Journal of the Korean Data & Information Science Society
2021, **32**(6), 1329–1341

http://dx.doi.org/10.7465/jkdi.2021.32.6.1329
한국데이터정보과학회지

단순 베이즈 분류에서 FDR 기반의 범주형 변수의 선택[†]

신지은[1] · 박창이[2]

[12]서울시립대학교 통계학과

접수 2021년 8월 12일, 수정 2021년 10월 25일, 게재확정 2021년 10월 28일

요 약

단순 베이즈 분류는 반응변수가 주어졌을 때 설명변수들이 조건부 독립이라는 단순 베이즈 가정에 기반한다. 비록 단순 베이즈 가정은 다소 강한 가정이지만 단순 베이즈 분류기는 고차원 데이터에서 합리적인 성능과 계산상의 이점을 보이고 있다. 고차원 데이터에는 보통 많은 잡음 변수들이 있기 때문에 변수선택은 분류기의 예측의 정확도와 해석을 향상시킬 수 있다. 이 논문에서는 단순 베이즈 분류에서 FDR (false discovery rate) 조절에 기반한 범주형 변수선택법을 제안한다. 모의실험과 실제 데이터 분석을 통해 제안한 방법과 변화점 분석에 기반한 또 다른 변수선택법과 비교하며 제안한 방법이 특히 희박한 혹은 고차원 데이터에 대하여 더 효율적임을 보인다.

주요용어: 고차원 데이터, 단순 베이즈 가정, 카이제곱 통계량.

1. 서론

단순 베이즈 분류 (naïve Bayes classification)는 반응변수가 주어졌을 때 설명변수들이 서로 독립이라는 단순 베이즈 가정에 기반하여 반응변수의 사후확률로 클래스를 예측하는 분류모형이다. 단순 베이즈 가정을 통해 클래스가 주어졌을 때 설명변수들의 결합 확률을 각 설명변수의 일차원 주변 확률들의 곱으로 계산이 가능하게 되며, 추정할 모수의 수를 대폭 감소시킴으로써 사후확률을 보다 간단하게 추정할 수 있도록 하는 계산상의 이점도 있다. 단순 베이즈 가정은 많은 경우 잘 성립하지 않는 다소 강한 가정임에도 불구하고 다양한 분야의 분류문제에서 합리적인 성능을 보여왔다 (Hand와 Yu, 2001). 특히 계산상의 이점으로 인해 많은 고차원 데이터 (high-dimensional data)의 분류문제에 적용되어 왔다. 가령 문서 분류에서 Ting 등 (2011)은 다른 분류방법인 지지벡터기계 (support vector machine), 신경망 (neural network), 의사결정나무 (decision tree)의 분류 성능과 비교하여 단순 베이즈 분류의 성능이 높음을 경험적으로 보였다. 또한 단순 베이즈 분류는 Kelemen 등 (2003)에서 효모의 마이크로어레이 (microarray) 데이터의 유전자발현 패턴의 분류에서 높은 예측 정확도를 보였다.

고차원 데이터에는 흔히 예측에 직접적인 영향을 미치는 신호변수 (signal variable)와 예측에 영향을 미치지 않는 잡음변수 (noise variable)가 혼재한다. 따라서 변수선택은 이러한 잡음변수를 제거하여 예측 성능과 모형 해석력을 향상시키는 것을 목적으로 한다 (Vidaurre 등, 2012). 문헌상에서 단순 베이즈 분류에서 변수선택에 대한 연구를 살펴보면 다음과 같다. Vidaurre 등 (2012)은 각 설명변수에 대한

[†] 이 성과는 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2020R1F1A1A01048268).

[1] (02504) 서울 동대문구 서울시립대로 163, 서울시립대 통계학과, 대학원생.
[2] 교신저자: (02504) 서울 동대문구 서울시립대로 163, 서울시립대 통계학과, 교수.
E-mail: park463@uos.ac.kr

반응변수의 클래스별 조건부 분포와 주변분포 간의 차이에 대하여 그룹 LASSO (group least absolute shrinkage and selection operator) 벌점화를 통한 변수선택을 고려하였다. 그러나 그룹 LASSO 벌점화에 의한 목적함수를 직접 최적화하기 어려워 근사적인 방법으로 AIC를 이용한 단계적 전진 선택법 (forward stagewise selection)을 제안하였다. Choi 등 (2014)은 정규 혼합분포에서 BIC를 기준으로 한 전진 선택법 (forward selection)을 제안하였다. Kim 등 (2015)은 단순 베이즈 분류에서 순서화된 카이제곱 통계량의 증가율이 변하는 지점까지를 후보모형으로 고려하는 변화점 분석 (change point analysis)에 기반한 범주형 변수선택을 제안하였다. Askari 등 (2019)은 반응변수가 주어졌을 때 설명변수가 가지는 확률의 차이가 미미한 갯수에 제약을 부여한 손실함수를 최적화 하는 방법을 제안하였다.

이제 FDR과 관련된 연구를 살펴보기로 하자. FDR은 Benjamini와 Hochberg (1995)가 소개한 다중비교 (multiple comparison)에서 정의되는 오류율로, 기각한 가설 중 오류가 발생할 기대비율을 의미한다. 본래 FDR은 대규모 (large-scale) 다중비교에서 오류율을 특정 수준 이하로 조절하여 유의한 대상을 선별하고자 할 때 사용된다. 예를 들어 마이크로어레이 데이터에서 수많은 유전자 중 발현에 직접적인 영향을 미치는 소수의 유전자를 선별하는 데 사용되었으며 (Dudoit, 2003), 신경영상 (neuroimaging)에서 뇌의 활성화된 위치를 찾기 위한 기준으로도 사용되었다 (Schwartzman 등, 2009). Yu 등 (2021)에서는 가우스 그래프 모형 (Gaussian graphical model)에서 유의한 간선 (edge)을 선별하는 데 이용하였다. 또한 본 연구와 직접적인 관련성은 없지만 Kang과 Jeon (2020)에서는 재생성 커널 힐버트 공간에서 랜덤 스케치 기법을 이용한 변수선택법을 제안하였고 Ahn과 Kim (2018)에서는 혼합정규분포의 가정하에 상호정보의 준모수적 추정량을 이용하여 고차원 데이터에 대한 변수선택을 연구한 바 있다.

본 논문에서는 범주형 변수로 이루어진 단순 베이즈 분류에서 각 설명변수와 반응변수 간의 카이제곱 통계량의 p-값 (p-value)에 FDR의 조절 절차를 적용하여 변수들을 선택하고자 한다. 참고로 Son (2020)에서는 불균형 텍스트 데이터에서 카이제곱 통계량을 이용하는 경우에 대하여 변수선택의 편향성을 연구한 바 있다. Kim 등 (2015)에서 고려한 변화점 분석도 카이제곱 통계량에 기반한 방법이지만 변수를 순차적으로 선택하지 않기 때문에 변수선택 시 설명변수와 잡음변수를 직접적으로 구분하지 않는다. 반면, FDR 조절 절차는 가장 작은 p-값부터 기각시키는 순차적인 검정을 시행하므로 신호변수를 선택하는 것을 우선적인 목표로 하며, 차원 및 신호변수 비율이 변하더라도 신호변수를 적절히 선택하는 동시에 잡음변수를 잘 제거할 수 있을 것으로 기대된다.

