해 $2^2 = 4$개의 모수가 필요하며 (c.f. Ross(1972), p.86) 일반적으로는 2^m개의 모수가 필요하게 된다. 둘째는 모형의 β_1이 군집의 존재성과 강도를 측정해 줄 수 있다. 모형 (2.4)에서 $\beta_1 = 0$이면 $\{X_t\}$는 성공확률이 $e^{\beta^0}/(1+e^{\beta^0})$ 인 n개의 독립적인 베르누이 변수들이며, $\beta_1 > 0$ 일때는 S_{t-1} = "과거 m개의 X값의 합"이 주어진 조건하에서 $X_t = 1$의 확률이 S_{t-1}을 따라서 증가한다. 결국 β_1은 X_t와 S_{t-1}의 상관계수로서 자기회귀계수와 유사한 역할을 하고 있으며 $\beta_1 > 0$은 "성공"(1)들의 군집(cluster) 존재성을 의미하고 있다. 더욱이 Cressie(1993, p.424)가 지적한대로 조건부 자기로지스틱 모형은 모형에 대한 가정이라기보다는 적절한 조건하에서 이항 종속 자료의 결과된 모형식(consequence of the algebra)으로 해석될 수 있으므로 식 (2.4)의 조건부 자기로지스틱모형은 실용적으로 타당성을 가지며 식의 의미와 해석은 Pickard(1987), Basawa(1996) 그리고 Cressie(1993)을 참조하기 바란다.

관측치 $X_1, \cdots, X_n$에 근거한 우도함수 $L_n(\beta)$는 다음과 같다.

$$L_n(\beta) = \prod_{t=m+1}^{n} p_{(t)}^{X_t} q_{(t)}^{1-X_t}, \quad \beta = (\beta_0, \beta_1)' \tag{2.5}$$

대수우도함수(log-likelihood function)

$$\ln(\beta) = \sum [\beta_0 X_t + \beta_1 X_t S_{t-1} + \ln q_{(t)}]$$

로부터 스코어(일차미분값) $J_n(\beta)$와 헤시안(부호를 바꾼 이차미분값) $H_n(\beta)$를 구한 결과는 다음과 같다.

$$J_n(\beta) = \begin{bmatrix} \sum(X_t - p_{(t)}) \\ \sum(X_t - p_{(t)}) S_{t-1} \end{bmatrix}, \tag{2.6}$$

$$H_n(\beta) = \begin{bmatrix} \sum p_{(t)}q_{(t)}, & \sum p_{(t)}q_{(t)}S_{t-1} \\ \sum p_{(t)}q_{(t)}S_{t-1}, & \sum p_{(t)}q_{(t)}S_{t-1}^2 \end{bmatrix}. \tag{2.7}$$

이제 β의 최대우도추정량(MLE) $\hat{\beta}_{ML}$은 식 (2.6)의 스코어를 0으로 놓고서 얻을 수 있으며, 즉, 추정방정식은 $J_n(\hat{\beta}_{ML}) = 0$이다.
또한 헤시안 $H_n(\beta)$는 다음과 같이 조건부 분산의 합으로 표시될 수 있다.

$$H_n = \sum_{t=m+1}^{n} Var\left[\begin{pmatrix} X_t - p_{(t)} \\ [X_t - p_{(t)}]\, S_{t-1} \end{pmatrix} \middle| X_{t-1}, \cdots, X_{t-m} \right].$$

따라서 $H_n(\beta)$는 언제나 양정치 행렬이며 추정방정식은 유일한 해 (unique solution)를 제공한다.

3. 제안된 판별분류 기준

이 절에서는 자료 $\{(X_t, y_t), t = 1, \cdots, n\}$와 y_{n+1}이 주어진 경우 미지의 X_{n+1}을 예측(판별)하는 절차에 대해 살펴 보기로 한다. X_{n+1}을 판별하는 데는 판별변수 y와 함께 $X_n, \cdots, X_{n-m}$이 이용된다. 집단변수에 계열상관이 존재하는 경우나(예를 들어, X_t가 시점 t에서의 (이진의)경제현상을 나타내고 y_t가 연관된 판별변수를 표시하는 경우) 군집현상(1 또는 0) 이 존재하는 경우(예를 들면, X_t가 t지점의 광물 매장 여부를 그리고 y_t가 적절한 판별변수인 경우)에는 자료점간의 독립이 가정된 통상적인 판별분류모형 보다는 상관 구조를 반영한 제안된 모형이 타당할 것으로 판단된다. 집단변수X의 종속구조를 반영한 식 (2.1)과 (2.2)에 의해서 X_t의 사후확률(posterior probability)구조는 y_t뿐만 아니라 $X_{t-1}, \cdots, X_{t-m}$에 의해서도 영향을 받는다. $f_1(y)[f_0(y)]$를 $X_t = 1[X_t = 0]$에 대한 판별변수 벡터 y의 조건부 pdf라 할 때 관찰치 y_t와 $X_{t-1}, \cdots, X_{t-m}$이 주어진 경우 X_t가 성공$(X_t = 1)$일 사후확률은 다음과 같이 정리 할 수 있다.

