The Korean Journal of Applied Statistics (2018) DOI: http://dx.doi.org/10.5351/KJAS.2018.31.6.721
31(6), 721–732

Variable selection for latent class analysis using clustering efficiency

Seongkyung Kim[a] · Byungtae Seo[b,1]

[a]Begas; [b]Department of Statistics, Sungkyunkwan University

(Received August 8, 2018; Revised September 20, 2018; Accepted October 6, 2018)

Abstract

Latent class analysis (LCA) is an important tool to explore unseen latent groups in multivariate categorical data. In practice, it is important to select a suitable set of variables because the inclusion of too many variables in the model makes the model complicated and reduces the accuracy of the parameter estimates. Dean and Raftery (*Annals of the Institute of Statistical Mathematics*, **62**, 11–35, 2010) proposed a headlong search algorithm based on Bayesian information criteria values to choose meaningful variables for LCA. In this paper, we propose a new variable selection procedure for LCA by utilizing posterior probabilities obtained from each fitted model. We propose a new statistic to measure the adequacy of LCA and develop a variable selection procedure. The effectiveness of the proposed method is also presented through some numerical studies.

Keywords: latent class analysis, clustering, variable selection, multivariate categorical data

1. 서론

다변량 범주형 데이터에서 잠재집단을 찾아내고 군집화하기 위한 방법 중 잠재집단 모형 혹은 latent class analysis (LCA)는 통계적 모형에 기반한 군집분석 방법이라 할 수 있다. LCA는 일종의 다항 혼합모형(multinomial mixture model) (McLachlan과 Peel, 2000)이라 볼 수 있는데 특히 특정 관측 값은 오직 한 집단에 속한다는 결정론적 군집분석이 아니라 각 관측값은 일정한 확률을 가지고 각 집단에 속한다는 확률론적 군집분석 방법이라고 할 수 있다. 이러한 LCA는 사회과학분야에서 이미 그 필요성과 효용성이 알려져 널리 쓰여지고 있으며 특히 심리학 분야에서는 매우 중요한 계량적 연구방법론으로 자리 잡아 오고 있다.

LCA는 특히 조사 분석에서 매우 유용한 도구로 쓰여지고 있는데 예를 들어 설문조사 분석시 설문 응답자들이 하나의 균질한 모집단에 속한 것이 아니라 관측되지 않은 서로 다른 특성을 가지는 여러 모집단으로부터 나왔다고 보고 분석하는 것이다. 이때 각 잠재그룹은 각 응답문항에 어떻게 답변했는지로부터 구분 및 파악될 수 있는데 이러한 결과를 토대로 어떠한 잠재집단이 어떠한 비율로 모집단에 혼합되어 있는지와 각 잠재집단의 특성을 파악할 수 있다.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Techonology (NRF-2016R1A2B4007373).
[1]Corresponding author: Department of Statistics, Sungkyunkwan University, Sungkyunkwan-ro 25-2, Jongno-gu, Seoul 03063, Korea. E-mail: seobt@skku.edu

기존의 많은 연구에서는 주어진 모든 변수를 이용하여 LCA 모형을 적합시키는데 이때 주어진 변수가 너무 많을 경우에는 변수들 중 일부 변수는 그룹간의 특성을 그다지 잘 설명하지 못하는 경우가 있다. 이 경우 모든 변수를 분석에 사용한다면 분석의 효율성을 매우 떨어트릴 수 있고 또한 각 잠재집단의 특성을 파악하기 어려울 수 있다. 따라서 모든 변수를 이용하기보다는 몇개의 변수를 선택하여 분석을 할 필요가 있는데 Raftery와 Dean (2006)은 연속형 변수들에 대한 모형기반의 군집분석에서 회귀분석에서의 단계적 변수 선택법과 유사한 방법을 제시하였고 이후 Dean과 Raftery (2010)은 Headlong search 알고리즘을 통해 이를 LCA로 확장하였다. 이러한 방법들은 군집화에 유용하다고 판단된 변수로 이루어진 모형과 후보변수를 포함한 모형을 비교함으로써 후보변수의 중요성 Bayesian information criteria (BIC) 값을 바탕으로 판단하는 방법이다.

본 논문에서는 Headlong search 알고리즘을 대체할 수 있는 새로운 변수선택 방법으로 서로 다른 변수들을 포함하는 두 모형을 적합시킨 후 이로부터 각 잠재집단에 속할 사후확률을 계산하여 이를 바탕으로 변수를 선택하는 방법을 제안한다. 이 방법은 어떠한 변수가 LCA에서 필요한 변수라면 이 변수는 각 그룹을 결정짓는데 유용해야 한다는 원칙에 기반한다. 즉, 현재 선택된 변수들로 적합된 LCA 모형에 비해 만약 새로운 변수를 추가한 새로운 LCA 모형이 이전의 모형보다 각 그룹을 더 명확하게 구분지을 때 새 변수를 추가하고 그렇지 않다면 추가하지 않는 방법이다. 즉 그룹을 나누는데 있어서의 기여도를 바탕으로 변수를 선택하는 것이다.

본 논문의 구성은 다음과 같다. 2장에서는 잠재집단 모형 즉 LCA에 대해 소개하고 EM algorithm을 이용하여 모수를 추정하는 방법에 대해 설명할 것이다. 3장에서는 Dean과 Raftery (2010)이 제시한 Headlong search 알고리즘을 소개하고 4장에서는 적합된 모형을 통해 얻어진 사후확률을 통해 변수를 선택하는 새로운 방법론을 소개할 것이다. 5장에서는 모의실험을 통해 새로 제안된 방법론의 효율성을 기존 방법과 비교 연구하였으며 6장에서는 실증자료를 바탕으로 제안된 방법론을 적용시켜 보았다. 끝으로 6장에서는 본 논문의 간략한 요약과 시사점들을 제시한다.