범주형 변수의 선택문제에서는 Vidaurre 등 (2012)의 단계적 전진선택법과 Kim 등 (2015)의 변화점 분석 기반의 방법과 비교해 볼 수 있다. Vidaurre 등 (2012)의 단계적 선택법에 대한 분석 코드를 얻을 수 없었고 범주형 변수에 대한 벌점화 방법에 의한 해의 계산상의 난점으로 인해 구현하지 못하여 본 연구에서는 카이제곱 통계량에 기반한 방법인 Kim 등 (2015)의 변화점 분석 기반의 방법과 비교만을 고려하였다. 또한 단순 베이즈 방법은 분류정확도가 매우 높지는 않지만 대용량 데이터베이스에서 범주의 개수를 세는 단순한 연산으로 구현 가능하다는 것이 장점이다. 카이제곱 통계량에 기반한 방법들의 계산량은 벌점화에 기반한 방법들에 비해 상대적으로 작아서 대용량 데이터베이스에서 단순 베이즈 분류의 장점을 그대로 유지하는 방법이라는 점도 고려되었다.

본 논문의 구성은 다음과 같다. 2절에서는 범주형 변수들에 대한 단순 베이즈 분류를 설명하고 FDR 조절 방법인 BH (Benjamini and Hochberg) 절차를 이용한 변수 선택법을 설명한다. 3절에서는 모의실험을 통해 카이제곱 통계량에 기반한 단순 베이즈 분류의 범주형 변수선택법인 BH 절차를 이용한 방법과 Kim 등 (2015)의 변화점 분석을 이용한 방법을 비교한다. 4절에서는 실제 데이터에 대하여 두 가지 방법을 적용하여 그 결과를 비교한다. 마지막으로 5절에서는 본 논문을 요약하고 향후 연구 방향에 대하여 논의한다.

2. 분석 방법론

이 절에서는 단순 베이즈 분류와 FDR 기반의 변수선택법에 대하여 설명한다. 우선 범주형 변수들로 구성된 단순 베이즈 분류에 대하여 소개하면 다음과 같다. 설명변수와 반응변수를 각각 $X = (X_1, \dots X_p)$와 $Y \in \mathcal{Y} = \{1, \dots, K\}$로 나타내기로 하자. 여기서 X_j는 범주형 변수로 j번째 순서형 또는 명목형 변수라고 하자. $\mathcal{X}_j$를 X_j가 가지는 범주들의 유한집합이라고 하면, 반응변수가 주어졌을 때 설명변수들은 서로 독립이라는 단순 베이즈 가정 하에서

$$\mathbb{P}(X_1 = m_1, \dots, X_p = m_p | Y = k) = \prod_{j=1}^{p} \mathbb{P}(X_j = m_j | Y = k), \quad k \in \mathcal{Y}, \quad m_j \in \mathcal{X}_j, \quad j = 1, \dots, p$$

를 만족한다.

$\{(\mathbf{x}_i, y_i)\}_{i=1}^N$를 (X, Y)로부터 독립적으로 관측된 N개의 데이터 쌍이라 하자. 여기서 $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$이다. 각 $k \in \mathcal{Y}$에 대해 $Y = k$의 확률과 $Y = k$이 주어졌을 때 각 X_j의 조건부 확률은 다음과 같이 훈련데이터에서 대응되는 상대도수로 추정할 수 있다:

$$\hat{\mathbb{P}}(Y = k) = \frac{1}{N} \sum_{i=1}^{N} I(y_i = k),$$
$$\hat{\mathbb{P}}(X_j = m_j | Y = k) = \frac{\sum_{i=1}^{N} I(x_{ij} = m_j, y = k) + 1}{\sum_{i=1}^{N} I(y_i = k) + |\mathcal{X}_j|},$$

여기서 $|\mathcal{X}_j|$은 $\mathcal{X}_j$의 원소의 갯수를 나타낸다. $\hat{\mathbb{P}}(X_j = m_j | Y = k)$의 추정값이 0이되는 문제를 피하기 위해 McCallum과 Nigam (1998)와 같이 분모와 분자에 각각 $|\mathcal{X}_j|$과 1을 더해주는 라플라스 스무딩 (Laplace smoothing) 을 적용하기로 한다.

각 $X_j, j = 1, \dots, p$이 주어졌을 때 Y의 사후확률 추정값은

$$\hat{\mathbb{P}}(Y = k | X_1 = m_1, \dots, X_p = m_p) \propto \prod_{j=1}^{p} \hat{\mathbb{P}}(X_j = m_j | Y = k) \hat{\mathbb{P}}(Y = k)$$

으로 나타낼 수 있고, 새로운 데이터 $\mathbf{x}^* = (m_1^*, \dots, m_p^*)$의 클래스는

$$\begin{aligned} k^* &= \underset{k \in \mathcal{Y}}{\operatorname{argmax}} \ln \hat{\mathbb{P}}(Y = k | X_1^* = m_1^*, \dots, X_p^* = m_p^*) \\ &= \underset{k \in \mathcal{Y}}{\operatorname{argmax}} \Big\{ \sum_{j=1}^{p} \ln \hat{\mathbb{P}}(X_j^* = m_j^* | Y = k) + \ln \hat{\mathbb{P}}(Y = k) \Big\} \end{aligned}$$

로 예측할 수 있다.

이제 이러한 단순 베이즈 분류에서 FDR기반의 변수 선택을 설명하기로 한다. $j(= 1, \dots, p)$번째 변수에 대응하는 귀무가설과 대립가설을 각각 H_{0j}과 H_{1j}라 하자. 변수선택의 관점에서 H_{0j}과 H_{1j}는 각각 X_j가 잡음변수 또는 신호변수인 경우를 나타낸다. 또한 $j(= 1, \dots, p)$번째 가설로부터 관측된 검정통계량을 u_j로 나타내자. H_{0j}가 참일 때 u_j에 대응되는 p-값 v_j은 독립적으로 다음의 균등분포를 따른다:

$$H_{0j} : v_j \sim U(0, 1).$$

주어진 임계값 (threshold) $C \in [0, 1]$하에서 전체 p개 중 신호변수로 판정된 변수들의 갯수를 $R(C) = \sum_{j=1}^{p} I(v_j < C)$라 하고, 실제로는 잡음변수지만 검정 결과 신호변수로 잘못 판정된 변수

들의 갯수를 $A(C) = \sum_{j=1}^{p}\{I(v_j < C)(1 - H_j)\}$라 하자. 여기서 H_j는 H_{0j}이 참이면 0, 그렇지 않으면 1의 값을 가진다. $A(C)$와 $R(C)$는 임계값 C에 의존하며 $R(C)$는 관측되지만 $A(C)$는 관측이 불가능하다. FDP (false discovery proportion)은 통계적 검정에서 신호변수로 판단하였지만 실제로는 잡음변수에 해당하는 비율을 의미하며 FDR은 FDP의 기댓값이다 (Benjamini와 Hochberg, 1995):

$$\text{FDP}(C) = \frac{A(C)}{R(C)}, \quad \text{FDR} = \mathbb{E}[\text{FDP}(C)].$$

만약 $R(C) = 0$으로 단 한 개의 가설도 기각하지 않은 경우 $\text{FDP}(C) = 0$으로 정의한다.