$$\begin{aligned} & P(X_t = 1 | y_t, X_{t-1}, \cdots, X_{t-m}) \\ = \; & P_*(X_t = 1 | y_t) \\ = \; & P_*(y_t | X_t = 1) \cdot P_*(x_t = 1)/P_*(y_t) \\ = \; & f_1(y_t) P_*(X_t = 1)/P_*(y_t) \end{aligned} \tag{3.1}$$

여기서 P_*는 $X_{t-1}, \cdots, X_{t-m}$이 주어진 경우의 조건부 확률을 의미한다.
마찬가지로 다음 식이 성립한다.

$$\begin{aligned} & P(X_t = 0 | y_t, X_{t-1}, \cdots, X_{t-m}) \\ = \; & f_0(y_t) P_*(X_t = 0)/P_*(y_t) \end{aligned} \tag{3.2}$$

따라서 $X_t = 1$의 사후확률이 $X_t = 0$의 사후확률보다 큰 영역은 식 (3.1)과 (3.2)로부터

$$\frac{f_1(y_t)P(X_t = 1|X_{t-1}, \cdots, X_{t-m})}{f_0(y_t)P(X_t = 0|X_{t-1}, \cdots, X_{t-m})} \geq 1$$

$$\leftrightarrow \quad ln\left[\frac{P(X_t = 1|X_{t-1}, \cdots, X_{t-m})}{P(X_t = 0|X_{t-1}, \cdots, X_{t-m})}\right] \geq ln\left[\frac{f_0(y_t)}{f_1(y_t)}\right]$$

이며 식 (2.4)로부터 위의 영역은 다음과 같이 표현할 수 있다.

$$\leftrightarrow \beta_0 + \beta_1 S_{t-1} \geq ln\left[\frac{f_0(y_t)}{f_1(y_t)}\right] \tag{3.3}$$

따라서 제안된, 최대사후확률기준에 근거한 분류(예측)기준은 다음과 같다.

<M1>

만일 $\beta_0 + \beta_1 S_n \geq ln\left[\frac{f_0(y_{n+1})}{f_1(y_{n+1})}\right]$이 성립하면 미지의 x_{n+1}을
$x_{n+1} = 1$로 분류하고 그렇지 않으면 $x_{n+1} = 0$으로 분류한다.

4. 모의실험 및 해석

본 절에서는 SAS/IML을 사용하여, 모수 β_1을 변화시켜 가면서 이산시계열을 따르는 집단변수(X)와 3변량 정규분포를 따르는 판별변수(y)벡터를 201개 생성시켜, 200개의 자료를 가지고 추정된(제안된) 분류기준에 따라 $\hat{X}_{201}$를 판별하여 실제 X_{201}값과 비교하였다. 이때 $\hat{\beta}_{ML}$을 구하기 위해 식 (2.6)과 (2.7)을 이용한 뉴튼-랩슨 수치해 기법을 사용하였으며 $f_0(y)$와 $f_1(y)$는 각각 평균이 $(0,0,0)'$,$(1,1,1)'$ 그리고 분산-공분산행렬

$$\sum = \begin{pmatrix} 1.0 & 0.8 & 0.6 \\ 0.8 & 1.0 & 0.4 \\ 0.6 & 0.4 & 1.0 \end{pmatrix}$$

를 갖는 3변량 정규분포를 가정하였으며 각각의 평균은 표본평균 그리고 $\sum$는 합동공분산행렬(pooled sample variance)로 추정한 후 분류기준을 만들었다. 이러한 판별과정을 500회 반복(replication)하여 제안된 분류기준 ⟨M1⟩의 오분류율(PER : predition error rate)을 계산하였다.

$$PER = \text{잘못분류된횟수}/500 \tag{4.1}$$

일반적으로 판별분류분석에서는 분류를 잘못했을 때 발생하는 오분류비용(misclassification cost)을 고려하여 분류기준을 만들고 있으나 여기서는 실제 0인데 1로 분류하는 비용과 실제 1인데 0으로 오분류하는 비용이 동일하다고 가정하기로 한다. 이런 경우 기존의 최소 오분류비용기준 (minimum ECM rule)은 다음과 같다.

<M2>

x_{n+1}을 1로 분류하는 영역은 다음과 같다.

$$f_0(y_{n+1})/f_1(y_{n+1}) \leq (p_1/p_0)$$

여기서 p_0, p_1은 집단변수에 대한 사전확률(*prior probability*)로서 $p_0(p_1)$은 x값 중 0(1)의 비율로 추정하여 사용한다.

$\beta_1 = 0$인 경우는 집단변수(X)들이 독립이므로 기존의 ⟨M2⟩분류기준이 최적(optimal)임은 자명하지만(c.f. Johnson & Wichern(1992)) β_1이 0을 벗어날 때는 종속구조를 반영한 제안된 분류기준(M1)이 더 우수할 것으로 예상할 수 있다.

다음 표 4.1은 X가 1차 마코프 연쇄구조를 가질때 ($m = 1$) 제안된 기준 ⟨M1⟩과 기존의 기준 ⟨M2⟩를 비교한 표이다.

