The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.3.407
28(3), 407–415

Categorical Variable Selection in Naïve Bayes Classification

Min-Sun Kim[a] · Hosik Choi[b] · Changyi Park[a,1]

[a]Department of Statistics, University of Seoul
[b]Department of Applied and Informational Statistics, Kyonggi University

(Received January 20, 2015; Revised March 3, 2015; Accepted March 10, 2015)

Abstract

Naïve Bayes Classification is based on input variables that are a conditionally independent given output variable. The Naïve Bayes assumption is unrealistic but simplifies the problem of high dimensional joint probability estimation into a series of univariate probability estimations. Thus Naïve Bayes classifier is often adopted in the analysis of massive data sets such as in spam e-mail filtering and recommendation systems. In this paper, we propose a variable selection method based on χ^2 statistic on input and output variables. The proposed method retains the simplicity of Naïve Bayes classifier in terms of data processing and computation; however, it can select relevant variables. It is expected that our method can be useful in classification problems for ultra-high dimensional or big data such as the classification of diseases based on single nucleotide polymorphisms(SNPs).

Keywords: big data, χ^2 statistic, Naïve Bayes assumption, SNP

1. 서론

단순 베이즈 분류(Naïve Bayes classification)는 출력변수(output variable)가 주어졌을 때 입력변수(input variable)들이 조건부 독립이라는 소위 단순 베이즈 가정에 기반한다. 또한 단순 베이즈 분류에서는 모든 확률이 대응되는 상대도수로 쉽게 추정되도록 모든 입력변수들을 범주화시키는 것이 일반적이다. 많은 경우에 단순 베이즈 가정은 비현실적이지만 고차원의 확률 추정 문제를 일련의 일차원 확률 추정 문제로 단순화 시킨다는 장점이 있다. Hand와 Yu (2001)에서 지적했듯이 단순 베이즈 분류는 비현실적인 가정에도 불구하고 합리적인 성능을 보이는 경우가 적지 않다.

요컨데 단순 베이즈 분류의 핵심은 문제를 단순화하여 모든 계산이 쉽도록 하는 것으로 방대한 데이터를 다루는 분야에서 흔히 사용되고 있다. 특히 스팸 메일 필터링, 추천 시스템(recommendation system) 등의 텍스트 마이닝 문제나 네트워크 마이닝 분야에서 많이 사용된다. 목적함수가 입력변수에 대하여 가법적으로 분해되므로 최적화 문제를 병렬 또는 분산 처리로 쉽게 구현할 수 있다. 따라서 최근 화두가 되고 있는 빅 데이터의 분류 문제에서 유용할 것으로 기대된다.

This work was supported by the 2014 Research Fund of the University of Seoul.
[1]Corresponding author: Department of Statistics, University of Seoul, 90 Jeonnong-Dong, Dongdaemun-Gu, Seoul 130-743, Korea. E-mail: park463@uos.ac.kr

그러나 단순 베이즈 분류는 최종 분류 모형에서 클래스(class)들을 구분하는 설명력 있는 입력변수들을 알 수 없다는 단점이 있다. 단순 베이즈 분류의 변수 선택 문제는 문헌상에서 그다지 다루어지지 않았는데, 예를 들어 Choi 등 (2014)에서는 가우스 혼합 모형에 기반하여 BIC 등을 이용한 전진 선택법을 제안하였다. 최근 Vidaurre 등 (2012)에서는 각 입력 변수에 대한 클래스별 조건부 분포와 주변 분포간의 차이에 대하여 그룹 LASSO(group least absolute shrinkage and selection operator) 벌점화를 통한 변수 선택을 고려하였는데, 그룹 LASSO 벌점화에 의한 목적함수를 직접 최적화하는 것은 어렵기 때문에 근사적인 방법으로 AIC를 이용하여 단계적 전진 선택법(forward stagewise selection)을 제안하였다. 후속 논문인 Vidaurre 등 (2013)에서는 마찬가지로 목적함수의 최적화가 어려운 문제를 피하기 위하여 단순 베이즈 분류를 회귀 문제 형태로 변형한 방법을 제안하였다.