2. 잠재집단 모형과 모수 추정

2.1. 잠재집단 모형

잠재집단 모형은 Lazarsfeld (1950a, 1950b)와 Lazarsfeld와 Henry (1968)에 의해 처음 제안된 모형으로 이 모형은 관찰변수와 잠재변수를 가지는데 이때 관찰변수는 직접 관측할 수 있는 변수이고, 잠재변수는 각 관찰값이 어느 집단에 속하는지를 나타내는 일종의 지시함수로서 직접 관찰이 불가능하다고 가정한다. 먼저 관찰 가능한 i번째 확률벡터를 $Y_i = (Y_{i1}, \ldots, Y_{iM})$라고 하자. 여기서 Y_{im}는 r_m개의 범주를 가지는 이산형 확률변수로 편의상 1부터 r_m까지의 정수값을 가진다고 가정하자. 또한 잠재변수 L_i는 관측되지 않은 이산형 확률변수로 C개의 범주를 가진다. L_i는 각 관찰벡터가 어느 집단에 속하는지를 나타내는 확률변수로 편의상 본 논문에서는 L_i는 1부터 C까지의 정수값을 가지는 확률변수라고 가정한다. 이때 Y_i의 확률 질량함수는 다음과 같이 표현될 수 있다.

$$P(Y_{i1} = y_{i1}, \ldots, Y_{im} = y_{iM}) = \sum_{l=1}^{C} P(Y_{i1} = y_{i1}, \ldots, Y_{iM} = y_{iM} | L_i = l) P(L_i = l). \tag{2.1}$$

보통의 잠재집단 모형은 집단 L_i가 주어졌을때 Y_{im}, $m = 1, \ldots, M$는 서로 독립이라고 가정하는데 이러한 독립성 가정하에서 식 (2.1)은 다시 다음과 같이 표현되어질 수 있다.

$$P(Y_{i1} = y_{i1}, \ldots, Y_{iM} = y_{iM}) = \sum_{l=1}^{C} \prod_{m=1}^{M} P(Y_{im} = y_{im} | L_i = l) P(L_i = l) = \sum_{l=1}^{c} \gamma_l \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(Y_{im}=k)}$$

여기서, $\gamma_l = P(L_i = l)$, $\rho_{mk|l} = P(Y_{im} = k|L_i = l)$이고 $I(Y_{im} = k)$은 $Y_{im} = k$일 경우 1 그렇지 않을 경우 0인 지시함수이다.

이러한 잠재집단 모형에서는 각 관찰벡터 Y_i가 주어졌을때 그 관찰값이 어느 집단에 속하는지를 다음의 사후확률을 통해 계산할 수 있다.

$$P(L_i = l|Y_i) = \frac{\gamma_l \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(Y_{im}=k)}}{\sum_{h=1}^{C} \gamma_h \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|h}^{I(Y_{im}=k)}}.$$

2.2. EM알고리즘을 이용한 모수추정

잠재집단 모형에서 최대 가능도 방법을 이용한 모수의 추정값은 보통 잠재변수 L_i를 결측값으로 보고 Expectation-Maximization (EM) 알고리즘 (Dempster 등, 1977)을 이용하여 얻어질 수 있다. 이를 설명하기 위해 먼저 Z_{il}을

$$Z_{il} = I(L_i = l) = \begin{cases} 1, & \text{if } L_i = l, \\ 0, & \text{otherwise} \end{cases}$$

라고 정의하면 Y_i와 Z_i의 결합 확률 질량함수는

$$P(Y_i, Z_i) = \prod_{l=1}^{C} \gamma_l \left\{ \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(Y_{im}=k)} \right\}^{Z_{il}}$$

로 표현될 수 있다. 여기서 $Y_i = (Y_{i1}, \ldots, Y_{iM})'$이고 $Z_i = (Z_{i1}, \ldots, Z_{iC})'$이다.

따라서 완전한 자료 (Y_i, Z_i), $i = 1, \ldots, n$이 주어졌을때의 우도함수는

$$L_c(\theta|Y, Z) = \prod_{i=1}^{n} \prod_{l=1}^{C} \left\{ \gamma_l \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(Y_{im}=k)} \right\}^{Z_{il}}$$

이고 로그우도함수는

$$l_c(\theta|Y, Z) = \sum_{i=1}^{n} \sum_{l=1}^{C} Z_{il} \log \left(\gamma_l \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(Y_{im}=k)} \right)$$

으로 주어진다. 이때 Z_{il}은 관찰되지 않은 변수이므로 다음의 EM 알고리즘을 이용하는데 먼저 γ_l^t와 $\rho_{mk|l}^t$을 t-번째 반복에서 얻어진 추정값이라고 하면 EM 알고리즘은 다음의 E-step과 M-step을 반복적으로 밟아가며 γ_l과 $\rho_{mk|l}$을 수렴할 때까지 갱신해 나간다.

E-step:

$$\begin{aligned} Q(\gamma_l, \rho_{mk|l}|\gamma_l^t, \rho_{mk|l}^t) &= \sum_{i=1}^{n} \sum_{l=1}^{C} E \left[Z_{il} \log \left\{ \gamma_l \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(Y_{im}=k)} \right\} \middle| Y_i, \gamma_l^t, \rho_{mk|l}^t \right] \\ &= \sum_{i=1}^{n} \sum_{l=1}^{C} \hat{Z}_{il} \log \left(\gamma_l \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(Y_{im}=k)} \right) \end{aligned}$$

단,

$$\hat{Z}_{il} = P\left(Z_i = l|Y_i, \gamma_l^t, \rho_{mk|l}^t\right) = \frac{\gamma_l^t \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{t\ I(Y_{im}=k)}}{\sum_{l=1}^{C} \gamma_l^t \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{t\ I(Y_{im}=k)}}$$

이다.