변수선택을 위한 BH 절차는 다음과 같다. 전체 가설 중 귀무가설이 참인 갯수를 $p_0 = \sum_{j=1}^{p} I(H_j = 0)$, 대립가설이 참인 갯수를 $p_1 = \sum_{j=1}^{p} I(H_j = 1)$이라 하자. 이 때, p_0는 모수 (parameter)로 간주한다. 또한 귀무가설과 대립가설이 참인 비율은 각각 $\pi_0 = p_0/p$과 $\pi_1 = p_1/p$로 나타내자. F_1을 실제 대립가설 하에서 생성된 p-값의 누적분포 함수라고 하면 p-값은 독립적으로 다음과 같은 혼합분포 (mixture distribution)를 따른다 (Genovese와 Wasserman, 2004):

$$F = \pi_0 U + \pi_1 F_1.$$

이에 대한 보다 자세한 설명은 Jang (2013)을 참고하기 바란다. p-값의 분포로부터 주어진 임계값 $c \in C$에 대해 아래의 부등식이 성립한다.

$$\text{FDR} = \mathbb{E}[\text{FDP(c)}] = \mathbb{E}\left(\frac{V(c)}{R(c)}\right) \leq \mathbb{E}\left(\frac{V(c)}{R(c)}|R(c) > 0\right) = \frac{\pi_0 c}{\hat{F}(c)} \leq \frac{c}{\hat{F}(c)},$$

여기서 $\hat{F}(c) = \sum_{j=1}^{p} I(v_j < c)/p$는 기각된 p-값의 비율을 의미한다. 이 때, 순서화된 p-값 $v_{(1)} \leq \cdots \leq v_{(p)}$의 j번째값을 임계점 $c = v_{(j)}$으로 놓으면 $\hat{F}(v_{(j)}) = \sum_{j=1}^{p} I(v_j < v_{(j)})/p = j/p$이다. 따라서 위 부등식의 우변을 조절하고자 하는 값 $\alpha \in [0, 1]$로 놓으면 모든 j번째 p-값에 대해 $\frac{v_{(j)}}{\hat{F}(v_{(j)})} = \alpha$로부터 관계식 $v_{(j)} = \alpha j/p$이 성립하므로 $\text{FDR} \leq \alpha$을 보장한다. 즉, 모든 j번째 가설에 대해 조정된 유의수준 $\alpha j/p$ 하에서 가장 작은 p-값으로부터 순차적으로 검정하여 $v_{(j)} \leq \alpha j/p$을 만족하는 가장 큰 j_{max}에 대해 $v_{(j)}, j = 1, \ldots, j_{max}$번째에 해당하는 귀무가설을 모두 기각한다.

변수선택 과정은 다음과 같다. 클래스가 $k \in \mathcal{Y}$이고 j번째 설명변수의 $m_j \in \mathcal{X}_j$범주가 가지는 결합 관측도수는 $O_j(k, m_j) = \sum_{i=1}^{N} I(y_i = k, x_{ij} = m_j)$이고, $O_j(\cdot, m_j) = \sum_{k\in\mathcal{Y}} O_j(k, m_j)$, $O_j(k, \cdot) = \sum_{m_j\in\mathcal{X}_j} O_j(k, m_j)$으로 나타내자. 결합 기대도수는 $E_j(k, m_j) = O_j(\cdot, m_j)\ O_j(k, \cdot)/N$으로 나타내어지므로 X_j와 Y의 카이제곱 통계량

$$\chi_j^2 = \sum_{k\in\mathcal{Y}} \sum_{m_j\in\mathcal{X}_j} \frac{(O_j(\cdot, m_j) - E_j(k, m_j))^2}{E_j(k, m_j)}$$

은 근사적으로 자유도 $(|\mathcal{X}_j| - 1)(k - 1)$인 카이제곱 분포 $\chi^2\{(|\mathcal{X}_j| - 1)(k - 1)\}$를 따른다. 훈련 데이터로부터 각 χ_j^2와 이에 대응하는 p-값 $v_j = \mathbb{P}(\chi^2 \geq \chi_j^2)$를 구하고 다음과 같이 BH 절차를 통해 j_{max}을 찾는다:

1. p-값들을 오름차순으로 정렬한다: $v_{(1)} \leq \cdots \leq v_{(p)}$.
2. 주어진 α에 대해 $j_{max} = \max\{j : v_{(j)} < \alpha j/p\}$을 찾는다.
3. $v_{(l)}, l = 1, \ldots, j_{max}$에 해당하는 모든 변수들을 선택한다.

α값은 교차검증 (cross validation)에 의해 선택한다. 또한 α의 선택시 Breiman 등 (1984)의 1-표준오차 규칙 (1-standard error rule)을 고려하였다. 즉, 가장 낮은 예측오차의 1-표준오차 범위내에서 변수

들을 가장 적게 선택하는 α값을 선택한다. 고차원 데이터에서는 변수선택 후에도 여전히 많은 잡음변수들이 남아 있을 수 있기 때문에, 1-표준오차 규칙은 추정한 모형의 예측오차가 다소 증가할지라도 더 적은 수의 변수들을 선택함으로써 단순한 모형을 준다는 장점이 있다.

3. 모의실험

이 절에서는 본 논문에서 제안하는 FDR 기반의 변수선택법 (BH-NB로 표기)을 Kim 등 (2015)의 변화점 기반의 변수 선택법 (CPT-NB로 표기)과 비교하기로 한다. 또한 변수 선택의 효과를 비교하기 위해 기준값으로서 변수 선택 기능이 없는 단순 베이즈 분류 (NB로 표기)도 같이 비교한다. 모든 분석은 R을 이용하였고 CPT-NB는 `changepoint`패키지의 `cpt.var`함수를 이용하였다. 또한 BH-NB의 경우 1-표준오차 규칙을 적용하였지만, Kim (2015)에서는 오분류율의 최솟값만을 기준으로 하였으므로 CPT-NB에는 1-표준오차 규칙을 적용하지 않았다.

모의실험 데이터는 Kim (2015)의 모의실험을 조금 변형하여 다음과 같이 생성하였다. Y는 성공률이 0.5인 베르누이 분포를 따르며, Y가 주어졌을 때 설명변수 $X_j, j = 1, \ldots, p$는 단순 베이즈 가정하에서 Table 3.1의 확률분포를 갖는다. p개의 변수 중 앞의 p_0개는 분류에 영향을 미치는 신호변수이고 나머지 $p - p_0$개는 분류에 영향을 미치지 않는 잡음변수이다.