사실 단순 베이즈 분류의 최대의 장점은 데이터 처리 및 계산의 단순성에 있다. 본 논문에서는 고차원의 범주형 입력변수로 이루어진 단순 베이즈 분류에서 입력변수와 출력변수간의 카이제곱 검정통계량에 기반한 변수의 순위에 의해 설명력있는 변수를 선택하고자 한다. Fan과 Lv (2008)에서는 LASSO 등의 벌점화에 의한 변수선택법을 직접 적용할 수 없는 초고차원 데이터에서 각 입력변수의 출력변수에 대한 상관계수의 크기 순으로 변수를 미리 스크린한 후 변수선택법을 적용할 것을 제안하고 있다. 본 논문에서 제안하는 방법은 각 범주형 입력변수에 출력변수와의 카이제곱 통계량을 이용한다는 점에서 SIS(sure independence screening)와 유사하다. SIS와의 차이점은 사전 스크린 단계 없이 직접 카이제곱 통계량에 기반한 변수의 선택이 이루어지므로 초고차원 데이터에 대해서도 직접적으로 적용이 가능하며 SNP(single-nucleotide polymorphism)에 의한 질병의 분류 등의 초고차원 혹은 빅데이터에서 단순하지만 매우 효과적일 것으로 기대된다.

본 논문의 구성은 다음과 같다. 2절에서는 단순 베이즈 분류에서 카이제곱 통계량을 이용한 범주형 입력변수의 선택법을 소개한다. 3절에서는 모의실험 및 실제 데이터에 대하여 변수 선택 기능이 없는 단순 베이즈 분류와 본 논문에서 제안하는 방법을 분류 정확도 및 변수 선택의 적절성 관점에서 비교한다. 마지막으로 4절에서는 본 논문을 요약하고 추후의 연구 방향에 대한 몇 가지 제언을 하고자 한다.

2. 단순 베이즈 분류에서 입력변수의 선택

단순 베이즈 분류에서의 변수선택법을 설명하기 전에 우선 몇 가지 기호 및 개념을 소개한다. $\boldsymbol{X} = (X_1, \ldots, X_p)^T$와 $Y \in \mathcal{Y}$는 각각 입력 및 출력 변수를 나타낸다. 여기서 $\mathcal{Y}$는 클래스를 나타내는 유한집합으로 원소의 개수가 K이다. 본래 단순 베이즈 분류에서 연속형 입력변수들은 일변량 정규분포를 이용하여 모형화하는데, 본 논문에서는 모든 입력변수가 범주형인 경우만을 고려한다. 즉, 모든 입력 변수들은 $X_j \in \mathcal{X}_j$, $j = 1, \ldots, p$와 같이 순서형 혹은 명목형으로 범주화 되어 있다고 가정한다. 그러면 단순 베이즈 가정은

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

로 표현할 수 있다.

$(\boldsymbol{x}_i, y_i), i = 1, \ldots, N$는 서로 독립적으로 $(\boldsymbol{X}, Y)$의 분포로부터 관측된 데이터라 하자. 각 $l \in \mathcal{Y}$에 대한 $Y = l$의 주변확률 및 각 $j(= 1, \ldots, p)$에 대한 $X_j = m_j$의 $Y = l$에 대한 조건부 확률은 훈련데이터에서 대응되는 상대도수

$$\hat{\mathbb{P}}(Y = l) = \frac{1}{N} \sum_{i=1}^{N} I(y_i = l),$$

$$\hat{\mathbb{P}}\left(X_j = m_j | Y = l\right) = \frac{\sum_{i=1}^{N} I(x_{ij} = m_j, y_i = l)}{\sum_{i=1}^{N} I(y_i = l)}$$

로 추정된다. 새로운 데이터 $\boldsymbol{x}$가 주어진 경우 Y에 대한 예측은

$$\arg\max_{l \in \mathcal{Y}} \prod_{j=1}^{p} \hat{\mathbb{P}}(X_j = m_j | Y = l) \hat{\mathbb{P}}(Y = l)$$

으로 한다.