M-step:

$$\gamma_l^{t+1} = \frac{\sum_{i=1}^n \hat{Z}_{il}}{\sum_{i=1}^n \sum_{h=1}^C \hat{Z}_{ih}}, \qquad \rho_{mk|l}^{t+1} = \frac{\sum_{i=1}^n \hat{Z}_{il} I(Y_{im}=k)}{\sum_{i=1}^n \hat{Z}_{il}}.$$

3. Headlong search 알고리즘

3장에서는 LCA에서의 변수선택 방법의 하나로 Dean과 Raftery (2010)에 의해 소개된 Headlong search 알고리즘을 알아보도록 하겠다. Headlong search 알고리즘을 설명하기 위해 먼저 확률벡터 Y의 각 원소 $Y_1, \ldots, Y_M$을 다음의 세 가지 그룹으로 나눈다.

Y^{clust} : clustering에 유용하다고 판단되어 이미 선택 된 관찰변수들의 집합

Y^{candi} : Y^{clust}에 포함될지 아닐지 결정되어야 하는 후보변수

Y^{other} : Y^{clust}와 Y^{candi} 이외의 변수들의 집합으로 clustering에 유용하다고 판단되지 않는 변수들의 집합

이제 이 세가지 벡터에 대하여 잠재변수 L이 주어졌을 때의 Y의 조건부 분포는

$$\begin{aligned} P(Y|L) &= P\left(Y^{\text{clust}}, Y^{\text{candi}}, Y^{\text{other}}|L\right) \\ &= P\left(Y^{\text{other}}|Y^{\text{clust}}, Y^{\text{candi}}, L\right) P\left(Y^{\text{candi}}, Y^{\text{clust}}|L\right) \end{aligned} \tag{3.1}$$

으로 표현할 수 있는데 여기서 Y^{other}는 Y^{clust}와 Y^{candi}가 주어진 경우 clustering에 영향을 주지 않는 즉 L과 독립이라고 가정하면 (3.1)은 다시

$$\begin{aligned} P(Y|L) &= P\left(Y^{\text{other}}|Y^{\text{clust}}, Y^{\text{candi}}\right) P\left(Y^{\text{candi}}, Y^{\text{clust}}|L\right) \\ &= P\left(Y^{\text{other}}|Y^{\text{clust}}, Y^{\text{candi}}\right) P\left(Y^{\text{candi}}|L\right) P\left(Y^{\text{clust}}|L\right) \end{aligned} \tag{3.2}$$

로 나타낼 수 있다. 식 (3.2)의 두 번째 등식은 잠재집단 모형에서 각 집단이 주어진 경우 관찰벡터의 모든 성분은 서로 독립이라는 가정에 기인한다. 이제 후보변수 Y^{candi}의 유용성 여부를 판단하기 위해 다음 모형 M_1과 M_2를 고려해보자.

$$\begin{aligned} M_1 &: P(Y|L) = P\left(Y^{\text{other}}|Y^{\text{clust}}, Y^{\text{candi}}\right) P\left(Y^{\text{candi}}\right) P\left(Y^{\text{clust}}|L\right) \\ M_2 &: P(Y|L) = P\left(Y^{\text{other}}|Y^{\text{clust}}, Y^{\text{candi}}, L\right) P\left(Y^{\text{candi}}|L\right) P\left(Y^{\text{clust}}|L\right) \end{aligned}$$

모형 M_1은 Y^{candi}가 잠재변수 L과 독립임을 가정하는 모형이고 모형 M_2는 Y^{candi}와 Y^{clust}가 잠재변수 L이 주어졌을 때만 서로 독립인 모형이다. 즉, M_1은 Y^{candi}가 그룹에 영향을 주지 않는 모형이라고 볼 수 있고 M_2는 Y^{candi}가 그룹에 영향을 주는 모형이라고 볼 수 있다. 이 경우 변수선택 문제는 모형선택 문제로 생각할 수 있고 만약 모형 M_2가 모형 M_1보다 통계적 의미에서 더 적합도가 높다고 판단된다면 이는 변수 Y^{candi}가 집단을 결정짓는데 유용한 변수라고 할 수 있을 것이다. 여기서 적합도의 판정은 우도함수를 이용할 수 있는데 M_2의 우도함수 값이 항상 M_1에서의 우도함수 값보다 크므로 우도함수를 모형이 필요로 하는 모수의 개수를 벌점 함수화하여 수정한 BIC값을 토대로 적합도를 판단한다.