표 4.1: 오분류율(PER) 비교
M1 : 제안된 분류기준, M2 : 기존의 기준

$(1)\beta_0 = -1.0$

β_1	M1	M2	
-1.0	0.174	0.174	
-0.8	0.192	0.194	
-0.6	0.186	0.190	
-0.4	0.156	0.164	
-0.2	0.168	0.172	
0.0	0.214	0.212	(*)
0.2	0.180	0.180	
0.4	0.200	0.208	
0.6	0.228	0.232	
0.8	0.194	0.216	
1.0	0.212	0.250	

$(2)\beta_0 = 1.0$

β_1	M1	M2	
-1.0	0.236	0.246	
-0.8	0.236	0.246	
-0.6	0.244	0.262	
-0.4	0.208	0.214	
-0.2	0.244	0.238	(*)
0.0	0.198	0.196	(*)
0.2	0.196	0.198	
0.4	0.168	0.172	
0.6	0.142	0.144	
0.8	0.144	0.152	
1.0	0.108	0.110	

이 표에서 보듯이, 대체로 제안된 분류기준이 더 낮은 PER을 보임으로써 우수하다고 생각되며, $\beta_1 = 0$인 경우, 즉 집단변수가 독립일 때는 기존의 최소오분류비용기준이 더 우수(* 표시)함을 알 수 있다. 즉 β_1이 0인 경우 $\{X_t\}$가 서로 독립적인 베르누이 시행이므로 제안된 분류기준 ⟨M1⟩을 사용할 필요성을 느끼지 못하게 된다. 또한 β_1이 0에 가까운 경우에는 ⟨M1⟩에서 β를 추정하는데 수반되는 오차 때문에 기존의 ⟨M2⟩가 더 우수하게 나타났으며((2)에서 $\beta_1 = -0.2$인 경우) 이러한 현상은 β_1이 0에 가까울 때는 ⟨M1⟩이 ⟨M2⟩보다 우수한 면을 β의 추정으로 발생하는 오차가 희석시키기 때문으로 판단된다. 결론적으로 $\{X_t\}$의 종속성이 클 때에만 (β_1이 0에서 멀 때) 제안된 판별분류 기준이 더 효율적임을 알 수 있다.

5. 결언

지금까지 우리는 자료가 추출된 기저모집단에 대한 정보를 나타내는 집단변수 $\{X_t\}$가 이산시계열을 따르는 경우에 그 구조를 반영하는 모형을 설정하고, 이를 바탕으로 기저모집단의 이산시계열 구조를 반영하는 판별분류기준을 유도하였다. 제안된 판별분류기준의 효율성을 검토하기 위하여 모수를 변화시켜가며 제시한 모형을 따르는 자료를 생성시켜 오분류율(PER)을 계산하여 기존의 분류방법과 비교하였다. 모의실험결과 $\{X_t\}$의 자기상관성이 높을 때는 기존의 분류기준에 비하여 제안된 분류기준이 효율적인 분류기준이 됨을 알 수 있었다.

실제자료의 분석에 있어서 집단변수 $\{X_t\}$가 자기상관성을 갖는 것으로 간주할 수 있을 때 그 구조를 반영하는 판별분류기준을 사용하는 것이 타당하며, 따라서 상관성이 높을 때는 제안된 판별기준을 사용하는 것이 기존의 방법에 비해서 효율적인 분류방법이 될 것으로 기대할 수 있다.

참고문헌

[1] Basawa, I. V. (1996). *Inference for a class of causal spatial models.* Preprint, Univ. of Georgia.

[2] Cressie, N. A. C. (1993). *Statistics for spatial data.* Wiley, N. Y.

[3] Dillon, W. R. and Goldstein, M. G. (1984). *Multivariate Analysis : Method and Applications.* Wiley, N. Y.

[4] Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis, 3rd. ed.* Prentice Hall.

[5] Morrison, D. F. (1990). *Multivariate Statistical Methods, 3rd. ed.* McGrow-Hill.

[6] Pickard, D. K. (1987). Inference for Discrete Markov field : The Simplest Nontrivial Case. *Journal of American Statistical Association*, vol. 82.

[7] Ross, S. M. (1972). *Introduction to Probability Models.* Academic Press.

[8] SAS Institute (1990). *SAS/IML Software : Usage and Reference, version 6.*

[9] 김기영, 전명식 (1989). 〈SAS 판별 및 분류분석〉. 자유아카데미.

[1997년 8월 접수, 1997년 12월 최종수정]

Classification of a Binary Group Variable with Dependence Structure *

Sun Y. Hwang †, Eun Jung Na ‡

ABSTRACT

Most of the research on discrimination and classification analysis has been directeted to the situation where the data consist of independent observations. However, it is often the case in practice that a dependence structure between objects does exist, in particular, for the time series data. This article is handling such a case and is concerned with the problem of classifying new object when the dependence can be modelled by a discrete time series via conditional autologistic transition probability.

*This research was supported by 1996-Grants from Sookmyung Women's Univ.

† Associate Professor, Department of Statistics, Sookmyung Women's Univ.

‡ Researcher, SAS Software Korea Ltd., Seoul 150-010, Korea.