클래스가 $l \in \mathcal{Y}$이고 $j\,(= 1, \ldots, p)$번째 입력변수의 값이 $m_j \in \mathcal{X}_j$인 경우의 관측도수는 $N_j(l, m_j) = \sum_{i=1}^{N} I(y_i = l, x_{ij} = m_j)$이며, $N_j(\cdot, m_j) = \sum_{l \in \mathcal{Y}} N_j(l, m_j)$과 $N_j(l, \cdot) = \sum_{m_j \in \mathcal{X}_j} N_j(l, m_j)$는 각각 주변 도수를 나타낸다. 그러면 기대도수는 $E_j(l, m_j) = N_j(l, \cdot) N_j(\cdot, m_j)/N$이다. 따라서 X_j와 출력변수 Y에 대한 카이제곱 통계량은

$$\chi_j^2 = \sum_{l \in \mathcal{Y}} \sum_{m_j \in \mathcal{X}_j} \frac{\left(N_j(l, m_j) - E_j(l, m_j)\right)^2}{E_j(l, m_j)}$$

으로 표현되며 근사적으로 자유도가 $(K-1)(d_j - 1)$인 카이제곱 분포를 따른다. 단 d_j는 $\mathcal{X}_j$의 원소의 개수를 나타낸다. S가 자유도가 df인 카이제곱 분포를 따르는 경우 피셔의 정규근사를 이용하면 $\sqrt{2S} - \sqrt{2df - 1}$는 근사적으로 표준정규분포를 따른다. 따라서 각 변수의 수준수가 다른 경우에

$$v_j = \left(\sqrt{2\chi_j^2} - \sqrt{2(K-1)(d_j - 1) - 1}\right)^2$$

를 구하여 비교할 수 있다. 이를 표준화된 카이제곱 통계량이라 하자.

주어진 훈련데이터를 $\mathcal{D}$이라 하자. $\mathcal{D}$를 미리 정한 비율(디폴트는 1 : 1)로 랜덤하게 둘로 나눈 데이터를 각각 $\mathcal{D}_T$, $\mathcal{D}_V$라 하자. 그러면 본 논문에서 제안하는 변수선택 절차는 다음과 같다.

1. $j = 1, \ldots, p$에 대하여 X_j와 Y간의 표준화된 카이제곱 통계량 v_j를 $\mathcal{D}_T$를 이용하여 구하고 $v_{(p)}, \ldots, v_{(1)}$과 같이 내림차순으로 정렬한다.
2. $v_{(p)}, \ldots, v_{(1)}$에 대하여 변화점 분석(change point analysis)을 실시하여 변화점들을 찾는다.
3. 각 변화점들에 대하여 대응되는 정렬된 카이제곱 통계량에 해당하는 변수로 이루어진 모형에 대하여 $\mathcal{D}_V$를 이용하여 검증오차(validation error)를 구한다.
4. 검증오차가 최소가 되는 변화점을 찾고 대응되는 변수들을 모형에 포함시켜 $\mathcal{D}$를 이용하여 최종모형을 적합한다.

위의 변수선택법은 범주형 입력변수를 갖는 경우에 단순 베이즈 분류 이외의 일반적인 분류문제에서도 적용이 가능하다. 본래 변화점 분석은 자료의 평균이나 분산 등의 통계적 성질이 변하는 지점을 찾아내기 위한 방법인데, 본 논문에서는 많은 수의 모형의 검증오차를 구하여 비교하는 것은 현실적으로 어렵기 때문에 검증오차를 비교할 적은 수의 후보모형을 찾기 위해 변화점 분석을 고려하였다. 구체적으로 말하면 카이제곱 통계량을 큰 순서대로 늘어 놓았을 때 값이 어느 시점을 벗어나면 통계량이 상대적으로 작은 값을 가지며 변동성이 적어질 것이라는 직관에 기초하여 분산에 대한 변화점 분석을 적용하였다. 변화점 분석 외에도 통계량에 대한 벌점화를 통한 모형 선택이나 모형선택기준을 이용하는 방법을 고려할 수도 있을 것이다.