Table 3.1 $P(X_j = m_j|Y = k)$ for generating simulated data

j	1∼10		11∼20		21∼30		31∼40		41∼50		51∼ p	
$m_j \setminus k$	0	1	0	1	0	1	0	1	0	1	0	1
1	0.3	0.5	0.4	0.3	0.2	0.1	0.2	0.2	0.3	0.3	1/3	1/3
2	0.3	0.4	0.2	0.4	0.5	0.5	0.3	0.5	0.3	0.4	1/3	1/3
3	0.4	0.2	0.4	0.3	0.3	0.4	0.5	0.3	0.4	0.3	1/3	1/3

데이터의 크기와 차원에 따른 변수선택의 성능을 비교하기 위해 훈련데이터의 크기를 $N \in \{100, 500, 1000\}$, 설명변수의 차원을 $p \in \{100, 500, 1000, 5000, 10000\}$의 수준에서 생성하였다. 또한 신호변수의 비율에 따른 변수선택 결과도 비교하기 위해 각 N과 p의 조합에서 신호변수의 갯수를 $p_0 \in \{10, 30, 50\}$의 수준에서 실험하였다. BH-NB의 경우 α에 대하여 등간격 로그 스케일 (log scale)로 20개의 값 $\{\exp(\log(10^{-4})), \exp(\frac{18}{19}\log(10^{-4})), \ldots, \exp(\log(10^{0}))\}$으로 이루어진 격자 (grid)를 고려한다. 모형 평가를 위한 시험데이터의 크기는 1000으로 고정하였고, 실험의 변동성을 고려하기 위해 데이터 생성, 모형 적합, 모형 평가에 해당하는 전 과정을 100회 반복하였다.

모형의 성능은 분류 정확도와 변수선택의 적절성의 두 측면에서 평가되었다. 분류 정확도는 오분류율로 측정하며 변수선택의 성능은 TP (true positive)와 FP (false positive)로 측정하였다. Table 3.2는 데이터의 크기 N, 설명변수의 차원 p, 신호변수 갯수 p_0의 각 수준에 대한 NB, BH-NB, CPT-NB의 평균 시험오분류율과 표준오차값을 보여준다. 데이터 크기와 설명변수 차원의 각 수준에 대하여 신호변수의 비율이 증가할 수록 오분류율은 감소하는 경향을 보인다. 두 가지 변수선택법 CPT-NB와 BH-NB 모두 변수선택을 하지 않는 NB에 비해 오분류율을 감소시켜주는데 BH-NB가 더 큰 폭으로 감소시켜주었다. 각 차원과 신호변수의 갯수의 수준에 대하여 데이터 크기가 증가하면 모든 방법에 대하여 오분류율이 감소하였다. 동일한 조건에서 데이터의 크기가 증가할 수록 단순 베이즈 분류에서 확률 추정의 정확도가 높아지기 때문으로 생각된다. 반면 다른 요인들의 수준이 고정되었을 때 설명변수의 차원이 증가함에 따라 각 방법의 오분류율은 증가한다. 다만 변수선택의 효과로 인해 NB에 비해 BH-NB, CPT-NB의 오분류율은 상대적으로 적은 폭으로 증가하며 BH-NB의 증가폭이 더 적게 나타났다. 특히 $N = 100$, $p = 100$, $p_0 = 50$의 경우 NB에 비해 BH-NB와 CPT-NB의 오분류율의 차이가 크지 않거나

Table 3.2 Average missclassification error rates for NB, BH-NB and CPT-NB with their standard errors in parentheses

p_0	p	N	NB	BH-NB	CPT-NB
10	100	100	0.3558(0.0022)	0.3412(0.0034)	0.3457(0.0023)
		500	0.2669(0.0017)	0.2252(0.0015)	0.2546(0.0016)
		1000	0.2421(0.0015)	0.2158(0.0015)	0.2352(0.0015)
	500	100	0.4317(0.0018)	0.4011(0.0045)	0.4211(0.0021)
		500	0.3496(0.0016)	0.2881(0.0040)	0.3259(0.0018)
		1000	0.3100(0.0016)	0.2303(0.0016)	0.2875(0.0016)
	1000	100	0.4518(0.0017)	0.4272(0.0042)	0.4446(0.0019)
		500	0.3863(0.0016)	0.2732(0.0030)	0.3650(0.0019)
		1000	0.3499(0.0016)	0.2870(0.0039)	0.3245(0.0015)
	5000	100	0.4824(0.0019)	0.4637(0.0040)	0.4772(0.0017)
		500	0.4471(0.0016)	0.3129(0.0045)	0.4349(0.0016)
		1000	0.4237(0.0017)	0.2549(0.0021)	0.4062(0.0015)
	10000	100	0.4882(0.0016)	0.4742(0.0036)	0.4855(0.0017)
		500	0.4632(0.0018)	0.3086(0.0047)	0.4541(0.0018)
		1000	0.4470(0.0018)	0.2903(0.0037)	0.4345(0.0017)
30	100	100	0.2380(0.0021)	0.2434(0.0031)	0.2339(0.0023)
		500	0.1524(0.0013)	0.1392(0.0014)	0.1453(0.0014)
		1000	0.1370(0.0012)	0.1280(0.0011)	0.1346(0.0012)
	500	100	0.3559(0.0023)	0.3407(0.0043)	0.3378(0.0024)
		500	0.2282(0.0014)	0.1676(0.0018)	0.2015(0.0017)
		1000	0.1859(0.0013)	0.1368(0.0015)	0.1675(0.0013)
	1000	100	0.3972(0.0020)	0.3744(0.0036)	0.3824(0.0230)
		500	0.2782(0.0015)	0.1947(0.0023)	0.2451(0.0017)
		1000	0.2283(0.0015)	0.1600(0.0023)	0.2014(0.0014)
	5000	100	0.4619(0.0023)	0.4379(0.0041)	0.4520(0.0022)
		500	0.3878(0.0017)	0.2590(0.0035)	0.3631(0.0018)
		1000	0.3424(0.0016)	0.1747(0.0017)	0.3134(0.0015)
	10000	100	0.4763(0.0020)	0.4551(0.0039)	0.4698(0.0021)
		500	0.4201(0.0019)	0.2157(0.0034)	0.4013(0.0019)
		1000	0.3870(0.0016)	0.2237(0.0029)	0.3620(0.0016)
50	100	100	0.1638(0.0018)	0.1861(0.0031)	0.1700(0.0023)
		500	0.0963(0.0010)	0.0940(0.0011)	0.0951(0.0010)
		1000	0.0887(0.0010)	0.0874(0.0010)	0.0881(0.0010)
	500	100	0.2946(0.0025)	0.2790(0.0043)	0.2775(0.0025)
		500	0.1558(0.0013)	0.1145(0.0016)	0.1352(0.0014)
		500	0.1215(0.0012)	0.0913(0.0011)	0.1076(0.0012)
	1000	100	0.3484(0.0024)	0.3239(0.0042)	0.3324(0.0023)
		500	0.2056(0.0014)	0.1453(0.0019)	0.1742(0.0016)
		500	0.1573(0.0013)	0.1030(0.0013)	0.1335(0.0014)
	5000	100	0.4430(0.0027)	0.4163(0.0045)	0.4300(0.0025)
		500	0.3373(0.0018)	0.2155(0.0028)	0.3059(0.0019)
		1000	0.2798(0.0014)	0.1278(0.0017)	0.2442(0.0014)
	10000	100	0.4661(0.0024)	0.4381(0.0043)	0.4556(0.0024)
		500	0.3820(0.0021)	0.1628(0.0021)	0.3565(0.0021)
		1000	0.3361(0.0015)	0.1811(0.0025)	0.3034(0.0015)

오히려 더 크게 나타났다. 이는 설명변수중 신호변수의 비율이 높고 데이터의 갯수가 충분치 않으면 변수선택이 그다지 도움이 되지 않는 것으로 보인다.