Dean과 Raftery (2010)은 이러한 아이디어를 바탕으로 반복적으로 변수를 선택하는 방법을 제시하고 이를 Headlong search 알고리즘이라고 명명하였다. 이 알고리즘은 Inclusion 단계와 Exclusion 단계를

Table 3.1. Headlong search algorithm

Inclusion 단계
(1) Y^{clust}만이 포함된 모형 M_1을 적합시켜 얻은 BIC값을 BIC(M_1)이라 하자.
(2) Y^{other}에 포함된 각 변수를 차례로 후보변수 Y^{candi}로 보고 $Y^{\text{clust}} \cup Y^{\text{candi}}$를 포함한 모형을 적합시켜 그중 가장 큰 BIC값을 주는 모형을 M_2라 하고 이때의 BIC값을 BIC(M_2)라 한다.
(3) BIC(M_2)−BIC(M_1)가 사전에 선택된 상한값 U 이상이면 BIC(M_2) 계산시 이용된 후보변수를 현재의 Y^{clust}에 추가하여 새로운 Y^{clust}를 만든다.
Exclusion 단계
(1) Y^{clust}만이 포함된 모형 M_1을 적합시켜 얻은 BIC값을 BIC(M_1)이라 하자.
(2) Y^{clust}에 포함된 변수를 하나씩 순차적으로 제거한 모형을 적합시켜 그중 가장 큰 BIC 값을 주는 모형을 M_2라 하고 이때의 BIC 값을 BIC(M_2)라 한다.
(3) BIC(M_2)−BIC(M_1)가 사전에 선택된 L 이상이면 BIC(M_2) 계산시 제거된 변수를 Y^{clust}에서 제거하여 새로운 Y^{clust}를 만든다.

거치며 유용한 변수를 BIC값을 바탕으로 선택 또는 제거해 나간다. 이 알고리즘에서는 먼저 모든 변수를 이용하여 잠재집단 모형을 적합시키고 잠재 집단간 변동이 가장 큰 변수들을 몇개 선택하여 이를 Y^{clust}라고 한다. 이때 초기 선택하는 변수의 개수 M은 모형의 identifiability를 만족하는 최소한의 개수 이상으로 정한다. 즉, C를 집단의 개수라고 할 때, 변수의 개수 M은 Goodman (1974)에 의해 제시된 identifiablity 조건

$$C\left[1+\sum_{m=1}^{M}(r_m-1)\right] \le \prod_{m=1}^{M} r_m$$

을 만족하는 가장 작은 값 혹은 그 이상으로 정한다.

Inclusion 단계는 일종의 전진선택법 그리고 Exclusion 단계는 후진 제거법을 의미하는데 이 두 단계를 새로 선택되거나 제거하는 변수가 없을 때까지 반복하는 Headlong search 알고리즘은 단계적 변수 선택방법과 유사한 방법으로 Table 3.1에 Inclusion 단계와 Exclusion 단계를 요약하였다.

4. 제안 알고리즘

Headlong search 알고리즘은 기본적으로 모형에 포함된 변수가 서로 다른 두 모형에 대하여 우도함수 (혹은 BIC)를 이용하여 그 적합성을 바탕으로 변수를 선택하는 방법이다. 하지만 군집분석에서 유용한 변수의 선택은 그 변수가 집단을 구분 짓는데 있어서 얼마나 기여를 하는가에 따라 결정하는 것이 보다 직관적인 방법이 될 수 있을 것이다. 이러한 생각을 바탕으로 본 논문에서는 먼저 선택된 두 개의 후보모형을 적합시키고 적합한 모형을 바탕으로 모든 관찰값들에 대하여 각각의 잠재집단에 포함될 확률을 계산한다. 모형기반의 군집분석의 특성상 각 개체가 어느 잠재집단에 속하는지는 결정론적으로 정하지 않고 확률론적으로 표현할 수 있는데 만약 모형이 잠재집단을 잘 구분 지을 수 있다면 각 개체가 어느 한 집단에 속할 확률이 다른 집단에 속할 확률에 비해 아주 높아야 할 것이다.

예를 들어 집단이 2개인 LCA 모형을 적합하고 이를 바탕으로 i번째 관찰값이 첫 번째와 두 번째 집단에 속할 사후확률을 각각 $P(L_i = 1|Y_i)$과 $P(L_i = 2|Y_i)$라고 하자. 이때 만약 $P(L_i = 1|Y_i) = P(L_i = 2|Y_i) = 0.5$라면 이는 i번째 관찰값에 대하여서는 현재의 모형이 집단을 구분짓는데 유용하지 않다고 판단할 수 있을 것이다. 반면에 만약 $P(L_i = 1|Y_i) = 0.9$이고 $P(L_i = 2|Y_i) = 0.1$이라면 이는 현재의 모형이 i번째 관찰값을 잘 구분 짓는 모형이라고 할수 있을 것이다. 다시 말해서 각 관찰값의 사후 확률이 0.5에서 멀어질 경우 가정한 모형이 집단을 결정짓는데 유용한 모형이라고 할 수 있을 것이다. 이 경우

Table 4.1. Proposed algorithm

Inclusion 단계
(1) Y^{clust}를 이용하여 잠재집단 모형 적합 후 식 (4.1)을 이용하여 $V(Y^{\text{clust}})$를 계산한다
(2) Y^{other}에 포함된 각 변수를 차례로 후보변수 Y^{candi}로 보고 잠재집단 모형을 적합시켜 추가할 변수 $Y^* = \text{argmax}_{Y^{\text{candi}} \in Y^{\text{other}}} V(Y^{\text{clust}} \cup Y^{\text{candi}})$를 찾는다.
(3) 만약 $V(Y^{\text{clust}} \cup Y^*) - V(Y^{\text{clust}}) > U$이면 Y^{clust}에 Y^*를 추가하여 새로운 Y^{clust}를 생성한다.
Exclusion 단계
(1) Y^{clust}를 이용하여 잠재집단 모형 적합후 식 (4.1)를 이용하여 $V(Y^{\text{clust}})$를 계산한다.
(2) Y^{clust}에 포함된 변수를 하나씩 순차적으로 제거한 모형을 적합시켜 제거할 변수 $Y^* = \text{argmax}_{Y^r \in Y^{\text{candi}}} V(Y^{\text{clust}}/Y^r)$를 찾는다.
(3) 만약 $V(Y^{\text{clust}}/Y^*) - V(Y^{\text{clust}}) > L$이면 Y^{clust}에 Y^*를 제거하여 새로운 Y^{clust}를 생성한다.