Table 3.1. $\mathbb{P}(X_j = k|Y = l)$ for generating simulated data

j	$1 \sim 10$		$11 \sim 20$		$21 \sim 30$		$31 \sim 40$		$41 \sim 50$		$51 \sim p$	
$k \setminus l$	0	1	0	1	0	1	0	1	0	1	0	1
1	0.2	0.5	0.4	0.3	0.3	0.1	0.2	0.2	0.3	0.3	1/3	1/3
2	0.3	0.3	0.2	0.4	0.5	0.5	0.2	0.6	0.3	0.4	1/3	1/3
3	0.5	0.2	0.4	0.3	0.2	0.4	0.6	0.2	0.4	0.3	1/3	1/3

Table 3.2. Average test error rate(s.e.) before and after variable selection for simulated data.

p	N	Before	After
100	100	0.0324(0.0015)	0.0597(0.0031)
	500	0.0140(0.0003)	0.0172(0.0005)
	1000	0.0127(0.0003)	0.0147(0.0004)
500	100	0.1051(0.0016)	0.0818(0.0025)
	500	0.0271(0.0005)	0.0160(0.0005)
	1000	0.0186(0.0004)	0.0142(0.0003)
1000	100	0.1756(0.0031)	0.1117(0.0028)
	500	0.0430(0.0006)	0.0179(0.0006)
	1000	0.0267(0.0005)	0.0138(0.0003)

3. 데이터 분석

모든 데이터분석은 R을 이용하였고 변화점 분석은 `changepoint` 패키지의 `spt.var` 함수에서 제공하는 PELT(pruned exact linear time) 알고리즘을 적용하였다. 변화점 분석에 대한 전반적인 소개는 Chen과 Gupta (2000)를 참조하기 바란다. 또한 `changepoint` R 패키지 사용법과 PELT 알고리즘에 대한 자세한 사항은 각각 Killick과 Eckley (2014)와 Killick 등 (2012)를 참조하기 바란다.

3.1. 모의실험

모의실험에서는 변수 선택 기능이 없는 단순 베이즈 분류와 본 논문에서 제안하는 방법을 분류 정확도 및 변수 선택의 적절성 관점에서 비교하고자 한다. 모의실험의 데이터 생성모형은 다음과 같다. Y는 성공률이 0.5인 베르누이분포를 따르며 Y가 주어졌을 때 X_j들의 조건부 분포는 Table 3.1과 같다. 이 실험에서 p개의 입력변수들 중 앞의 50개는 설명력이 있는 신호변수이고 나머지 $p-50$개는 설명력이 없는 잡음변수이다.

본 모의실험에서는 훈련데이터의 크기와 차원에 따른 변수선택의 효과를 비교하기 위하여 N과 p를 각각 100, 500, 1000의 세 수준에서 실험하였고, 모형의 평가를 위한 시험데이터의 크기는 1000으로 고정하였다. 또한 실험의 변동성을 고려하여 데이터 생성, 모형적합, 시험오차의 계산 등 전과정을 100회 반복하였다.