Table 3.3은 N, p, p_0의 각 수준에 대하여 단순 베이즈 분류에 대한 두 변수 선택법 BH-NB와 CPT-NB의 변수선택의 측도인 TP와 FP의 평균값을 보여준다. 먼저 TP의 결과를 살펴보자. 데이터의 크기가 100인 경우와 $500, 1000$에서 TP값이 다르게 나타난다. $N = 100$의 경우 BH-NB는 약 30 $\sim$ 60%의 실제 신호변수를 선택하는 경향이 있는데, 이는 상대적으로 반응변수가 주어졌을 때 조건부 분포가 클래스에 대하여 차이가 크지 않은 변수들인 $X_{21}, \ldots, X_{30}$ 그리고 $X_{41}, \ldots, X_{50}$의 선택이 적게 되기 때문으로 보인다. 반면 CPT-NB는 신호변수의 비율에 따라서도 결과가 달라지는데, $p_0 = 10$에서는 약 80 $\sim$ 90%를, $p_0 = 30, 50$에서는 약 70%의 신호변수들을 선택한다. 상대적으로 크기가 큰 $N = 500, 1000$에서는 BH-NB와 CPT-NB 모두 $N = 100$의 결과와 달리 조건부 분포가 비슷한 신호변수들도 데이터 크기가 커지면서 p-값의 정확도가 높아져서 잘 선택되는 것으로 보인다.

이제 FP의 측면에서 두 변수선택법을 비교해보자. BH-NB는 데이터의 크기가 100인 경우, 설명변수의 차원 p가 증가함에 따라 FP값이 전체적으로 증가하는 경향을 보이지만 적어도 전체 잡음변수의 약 50%정도를 줄일 수 있었다. 또한 $p = 5000, 10000$으로 설명변수의 차원이 큰 경우 표준오차가 매우 크게 나타났다. 이는 차원이 큰 경우 신호변수와 잡음변수에서 계산되는 p-값의 구분이 어려워 BH 절차에서 거의 대부분의 변수들을 선택하기 때문으로 보인다. 반면 BH-NB는 상대적으로 데이터의 크기가 큰 $N = 500, 1000$에서는 일부 결과를 제외하면 잡음변수를 평균적으로 10% 이하에서 선택하는 경향을 보인다.

CPT-NB가 일부 결과를 제외하면 데이터 크기나 차원, 신호변수의 크기에 상관없이 잡음변수를 20 $\sim$ 30%를 선택하는 것을 고려하면, BH-NB는 데이터 크기와 신호변수 차원에 따라 변동이 있는 편이지만 전체적으로 더 적은 수의 잡음변수를 선택한다고 볼 수 있다. 다만 $p = 10000$이고 $N = 100$인 경우 BH-NB의 CPT-NB에 비해 FP의 값이 크게 나타났다. $N = 1000, 5000, 10000$의 경우를 비교해 보면 N에 따라 BH-NB의 FP의 표준오차값이 점점 커지는 것으로 보아 고차원성이 심화되면 BH-NB에 의한 변수 선택이 불안정한 경우가 있음을 알 수 있다. 따라서 고차원 데이터에 대하여 BH-NB를 적용할 때에는 주의를 요한다. 더 명확한 가이드 라인을 제시하기 위해서는 추가적인 연구가 더 필요할 것이다.

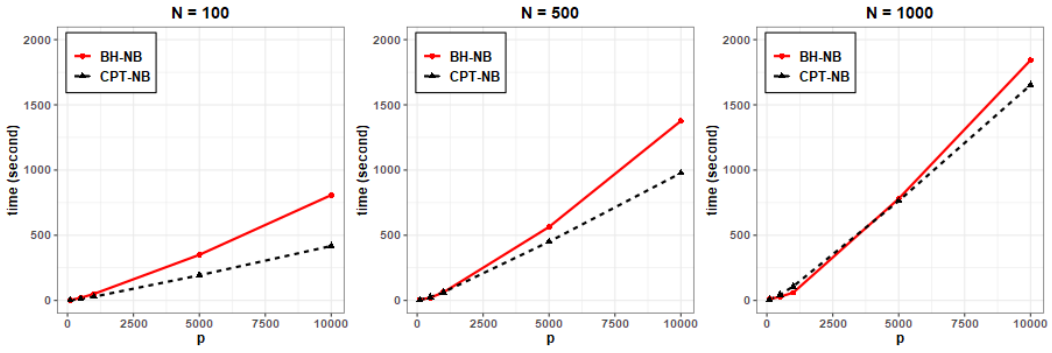
Figure 3.1은 모의실험에서 데이터 크기를 고정시킨 후 설명변수의 차원에 따른 BH-NB와 CPT-NB의 평균 소요시간과 그 표준오차를 신호변수 크기 별로 나타낸 것이다. 소요시간은 Window10 (Intel (R) Core (TM) i5-9500 CPU at 3.00GHz, 8GB RAM)에서 5개의 코어 (core)로 병렬처리 하여 측정하였으며 단위는 초 (second)이다. 먼저 전체적으로 BH-NB과 CPT-NB 모두 소요시간이 p가 증가함에 따라서 증가 추세를 보이며, BH-NB의 소요시간이 CPT-NB보다 크게 나타난다. 그리고 설명변수의 차원이 작으면 BH-NB와 CPT-NB의 소요시간 차이는 미미하지만, 설명변수의 차원이 커지면 두 방법의 소요시간 차이가 발생하며 데이터 크기가 작을 때 더 많은 차이가 났다. 아마도 차원에 따른 계산의 비용이 BH-NB에서 사용되는 정렬알고리즘의 경우 안정적으로 증가하는데 반해 CPT-NB은 회귀분석에 기반하기 때문에 차이가 발생하는 것으로 보인다.