Table 5.1. Simulation models

	Class	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}
Model 1	1	0.2	0.5	0.7	0.4	0.9	0.1	0.6	0.2	0.8	0.7
	2	0.2	0.4	0.8	0.2	0.4	0.7	0.1	0.7	0.8	0.7
Model 2	1	0.2	0.5	0.7	0.4	0.9	0.1	0.6	0.2	0.8	0.7
	2	0.2	0.4	0.8	0.2	0.6	0.5	0.2	0.6	0.8	0.7

각 관찰값이 각 잠재집단에 속할 사후확률의 차가 클수록 집단을 구분 짓는데 더 유용한 모형 혹은 변수라고 할 수 있을 것이다. Kim (2013)은 이와 유사한 아이디어를 가지고 잠재 집단이 두개일 때 각 관찰값에 대한 사후확률의 차이를 이용하여 변수를 선택을 하는 방법을 제시했으나 이는 잠재집단이 두개이고 또한 두 잠재집단의 크기가 같은 경우에만 사용이 가능한 방법으로 현실적인 사용에는 큰 제약이 있었다.

본 논문에서는 여러 개의 잠재집단이 임의의 비율로 혼합되어 있을 때에도 사용 가능한 방법을 제안하는데 이를 위해 먼저 다음의 통계량을 변수 혹은 모형의 유용성을 판단하기 위해 사용할 것을 제안한다.

$$V(Y) = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{C}\gamma_l(P(L_i = l|Y_i) - \mu_i)^2, \tag{4.1}$$

여기서 μ_i는 잠재 집단간 사후확률의 가중평균 즉 $\mu_i = \sum_{l=1}^{C}\gamma_l P(L_i = l|Y_i)$이다. 식 (4.1)에서 V는 사후확률의 잠재집단별 가중분산의 평균을 나타내는 통계량으로 가중분산을 사용한 이유는 각 그룹의 혼합비율이 다르기 때문에 이를 반영하기 위함이다. 따라서, 이 통계량은 각 관찰치의 사후확률이 아무런 정보가 없을 때의 평균값인 μ_i로부터 얼마나 멀리 떨어졌는지를 나타내는 통계량이라고 할 수 있다. 이제 통계량 V를 이용한 제안 알고리즘은 Headlong search 알고리즘의 단계적 선택방법과 유서사하게 구성할 수 있는데 이를 Table 4.1에 정리하였다. 단, Table 4.1에서 $A/B = A \cap B^c$을 의미한다.

5. 모의실험

5장에서는 모의실험을 통해 Headlong search 알고리즘과 제안한 알고리즘의 변수선택의 효율을 비교하고자 한다. 모의실험을 위해서 10개의 이진변수를 가지는 관찰벡터 $Y = (Y_1, \ldots, Y_{10})$를 두개의 집단(Class = 1, 2)으로부터 확률 γ_1과 γ_2의 비율로 생성하였다. 이때 첫 번째 집단과 두 번째 집단에서 각 Y_m, $m = 1, \ldots, 10$이 1의 값을 가질 확률은 Table 5.1과 같이 두 가지 모형을 고려하였다. Table 5.1에서 ρ_m은 Y_m이 1의 값을 가질 확률을 의미한다. 고려된 두 모형 모두에서 Y_5, Y_6, Y_7, Y_8은 class별

Table 5.2. The number of selected variables for Model 1

n	(γ_1, γ_2)	Method	α	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
1000	(0.5, 0.5)	제안	0.005	1	29	42	89	99	100	99	100	1	2
			0.01	0	3	8	86	100	100	100	100	0	0
			0.03	0	0	0	2	100	100	100	100	0	0
		Headlong		0	0	1	99	81	100	76	43	0	0
	(0.8, 0.2)	제안	0.005	10	25	27	50	94	99	90	93	18	13
			0.01	7	10	13	40	95	99	95	96	9	5
			0.03	0	1	1	12	99	100	97	97	3	0
		Headlong		0	0	8	87	83	99	50	63	0	0
2000	(0.5, 0.5)	제안	0.005	0	26	27	99	100	100	100	100	0	0
			0.01	0	2	1	88	100	100	100	100	0	0
			0.03	0	0	0	1	100	100	100	100	0	0
		Headlong		0	0	0	100	80	100	85	35	0	0
	(0.8, 0.2)	제안	0.005	1	27	27	62	98	100	98	100	2	5
			0.01	2	14	12	48	98	100	98	100	0	2
			0.03	0	0	0	4	100	100	99	100	0	0
		Headlong		0	1	0	99	91	100	48	61	0	0

Table 5.3. The number of selected variables for Model 2

n	(γ_1, γ_2)	Method	α	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
1000	(0.5, 0.5)	제안	0.005	12	52	61	96	96	99	99	99	10	15
			0.01	3	28	42	93	96	99	99	99	2	4
			0.03	0	5	9	56	97	99	99	100	0	0
		Headlong		0	0	5	88	48	92	87	80	0	0
	(0.8, 0.2)	제안	0.005	19	39	43	69	88	94	89	88	20	20
			0.01	12	30	35	60	91	94	91	90	8	14
			0.03	0	14	7	39	95	98	97	94	1	6
		Headlong		0	3	9	59	77	95	78	79	0	0
2000	(0.5, 0.5)	제안	0.005	3	65	69	98	100	100	99	100	1	4
			0.01	1	36	35	97	100	100	99	100	0	0
			0.03	0	1	1	79	100	100	100	100	0	0
		Headlong		0	1	1	96	34	97	89	82	0	0
	(0.8, 0.2)	제안	0.005	13	48	38	75	96	99	94	99	7	11
			0.01	4	21	24	70	97	99	95	99	1	4
			0.03	1	2	1	34	100	99	99	100	1	4
		Headlong		0	4	2	88	56	99	68	83	0	0