Table 3.2는 p와 N의 각 수준에 대하여 변수선택을 하지 않은 경우와 변수선택을 한 경우의 시험오차를 요약한다. N이 고정되었을 때 $p = 100, 500, 1000$에 대하여 신호변수의 비율은 각각 50%, 10%, 5%로 점점 더 희박(sparse)해진다. 신호변수의 비율이 희박해질수록 변수선택은 시험오차를 더 많이 줄여주는 효과가 있고 따라서 변수선택이 의미가 있음을 알 수 있다. 반면 신호변수의 비율이 50%로 조밀한 경우에는 오히려 변수선택후에 시험오차가 증가하기도 한다. p가 고정된 경우를 보면 N이 증가함에 따라 변수선택을 한 경우와 하지 않은 경우 모두 시험오차가 줄어드는 경향이 보인다. 이는 표본의 개수가 늘어남에 따라 추정오차가 줄어들어 나타나는 현상으로 볼 수 있다.

Table 3.3. Performance measures(s.e.) of selectivity for simulated data.

p	N	True positive	False positive	# of selected variables
100	100	21.82(0.8403)	4.13(0.5244)	25.95(1.2860)
	500	33.58(0.5007)	0.15(0.0435)	33.73(0.5157)
	1000	36.78(0.3891)	0.01(0.0100)	36.79(0.3903)
500	100	21.03(0.9216)	38.42(6.0136)	59.45(6.7794)
	500	37.59(0.4765)	10.73(2.6543)	49.09(2.9387)
	1000	40.91(0.4874)	7.84(2.0019)	48.75(2.2806)
1000	100	18.25(0.9919)	70.33(11.7744)	88.58(12.6319)
	500	35.95(0.5746)	11.50(2.0459)	48.37(2.4162)
	1000	41.32(0.4054)	7.21(1.7895)	48.53(1.9954)

Table 3.3은 N과 p의 각 수준별 변수선택의 적절성과 관련된 측도인 TP(true positive), FP(false positive), 선택된 변수의 개수를 보여준다. 모든 N과 p의 수준에 대하여 신호변수의 개수는 항상 50으로 고정되어 있다. N이 고정된 경우 p가 증가함에 따라 TP는 큰 차이를 보이지 않지만 FP가 커지는 경향을 볼 수 있다. 이는 추정할 모수에 비해 상대적으로 데이터의 개수가 적어지면 잡음변수를 선택할 가능성이 증가하는 것으로 볼 수 있다. p가 고정된 경우 N이 증가하면 TP는 증가하는 반면 FP는 감소한다. $n = 1000$인 경우 설명력이 있는 변수들 중 대략 40개를 선택하는데 그 이유는 $X_{41}, \ldots, X_{50}$의 클래스별 분포가 잡음변수의 분포와 구별하기 어렵기 때문일 것으로 추측할 수 있다. 비록 변수선택의 일치성의 증명이나 광범위한 시뮬레이션을 통해 검증되지는 않았지만 모의실험 결과로부터 본 논문에서 제안하는 방법은 합리적인 변수선택을 함을 알 수 있다.

3.2. 실제데이터

변수 선택 기능이 없는 기존의 단순 베이즈 분류와 제안된 변수선택법의 성능을 다음의 데이터에 대하여 비교하였다.

- 은행 대출(bank loan): 국내 어느 은행의 대출 관련 데이터로 하재환, 박창이 (2009)에서 분석된 바 있다. 전체 데이터 수는 1920개이며 입력변수 27개 중 연속형과 범주형이 각각 23개, 4개이고 출력변수는 신용상태를 나타낸다.
- 스팸(spam): Hastie 등 (2009)의 데이터로 그 크기는 4601개이다. 총 57개의 입력변수들중 48개는 이메일에서 특정 단어의 발생 비율, 6개는 특정 문자가 나타나는 비율, 나머지 3개는 연속된 대문자열의 길이를 나타낸다. 출력변수는 스팸 메일 여부를 나타낸다.
- 접착접합 유전자 시퀀스(splice-junction gene sequence): UCI 기계학습 저장소(UCI machine learning repository)의 데이터로서 고등생물의 DNA에 있는 유전자에 의해 단백질이 생성되는 과정과 관련된다. 접합효소(splicing enzyme)는 유전자 조합으로부터 쓸모없는 부분인 인트론(intron)을 제거하고 유전정보가 들어있는 부분인 엑손(exon)을 연결하는 역할을 한다. 데이터의 크기는 3190개이며 입력변수는 60개의 범주형 변수들로 이루어져 있다. 출력변수는 DNA 유전자 결합구조에 따라 그 경계를 EI, IE로 구분한 값이다.
- 은행 마케팅(bank marketing): 포르투갈 금융기관의 직접 마케팅 캠페인의 결과 고객의 금융상품(은행 단기 예금)의 구매 여부에 대한 데이터로 출처는 UCI 기계학습 저장소이다. 전체 데이터의 크기는 4521개이며, 연속형 7개와 범주형 10개의 총 17개의 입력변수로 구성되어 있고, 출력변수는 고객의 구매여부를 나타낸다.