4. 실제 데이터 분석

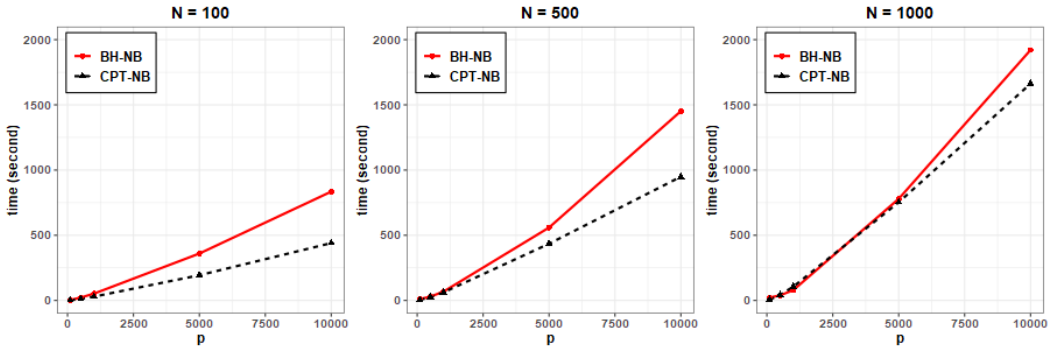
이 절에서는 FDR기반 범주형 변수의 선택방법인 BH-NB를 실제 데이터에 적용하고 변화점 분석에 기반한 변수선택법 CPT-NB와 성능을 비교한다. 다음과 같은 미국 생물공학 정보센터 (National Center for Biotechnology Information)와 UCI 기계학습 저장소 (UCI Machine Learning Repository)에서 제공하는 데이터를 분석하였다:

Table 3.3 Average percentage of TP, FP for BH-NB and CPT-NB with their standard errors in parentheses

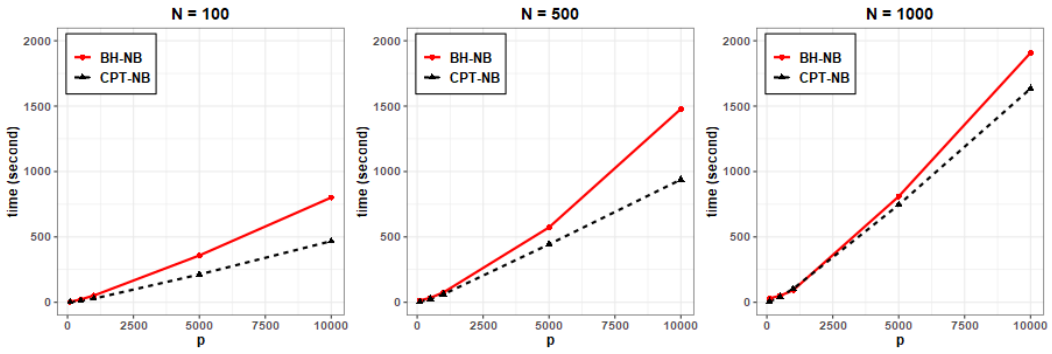
p_0	p	N	BH-NB		CPT-NB	
			TP	FP	TP	FP
10	100	100	66.0000(0.2340)	15.2667(1.5926)	90.5000(0.0978)	36.9000(1.4529)
		500	97.3000(0.0529)	1.0444(0.1619)	100.0000(0.0000)	33.4222(1.0054)
		1000	99.9000(0.0100)	0.6000(0.1019)	100.0000(0.0000)	35.4000(1.2501)
	500	100	54.3000(0.3242)	25.2306(20.2092)	88.3000(0.1016)	30.1673(3.3018)
		500	99.6000(0.0243)	14.3612(6.9356)	100.0000(0.0000)	29.7469(3.0938)
		1000	100.0000(0.0000)	1.6878(0.7795)	100.0000(0.0000)	28.8143(2.7560)
	1000	100	56.0000(0.3803)	39.4869(48.1481)	89.1000(0.1006)	30.8869(7.6504)
		500	99.8000(0.0141)	2.9596(2.4510)	100.0000(0.0000)	29.2313(5.6413)
		1000	100.0000(0.0000)	15.6606(16.4194)	100.0000(0.0000)	27.5960(3.5467)
	5000	100	61.3000(0.4380)	55.0483(246.2324)	86.9000(0.0992)	26.6248(2.8193)
		500	99.6000(0.0197)	1.2060(5.9690)	100.0000(0.0000)	28.5675(22.0045)
		1000	100.0000(0.0000)	0.4583(1.3582)	100.0000(0.0000)	27.6549(10.6484)
	10000	100	60.2000(0.4495)	55.0279(499.1904)	86.9000(0.0992)	26.7196(4.7710)
		500	98.3000(0.0403)	0.6130(12.9951)	100.0000(0.0000)	31.0219(68.9992)
		1000	100.0000(0.0000)	0.6286(5.7764)	100.0000(0.0000)	28.5399(41.0655)
30	100	100	62.3333(0.5096)	17.8286(1.0630)	79.0000(0.3323)	35.5143(1.3666)
		500	91.9333(0.2527)	3.3429(0.2602)	99.5000(0.0435)	31.9714(1.5822)
		1000	96.7000(0.1291)	0.8143(0.1166)	100.0000(0.0000)	43.4286(2.1576)
	500	100	37.9667(0.7860)	12.8106(13.8054)	77.0333(0.2821)	29.5957(3.3900)
		500	96.4333(0.1810)	8.1489(1.8932)	99.4667(0.0443)	27.6000(3.2370)
		1000	98.6000(0.1075)	3.5319(1.4198)	100.0000(0.0000)	29.5404(3.8542)
	1000	100	38.0333(0.9570)	18.0052(34.9925)	77.5667(0.2601)	30.1134(7.6651)
		500	97.2333(0.0933)	7.5557(3.2525)	99.5000(0.0411)	27.3701(5.1852)
		1000	99.7000(0.0321)	7.7423(9.2353)	100.0000(0.0000)	26.6763(4.3458)
	5000	100	47.9000(1.3719)	43.0815(246.9394)	75.1667(0.2434)	26.5085(2.7939)
		500	94.4333(0.1450)	3.3990(10.0247)	99.4667(0.0443)	29.2109(26.7689)
		1000	99.3667(0.0394)	1.3362(2.5539)	100.0000(0.0000)	27.7280(14.5663)
	10000	100	47.6333(1.4051)	44.0445(496.9966)	75.2333(0.2446)	26.6529(4.7680)
		500	86.6333(0.1828)	0.6606(9.5459)	99.6333(0.0345)	31.1631(70.0790)
		1000	99.4333(0.0403)	1.9012(10.3732)	100.0000(0.0000)	28.7509(43.7714)
50	100	100	59.0400(0.8666)	19.4000(0.7428)	72.3200(0.6101)	31.9000(0.9808)
		500	86.9000(0.4352)	4.4000(0.2335)	98.6200(0.1203)	54.3000(1.3381)
		1000	92.8800(0.3151)	1.6400(0.1344)	99.8200(0.0379)	55.8200(1.2249)
	500	100	40.3800(0.9796)	9.3711(8.6045)	72.8000(0.3957)	30.1711(3.9317)
		500	91.9800(0.2942)	9.7133(2.7628)	97.6600(0.1055)	28.2356(4.2809)
		1000	97.3000(0.1743)	5.9156(1.8981)	99.7400(0.0367)	27.2467(3.5736)
	1000	100	38.6400(1.2939)	12.6337(25.0876)	73.1600(0.3782)	29.9600(7.3008)
		500	93.9800(0.1839)	11.4653(4.1069)	97.6600(0.0980)	28.2284(6.9091)
		1000	98.0200(0.1307)	5.8874(3.8268)	99.7400(0.0345)	26.4989(5.4944)
	5000	100	42.3800(2.2335)	37.1368(239.67)	70.7400(0.3237)	26.3952(2.7238)
		500	90.7800(0.2155)	5.3859(11.7355)	97.9000(0.0978)	29.9552(30.6455)
		1000	97.0400(0.1185)	2.1788(3.9287)	99.7600(0.0356)	27.8236(17.8996)
	10000	100	44.2600(0.3440)	41.0473(491.45)	70.8800(0.3230)	23.5850(4.7070)
		500	79.7200(0.2570)	0.6602(2.7434)	98.0400(0.0899)	30.8489(69.5528)
		1000	97.6600(0.1006)	3.1608(13.8889)	99.7800(0.0345)	28.4394(40.9459)