로 큰 차이를 주는 변수 즉 잠재집단을 구분 짓는데 있어서 유용한 변수이고 Y_1, Y_9, Y_{10}은 잠재집단별 차이가 없는 변수이고 Y_2, Y_3, Y_4는 잠재집단별로 약간의 차이가 있는 변수이다. 또한 Model 1은 Y_5, Y_6, Y_7, Y_8이 1이 될 확률이 잠재집단에 따라 큰 차이를 보이는 모형이고 Model 2는 Model 1보다는 작은 차이를 보이는 모형이라고 할 수 있는데 보통의 혼합모형의 특성상 Model 1은 안정적인 추정이 가능하고 Model 2에서의 추정은 특히 관찰개수가 적을 경우 다소 불안정적일 수 있다.

모의실험에서는 혼합비율 (γ_1, γ_2)가 (0.5, 0.5)인 경우와 (0.8, 0.2)인 경우를 자료개수 $n = 1000, 2000$인 경우에 대하여 100번 반복하였다. Tables 5.2와 5.3는 각 변수별로 100번의 반복실험에서 선택된 횟수를 나타낸 표이다. Headlong search 알고리즘과 제안 알고리즘 모두 처음 변수로 4개를 선

Table 6.1. Questionnaire

변수	질문내용
Y_1	학교 수업 시간이 재미있나요?
Y_2	학교 숙제를 빠뜨리지 않고 잘 하나요?
Y_3	수업 시간에 배운 내용을 잘 알고 있나요?
Y_4	모르는 것이 있을 때 다른 사람(부모님이나 선생님 또는 친구들)에게 물어보나요?
Y_5	공부 시간에 딴 짓을 하나요?
Y_6	당번이나 1인 1역 등 반에서 맡은 활동을 열심히 하나요?
Y_7	복도와 계단을 다닐 때 뛰지 않고 조용히 다니나요?
Y_8	학교 물건을 내 것처럼 소중히 사용하나요?
Y_9	화장실이나 급식실 등에서 차례를 잘 지키나요?
Y_{10}	휴지나 쓰레기를 버릴 때 꼭 휴지통에 버리나요?
Y_{11}	반 아이들과 잘 어울리나요?
Y_{12}	친구와 다투었을 때 먼저 사과하나요?
Y_{13}	내 짝이 교과서나 준비물을 안 가져왔을 때 함께 보거나 빌려 주나요?
Y_{14}	친구가 놀이를 하거나 공부를 할 때 방해하나요?
Y_{15}	놀이나 모둠활동을 할 때 친구들이 내 말에 잘 따라 주나요?
Y_{16}	선생님을 만나면 반갑게 인사하나요?
Y_{17}	선생님과 이야기하는 것이 편한가요?
Y_{18}	학교 밖에서 선생님을 만나면 반가운가요?
Y_{19}	우리 선생님께서는 나에게 친절하신가요?
Y_{20}	내년에도 지금 선생님께서 담임 선생님을 해 주신다면 어떤 기분이 들까요?

택하였으며 Headlong search 알고리즘에서는 $U = L = 0$을 사용하였고 제안 알고리즘에서는 $\alpha = 0.005, 0.01, 0.03$에 대하여 $U = L = \alpha V(Y^{\text{clust}})$를 사용하였다.

두 모형 모두 제안 알고리즘의 경우 모든 α에 대하여 선택되어야 하는 변수 Y_5, Y_6, Y_7, Y_8을 Headlong search 알고리즘에 비해 잘 선택하는 것을 알 수 있다. Headlong search 알고리즘은 전반적으로 잠재집단별 비교적 작은 차이를 보이는 변수 Y_2, Y_3, Y_4를 n이 커져도 잘 선택하지 못하는 경향이 있다. 제안한 알고리즘의 경우 Y_2, Y_3, Y_4의 선택빈도는 α값에 따라 달라지는데 α값이 작아질수록 경우 비교적 많은 변수를 선택한다. 또한 선택되지 않아야 할 변수인 Y_1, Y_9, Y_{10}은 Headlong search 알고리즘은 전혀 선택하지 않았고 제안 알고리즘에서는 $\alpha = 0.005$인 경우를 제외하고는 약간 선택되었다. 하지만 선택빈도는 n이 커짐에 따라 줄어드는 것을 알 수 있다.

6. 사례연구

이 장에서는 Headlong search 알고리즘과 제안 알고리즘을 이용하여 잠재집단 모형에서 변수선택을 해보았다. 사용된 자료는 한국청소년정책연구원의 한국 아동 청소년 패널조사 중 2012년 초등학교 3학년 학생들을 대상으로 한 설문조사 결과이다. 분석에 사용된 총 20개의 설문문항은 Table 6.1에 정리되어 있는데 설문문항은 크게 네 가지 범주로 나누어 볼 수 있다. 먼저 Y_1–Y_5는 주로 학습활동에 대한 문항이고 Y_6–Y_{10}은 학교규칙, Y_{11}–Y_{15}은 교우관계, Y_{16}–Y_{20}은 선생님과의 관계에 대한 문항이다. 설문 참여자는 총 2,342개이고 이중 결측값이 있는 자료를 제외하고 2,172명의 응답내용을 이용하여 분석을 진행하였다. 설문 문항에는 성별이나 부모님 학력 등의 자료가 포함되어 있으나 본 연구에서는 이러한 자료는 제외하고 Table 6.1에 있는 20개의 문항에 대한 응답 결과만을 이용하였다.