Table 3.4. Results before and after variable selection for real data sets

Data	Average error rate (s.e.)		# of selected
	Before	After	variables (s.e.)
bank loan	0.2677(0.0016)	0.2761(0.0017)	13.9600(0.2117)
spam	0.1078(0.0008)	0.1107(0.0010)	16.5900(0.6911)
splice-junction gene sequence	0.0449(0.0007)	0.0501(0.0015)	19.3100(0.4556)
bank marketing	0.2936(0.0026)	0.2801(0.0027)	2.6800(0.1595)

본 논문에서는 모든 입력변수가 범주형인 경우만을 고려하므로 연속형 변수들은 사분위수를 이용하여 적절히 범주화를 하였다. 참고로 연속형 변수에 대한 범주화는 Jin 등 (2012)에서 처럼 상수 스플라인(constant spline)에 대한 매듭점(knot) 선택을 통해 접근할 수 있으며 사분위수를 이용한 범주화 등 연속형 변수의 범주화에 대한 이슈들은 Jin 등 (2012) 및 인용 논문들을 참고하기 바란다.

각 데이터에 대하여 7 : 3의 비율로 훈련데이터와 시험데이터로 랜덤하게 분할한 후 모형의 적합 및 평가를 하였다. 또한 실험의 변동성을 고려하여 이러한 전 과정을 100회 반복 시행하였다.