(a) $p_0 = 10$



(b) $p_0 = 30$



(c) $p_0 = 50$

Figure 3.1 Average running times for learning BH-NB and CPT-NB on simulated data

- 경구 독성 (oral toxicity) 데이터 (`https://archive.ics.uci.edu/ml/datasets/QSAR+oral+toxicity`)
  8992개의 화학물질에 대한 데이터로 반응변수는 각 화학물질을 경구 복용했을 때 741개의 매우 독성이 강한 (very toxic) 혹은 8251개의 그렇지 않은 (not very toxic) 두 가지 클래스를 갖는다. 설명변수는 1024개의 분자 지문 (molecular fingerprint) 으로 0 또는 1의 값을 가진다.
- SNP (single nucleotide polymorphism) 유전자형 (genotype) 데이터 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39428`)
  미국 생물공학 정보센터에서 운영하는 데이터베이스인 GEO (Gene Expression Omnibus)에서 제공하는 GSE39428 데이터이다. 전체 데이터의 크기는 480이며 266개의 류마티스 관절염 (rheumatoid arthritis), 51개의 강직성 척추염 (ankylosing spondylitis), 163개의 건강한 대조군 (healthy controls)으로 이루어진 반응변수를 가진다. 설명변수는 성별과 831개의 SNP 유전자형 (AA, AB, BB, non-class)으로 구성되어 있다.
- 대장내시경과 위장병 (gastoenterology) 데이터 (`https://archive.ics.uci.edu/ml/datasets/Gastrointestinal+Lesions+in+Regular+Colonoscopy`)
  위장병변을 감지하는데 사용된 대장내시경 영상에서 추출한 크기 152개의 데이터로 42개의 과형성(hyperplasic), 80개의 선종 (adenoma), 30개의 톱니모양 선종 (serrated adenoma)으로 구분된다. 변수들로는 영상에서 수집한 2D 질감, 2D 색상, 3D 색상으로 구성되어 있는데, 모든 값이 0으로 구성된 변수는 분석에서 제외하여 201개의 2D 질감, 54개의 2D 색상, 184개의 3D 형상을 나타내는 설명변수들만을 분석에 사용하였다.
- 질량 분석 (mass spectrometric) 데이터 (`https://archive.ics.uci.edu/ml/datasets/Micro Mass`)
  100명의 암 환자 (난소 암 또는 전립선 암)와 정상 환자로 구분되어 있다. 분석에서는 암 환자를 양성 (positive), 정상 환자를 음성 (negative)의 클래스로 구분하였고 2554개의 설명변수들은 혈청 내의 단백질의 질량 발현정도를 나타내는 연속형 변수들이다.

Table 4.1 Results for NB, BH-NB, and CPT-NB on real datasets (a) Average missclassification error rates with their standard errors in parentheses

Method	Dataset			
	oral toxicity	SNP	gastroenterology	mass spectrometry
NB	0.2158(0.0022)	0.4184(0.0044)	0.3471(0.0080)	0.3340(0.0051)
BH-NB	0.1357(0.0012)	0.3421(0.0039)	0.3631(0.0079)	0.3287(0.0049)
CPT-NB	0.2113(0.0022)	0.3981(0.0049)	0.3551(0.0077)	0.3320(0.0052)

(b) Average numbers of selected variables with their standard errors in parentheses

Method	Dataset			
	oral toxicity	SNP	gastroenterology	mass spectrometry
BH-NB	124.0700(0.9284)	30.0500(1.8447)	335.1000(11.6478)	358.3700(45.0063)
CPT-NB	550.8400(5.8456)	148.8500(5.1812)	354.6500(6.1627)	1784.4600(31.1311)
BH-NB and CPT-NB	124.0700(0.9284)	30.0500(1.8447)	297.74(10.2624)	352.97(42.6090)

본 연구에서는 설명변수가 범주형인 경우만을 고려하므로 연속형 설명변수의 경우 Kim (2015)에서 분석한 바와 같이 사분위수를 이용하여 범주화 하였다. 각 데이터는 7:3의 비율로 훈련데이터와 시험데이터로 랜덤하게 분할한 후 모형을 적합 및 평가하였다. 또한 실험의 변동성을 고려하기 위해 데이터 분

할 및 모형 적합과 평가의 전 과정을 100번 반복하였다. Table 4.1은 앞서 설명한 네 가지 실제 데이터를 NB, BH-NB, CPT-NB를 적용하여 얻은 평균 시험오분류율과 선택된 변수의 평균 갯수를 보여준다. 변수선택의 전과 후를 비교하면 경구독성과 SNP 데이터의 경우 변수선택 후 오분류율이 감소하였고, 질량 분석 데이터는 변수선택 후 오분류율이 변수선택 전과 거의 차이가 없으며, 위장병 데이터의 경우에는 변수선택 후에 오분류율이 증가하였다.

경구 독성 데이터는 데이터 크기에 비해 설명변수의 차원이 가장 작은 데이터이다. 전체 1024개 변수 중 BH-NB는 약 124개의 변수를 선택하였고, CPT-NB에서 선택된 약 551개보다 더 적은 변수를 선택했는데 오분류율은 더 작았다. SNP와 위장병 데이터는 데이터 크기와 설명변수의 차원이 비슷하다. SNP 데이터의 경우 BH-NB가 선택한 변수는 약 30개로 CPT-NB가 선택한 149개보다 더 적은 변수를 선택했으며 더 낮은 오분류율을 보였다. 위장병 데이터에서는 BH-NB와 CPT-NB는 335개와 255개의 변수들을 선택하여 선택된 변수들의 갯수가 대략 비슷하였고, 오분류율도 두 방법 모두 조금 증가하였다. 질량 분석 데이터는 데이터 크기에 비해 설명변수의 차원이 가장 크다. BH-NB는 약 358개의 변수를 선택하였는데, CPT-NB로 선택한 1784개보다 훨씬 적은 갯수의 변수를 선택하였지만 더 낮은 오분류율을 보였다. 또한 각 데이터에 대하여 BH-NB와 BH-CPT에 의해 공통적으로 선택된 변수들의 갯수와 BH-NB에 의해 선택된 변수들의 갯수가 거의 비슷한 것으로 보아 CPT-NB는 BH-NB에 비해 불필요한 노이즈 변수들을 더 많이 선택하는 것으로 여겨진다.