각각의 설문항목에 대하여 응답자는 "1.매우 그렇다", "2.그런 편이다", "3.그렇지 않은 편이다", "4.전혀 그렇지 않다"의 4개의 범주중 하나를 선택하였는데 Sung 등 (2016)에서는 같은 자료를 이용하여 초

Table 6.2. Estimated class probability

Method	Class 1	Class 2	Class 3	Class 4
Headlong	0.4758	0.2740	0.0715	0.1787
제안($\alpha = 0.01$)	0.3788	0.4109	0.0913	0.1190
제안($\alpha = 0.03$)	0.4803	0.3890	0.0128	0.1179

Table 6.3. Response probabilities for selected models

변수	Class	Headlong				제안($\alpha = 0.01$)				제안($\alpha = 0.03$)			
		답변문항				답변문항				답변문항			
		1	2	3	4	1	2	3	4	1	2	3	4
Y_1	1					0.57	0.43	0.00	0.00	0.56	0.39	0.02	0.01
	2					0.11	0.80	0.08	0.01	0.12	0.82	0.06	0.01
	3					0.06	0.43	0.31	0.21	0.13	0.18	0.13	0.56
	4					0.56	0.33	0.09	0.02	0.06	0.52	0.32	0.10
Y_8	1					0.66	0.33	0.01	0.00	0.71	0.27	0.01	0.00
	2					0.20	0.76	0.04	0.00	0.20	0.77	0.03	0.00
	3					0.19	0.56	0.18	0.07	0.34	0.23	0.00	0.43
	4					0.85	0.13	0.00	0.02	0.20	0.62	0.19	0.00
Y_{12}	1					0.19	0.70	0.09	0.02				
	2					0.07	0.65	0.26	0.02				
	3					0.10	0.26	0.35	0.28				
	4					0.77	0.10	0.09	0.04				
Y_{16}	1	0.71	0.27	0.02	0.00	0.68	0.31	0.00	0.00	0.76	0.23	0.01	0.00
	2	0.24	0.69	0.06	0.00	0.21	0.72	0.07	0.00	0.22	0.73	0.05	0.00
	3	0.20	0.35	0.37	0.08	0.14	0.40	0.39	0.06	0.17	0.24	0.34	0.25
	4	0.31	0.59	0.09	0.01	0.97	0.00	0.02	0.01	0.17	0.46	0.35	0.03
Y_{18}	1	0.98	0.02	0.00	0.00	0.97	0.08	0.00	0.00	0.91	0.09	0.00	0.00
	2	0.31	0.67	0.02	0.00	0.41	0.58	0.01	0.00	0.42	0.57	0.00	0.00
	3	0.06	0.20	0.50	0.24	0.04	0.37	0.40	0.18	0.00	0.13	0.16	0.71
	4	0.34	0.66	0.00	0.00	0.82	0.17	0.01	0.00	0.11	0.49	0.33	0.07
Y_{19}	1	0.81	0.19	0.00	0.00	0.82	0.17	0.00	0.01	0.80	0.19	0.01	0.01
	2	0.50	0.49	0.01	0.01	0.30	0.67	0.03	0.00	0.34	0.66	0.00	0.00
	3	0.02	0.37	0.35	0.26	0.08	0.44	0.29	0.19	0.11	0.14	0.06	0.69
	4	0.00	0.91	0.09	0.00	0.68	0.29	0.01	0.01	0.06	0.54	0.32	0.08
Y_{20}	1	0.89	0.10	0.01	0.00								
	2	0.64	0.31	0.05	0.00								
	3	0.05	0.20	0.39	0.36								
	4	0.00	0.69	0.27	0.04								

등학교 3학년 아동의 학교적응 유형을 BIC등의 값을 통해 4개의 class를 가지는 잠재집단 모형을 선택하였다. 본 논문에서도 4개의 class를 가정하고 Headlong search 알고리즘과 제안 논문을 통해 변수를 선택해 보았다.

Headlong search 알고리즘을 이용하였을 때 선택된 변수는 $Y_{16}, Y_{18}, Y_{19}, Y_{20}$이고 이 변수들은 주로 선생님과의 관계를 나타내는 변수이다. 제안된 알고리즘을 이용했을 때 선택된 변수는 $\alpha = 0.03$일 경우는 $Y_1, Y_8, Y_{16}, Y_{18}, Y_{19}$으로 학습활동과 학교규칙에 대한 질문도 포함된 것을 알 수 있다. 또한 α값이 0.01과 0.005인 경우는 $Y_1, Y_8, Y_{12}, Y_{16}, Y_{18}, Y_{19}$이 선택되었고 이들은 네 개의 질문범주 안에 들어가는

변수들을 하나 이상씩 포함한 것을 알 수 있는데 이들 변수를 이용하여 적합한 결과를 Table 6.3에 나타내었다. 모든 모형에서 Class 1은 학교적응 우수 집단으로 파악할 수 있으나 그 외의 Class는 선택된 변수에 따라 다소 다른 의미를 가진다고 할 수 있다. 예를 들어 $\alpha = 0.01$을 이용한 제안 알고리즘을 통해 선택된 변수를 사용한 잠재집단모형에서는 Class 4는 Class 1과 유사한 패턴을 보여주지만 12번 문항에서 class 1과 극명한 차이를 보이는데 Class 4에 속하는 학생은 Class 1에 속한 학생들에 비해 친구와 다툰 후 보다 적극적으로 화해를 시도한다고 볼 수 있다. 하지만 Headlong search 알고리즘으로부터 선택된 변수를 이용하여 잠재집단 모형을 적합시켰을 때의 Class 4는 Class 2와 유사한 집단으로 Y_{19}와 Y_{20} 설문문항에 대해 Class 2보다는 다소 약한 긍정을 보여준 집단이라고 할 수 있다.