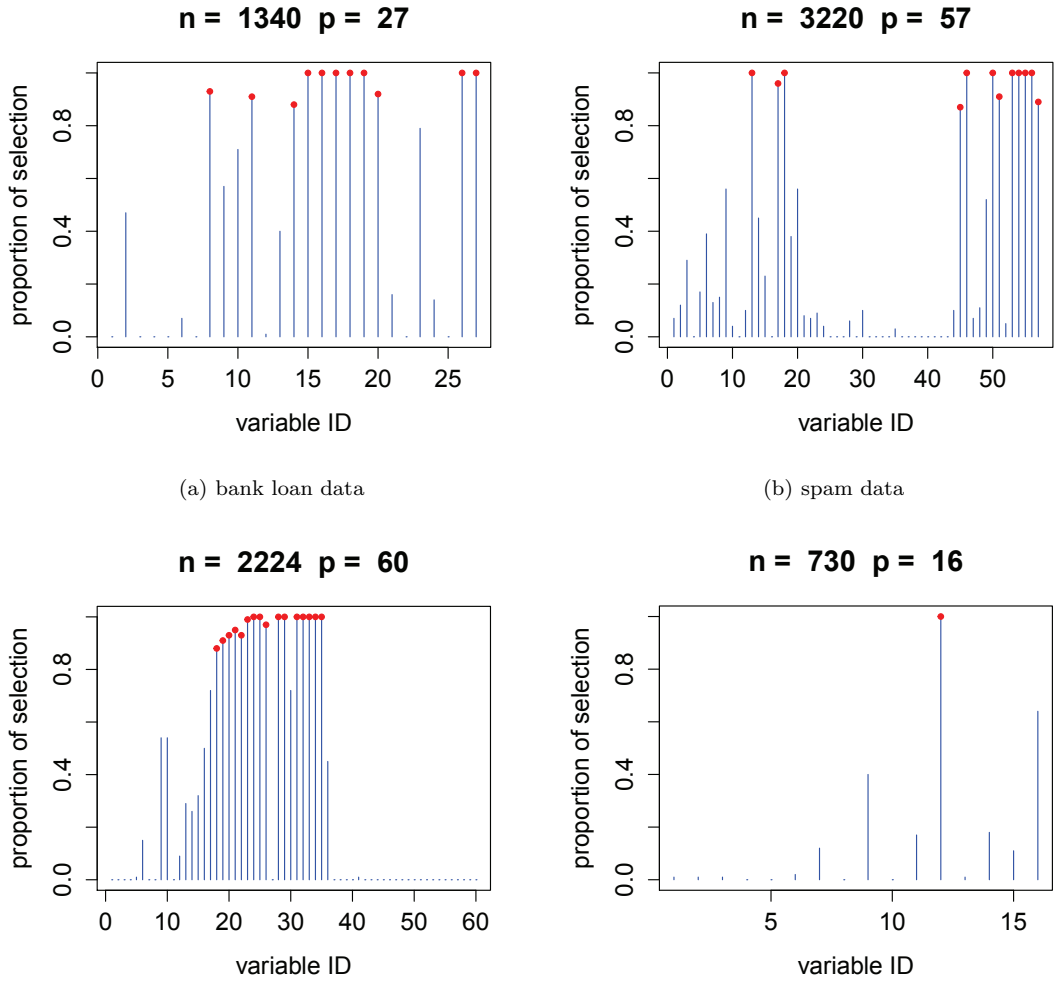
Table 3.4는 네 가지 실제 데이터에 대하여 변수선택을 하기 전과 후의 평균 오분류율과 선택된 변수들의 개수들을 요약한다. 데이터에 따라 변수선택을 한 경우 오분류율이 감소한 경우도 있고 그렇지 않은 경우도 있는데 그 차이는 그리 크지는 않다. 변수선택에 따른 예측력의 차이는 데이터의 특성에 따라 달라질 것으로 생각된다. 적어도 예측력이 심각하게 떨어지지 않으면서도 변수의 개수를 줄여준다는 점에서 본 논문에서 제안된 변수선택은 최종모형의 해석력을 향상 시킨다고 볼 수 있다.

마지막으로 Figure 3.1은 각 데이터에 대하여 각 변수에 대하여 100번의 반복중 선택된 비율을 보여준다. 80%이상 선택된 변수들은 설명력이 있는 변수들로 간주하여 비율값 끝부분에 점을 찍어 표시하였다. 은행 대출 데이터의 경우 기업 데이터이므로 자세히 설명할 수는 없지만 대출관련 조회 건수와 연체와 관련된 변수들이 고객의 신용등급을 분류하는데 설명력이 있는 변수들로 선택되었다. 스팸 데이터에서는 점으로 표시된 변수들은 해당되는 문자의 발생 비율 및 길이가 스팸 메일 여부와 관련이 큰 것으로 생각할 수 있다. 접착접합 유전자 시퀀스 데이터의 경우 60개의 입력변수들 중 점으로 표시된 16개의 변수들에 해당하는 DNA 유전자 결합구조가 인트론과 엑손의 경계를 구분하는데에 영향을 미친다고 볼 수 있다. 은행 마케팅 데이터의 경우 12번째 변수인 캠페인의 지속시간이 고객의 제품 구매여부에 영향을 주는 변수로 나타났다.