5. 결론

본 연구에서는 단순 베이즈 분류에서 FDR기반의 범주형 변수의 선택 방법을 제안하였고 모의실험과 실제 데이터의 분석을 통해 Kim 등 (2015)의 변화점 분석 기반의 방법과 예측 정확도 및 변수선택의 성능의 두 가지 관점에서 비교하였다. FDR기반의 방법에서는 설명변수 차원이 데이터의 크기보다 큰 고차원의 경우에서 데이터가 충분할 때에는 잡음변수를 거의 선택하지 않아 변화점 분석 기반의 방법에 비해 예측 오차 뿐만 아니라 FP 측면에서 더 효율적으로 나타났다. 또한 데이터가 충분하지 않은 경우에도 변화점 기반의 방법보다는 잡음변수들을 훨씬 덜 선택하여 변수선택 측면에서는 더 효과적이었다. 반면, TP 측면에서는 변화점 분석 기반의 방법이 FDR 기반의 방법보다 조금 더 많은 신호변수들을 선택하였지만 그 차이는 그리 크지 않고 오분류율 측면에서는 오히려 FDR 기반의 방법이 더 좋았다. 또한 실제 데이터의 분석 결과를 통해서도 FDR기반의 방법이 더 적은 변수를 선택하면서도 오분류율이 더 낮거나 큰 차이가 나지 않는 것을 확인하였다.

FDR을 조절하기 위한 검정은 p-값들이 서로 독립이라는 가정 하에서 이루어진다. 그러나 실제로 많은 문제들은 p-값들간의 독립성을 보장할 수 없는 경우가 많으며, 이에 따라 p-값들간 비독립 구조에 관한 연구가 많이 이루어지고 있다. 따라서 향후 과제로써 Benjamini와 Yekutieli (2001)에서 소개한 바와 같이 p-값이 다양한 비독립 구조를 가질 때 개발된 FDR을 단순 베이즈 분류에서의 변수선택 문제에 적용해 볼 수 있을 것이다. 혹은 G'Sell 등 (2016)이 순차적으로 모형을 선택하는 방식으로 개발한 FDR 조절 절차를 단순 베이즈 분류의 변수선택 방법에 맞게 적용하여 볼 수도 있을 것이다.

References

Ahn, C. and Kim, D. (2018). Variable selection based on semi-parametric estimator of conditional mutual information assuming normal mixture ih high-dimensional data. *Journal of the Korean Data & Information Science Society*, **29**, 1339-1351.

Askari, A., d'Aspremont, A. and Ghaoui, L. E. (2019). Naïve featureselection: Sparsity in naïve Bayes. arXiv:1905.098842v2.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**, 289-300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165-1188.

Choi, B. J., Kim, K. R., Cho, K. D., Park, C. and Koo, J. Y. (2014). Variable selection for naïve Bayes semisupervised learning. *Communications in Statistics-Simulation and Computation*, **43**, 2702-2713.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71-103.

Friedman, J., Hastie, T. and Tibshirani, R. (2001). The elements of statistical learning. Springer, New York.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics*, **32**, 1035-1061.

G'Sell, M. G., Wager, S., Chouldechova, A. and Tibshirani, R. (2016). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society Series B*, **78**, 423-444.

Hand, David J. and Keming Yu. (2001). Idiot's Bayes?not so stupid after all?. *International Statistical Review*, **69**, 385-398.

Jang, W. C. (2013). Multiple testing and its applications in high-dimension. *The Korean Journal of Applied Statistics*, **24**, 1063-1076.

Kang, J. and Jhun, M. (2020). Variable selection in reproducing kernel Hilbert space using random sketch method. *Journal of the Korean Data & Information Science Society*, **31**, 501-511.

Kelemen, A., Zhou, H., Lawhead, P. and Liang, Y. (2003). Naïve Bayesian classifier for microarray data. In *Proceedings of the International Joint Conference on Neural Networks*, 1769-1773, doi: 10.1109/IJCNN.2003.1223675.

Kim, M. S., Choi, H. S. and Park, C. (2015). Categorical variable selection in naïve Bayes classification. *The Korean Journal of Applied Statistics*, **28**, 407-415.

McCallum, A. and Nigam, K. (1998). A comparison of event models for naïve bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, 41-48.

Schwartzman, A., Dougherty, R. F., Lee, J., Ghahremani, D. and Taylor, J. E. (2009). Empirical null and false discovery rate analysis in neuroimaging. *Neuroimage*, **44**, 71-82.

Son, W. (2020). Skewness of chi-square statistic for imblanaced text data. *Journal of the Korean Data & Information Science Society*, **31**, 807-821.

Ting, S. L., Ip, W. H. and Tsang, A. H. (2011). Is naïve Bayes a good classifier for document classification? *International Journal of Software Engineering and Its Applications*, **5**, 37-46.

Vidaurre, D., Bielza, C. and Larranaga, P. (2012). Forward stagewise naïve Bayes. *Progress in Artificial Intelligence*, **1**, 57-69.

Yu, L., Kaufmann, T. and Lederer, J. (2021). False discovery rates in biological networks. arXiv:1907.03808v3.

Journal of the Korean Data & Information Science Society
2021, **32**(6), 1329–1341

http://dx.doi.org/10.7465/jkdi.2021.32.6.1329
한국데이터정보과학회지

FDR-based categorical variable selection in naïve Bayes classification[†]

Jieun Shin[1] · Changyi Park[2]

[12]Department of Statistics, University of Seoul

Received 12 August 2021, revised 25 October 2021, accepted 28 October 2021

Abstract

Naïve Bayes classification is based on the naïve Bayes assumption that explanatory variables are conditionally independent given the response variable. Although the naïve Bayes assumption is rather strong, the naïve Bayes classifier shows reasonable performances and has computational advantages on high-dimensional data. Since high-dimensional data sets usually have many noisy variables, variable selection can improve the accuracy in prediction and the interpretation of the classifier. In this paper, we propose a categorical variable selection method based on FDR control in naïve Bayes classification. Through simulations and real data analysis, the proposed method is compared with another variable selection method based on change point analysis and the proposed methods is illustrated to be more effective, particularly, for sparse or high-dimenional data.

Keywords: Chi-square statistic, high-dimensional data, naïve Bayes assumption.

[†] This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1F1A01048268).

[1] Graduate student, Department of Statistics, University of Seoul, Seoul 02504, Korea.
[2] Corresponding author: Professor, Department of Statistics, University of Seoul, Seoul 02504, Korea. E-mail: park463@uos.ac.kr