7. 결론

본 논문에서는 잠재집단 모형에서의 변수선택 방법으로 기존의 Headlong search 알고리즘을 살펴보고 이를 대체할 수 있는 새로운 알고리즘을 제시하였다. 두 알고리즘을 모의실험과 실증자료 분석을 통해 비교하였는데 두 방법은 서로 다른 기준에 따라 변수를 선택하고 또한 이를 조율하는 일종의 조율모수가 있기 때문에 완전히 객관적인 비교는 불가능하지만 대체적으로 두 방법 모두 매우 유의한 변수는 잘 선택하는 것을 알 수 있다. 그 외 약간 유의한 변수 선택에 있어서는 제안된 알고리즘은 Headlong search 알고리즘보다 대체적으로 더 우수한 성능을 보여주는 것을 알 수 있다. 이러한 성능 차이는 사용하는 방법의 차이에서 기인하기도 하지만 조율모수의 선택에 따라 달라질 수 있는데 연구자가 연구목적이나 자료의 형태에 따라 적절한 조율모수를 선택함으로써 적당한 크기의 변수집단을 선택할 수 있을 것이다. 하지만 보다 일반적이고 객관적인 조율모수의 선택에 대하여서는 향후 더 깊은 연구가 필요할 것이다.

Headlong search 알고리즘은 근본적으로 우도함수를 이용하는 관계로 보통의 잠재집단 모형이 가지는 모형적 가정에 강하게 의존하므로 이러한 가정이 틀릴 경우, 예를 들면 잠재집단이 주어졌을 때의 독립성 등이 만족하지 않을 경우 효과적인 변수선택을 하기 어려울 것이다. 반면에 제안된 방법은 우도함수를 통한 모형 적합도의 의존하지 않고 순수하게 적합된 모형의 군집효율만을 바탕으로 변수선택을 하는 방법으로 모형의 가정에 덜 만감할 것으로 판단된다. 또한 제안한 방법에서 사용된 V 통계량은 결정론적 분류 효율을 이용한 것이 아니라 확률론적 효율을 이용함으로써 모형기반 군집분석에 보다 적합한 방법이 될 수 있고 향후 이를 이용 또는 응용한 방법을 통해 변수 선택뿐 아니라 군집의 개수 추정에도 이용될 수 있을 것으로 기대된다.

References

Dean, N. and Raftery, A. E. (2010). Latent class analysis variable selection, *Annals of the Institute of Statistical Mathematics*, **62**, 11–35.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B*, **39**, 1–38.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika*, **61**, 215–231.

Kim, S. (2013). *Variable selection in latent class analysis* (Master thesis), Sungkyunkwan University, Seoul.

Lazarsfeld, P. F. (1950a). The logical and mathematical foundations of latent structure analysis. In S. A. Stouffer (Ed.), *Measurement and prediction, the American soldier: studies in social psychology in World War II* (Vol. IV, Chap. 10, pp. 362–412). Princeton, Princeton University Press, NJ.

Lazarsfeld, P. F. (1950b). The interpretation and computation of some latent structures. In S. A. Stouffer (Ed.), *Measurement and prediction, the American soldier: studies in social psychology in World War II* (Vol. IV, Chap. 11, pp. 413-472). Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*, Houghton Mifflin, Boston.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*, Wiley, New York.

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering, *Journal of the American Statistical Association*, **101**, 168–178.

Sung, M., Chang, Y. E., and Seo, B. (2016). The roles of study habits and emotional behavioral problems in predicting school adjustment classification among 3rd graders, *Korean Journal of Childcare & Education*, **12**, 79–102.

잠재변수 모형에서의 군집효율을 이용한 변수선택

김성경[a] · 서병태[b,1]

[a]베가스, [b]성균관대학교 통계학과

(2018년 8월 8일 접수, 2018년 9월 20일 수정, 2018년 10월 6일 채택)

요 약

잠재집단 모형은 다변량 범주형 자료 안에 숨겨진 집단을 찾는 매우 중요한 도구종의 하나이다. 하지만 실제 자료 분석에서 너무 많은 관찰변수들을 포함시킨 모형은 모형을 복잡하게 만들고 또한 모수추정의 정확도에 영향을 주기 때문에 정보가 손실되지 않는 내에서 유용한 변수를 찾는 것은 중요한 문제이다. Dean과 Raftery (2010)은 잠재집단 모형에서의 변수선택을 위해 BIC를 이용한 Headlong search 알고리즘을 제시하였는데 본 논문에서는 이 방법을 대체할 수 있는 방법으로 적합한 모형으로부터 계산된 잠재집단에 속할 사후확률을 이용하여 변수 선택을 하는 방법을 제안하고자 한다. 이를 위하여 잠재집단 모형의 적합성을 측정할 수 있는 새로운 통계량과 이를 이용한 변수선택 알고리즘을 제시할 것이다. 또한 제안된 방법의 효율성을 모의실험과 실증자료 분석을 통해 살펴보고자 한다.

주요용어: 잠재집단모형, 군집분, 변수선택, 다변량 범주형 자료

[1]교신저자: (03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: seobt@skku.edu