4. 결론

본 연구에서는 단순 베이즈 분류에서 카이제곱 통계량을 이용한 범주형 입력변수의 선택법을 소개하고 모의실험과 실제 데이터에 대한 분석을 통해 기존의 단순 베이즈 분류와 본 논문의 변수선택법의 예측력과 해석력을 비교하였다. 본 논문에서 제안한 방법은 경우에 따라서는 예측력을 향상시킬 수도 있으며 변수선택을 통해 적어도 해석력은 향상시키는 것을 확인할 수 있었다. 본 논문의 변수선택법은 단순 베이즈 분류의 최대의 장점인 데이터 처리 및 계산의 단순성을 유지하므로 SNP에 의한 질병의 분류 등 초고차원 또는 빅데이터의 분류문제에 유용할 것으로 기대된다.

본 연구와 관련된 후속 연구에 대한 제언으로는 다음과 같은 것들이 있다. 첫째, 단순 베이즈 분류의 경우 결합확률은 클래스 변수의 주변확률과 클래스 변수가 주어졌을 때 입력변수들의 조건부 확률의 곱으로 표현되므로 클래스 불균형(class imbalance)이 심각한 데이터에서는 적용하기 힘들다. 따라서 클래스 불균형이 심각한 데이터에서 단순 베이즈 분류의 적용이 가능하도록 하는 방안에 대한 연구가 필요하



(a) bank loan data

(b) spam data

(c) splice-junction gene sequence data

(d) bank marketing

Figure 3.1. Proportions of variables selected on real data sets.

리라고 생각된다. 둘째, 단순 베이즈 모형에서 2차 교호작용을 허용하는 모형의 개발을 고려할 수 있다. 입력변수들을 그래프의 노드로 보면 노드들 간의 간선(edge)의 유무를 추정하는 그래프 모형(graphical model)과 관련된 단순 베이즈 분류 기반의 방법을 생각해 볼 수 있을 것이다. 본 논문에서 제안한 방법은 일종의 후진 선택법(backward selection)으로 p개의 주효과 변수들만 다루므로 계산에 큰 무리가 없지만, 교호작용이 있는 경우에는 $p(p-1)/2$개의 교호작용이 존재하므로 단계적 전진 선택법(forward stagewise selection) 형식의 효율적인 계산 알고리즘이 필요할 것으로 기대된다. 셋째, 제안한 방법은 변수선택이 변화점 분석 결과에 크게 영향 받을 수 있으므로 여러가지 변화점 분석법에 대한 비교, 벌점화를 통한 모형 선택 혹은 모형선택기준을 이용하는 방법을 고려할 수도 있을 것이다. 또한, FDR(false discovery rate)관점에서 유의확률을 이용하는 방법을 고려할 수 있는데 범주형 변수들의 수준수가 상이

한 경우나 연속형 변수가 혼재된 경우에도 손쉽게 적용할 수 있다는 장점이 있다. 유의확률을 이용한 방법과 제안된 방법의 비교도 흥미로울 것이다.

References

Chen, J. and Gupta, A. K. (2000). *Parametric Statistical Change Point Analysis*, Birkhauser.

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society*, **70**, 849–911.

Ha, J. H. and Park, C. (2009). Variable selection in linear discriminant analysis, *Journal of the Korean Data Analysis Society*, **11**, 381–389.

Hand, D. and Yu, K. (2001). Idiot's Bayes-not so stupid at all?, *International Statistical Review*, **69**, 385–399.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, (2nd Edition), Springer, New York.

Jin, S. K., Kim, K.-R. and Park, C. (2012). Cutpoint Selection via penalization in credit scoring, *The Korean Journal of Applied Statistics*, **25**, 261–267.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost, *Journal of the American Statistical Association*, **107**, 1590–1598.

Killick, R. and Eckley, I. A. (2014). Changepoint: An R package for changepoint analysis, *Journal of Statistical Software*, **58**.

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

Vidaurre, D., Bielza, C. and Larrañaga, P. (2013). An L_1-regularized naïve Bayes-inspired classifier for discarding redundant and irrelevant predictors, *International Journal on Artificial Intelligence Tools*, **22**, 1350019.

단순 베이즈 분류에서의 범주형 변수의 선택

김민선[a] · 최호식[b] · 박창이[a,1]

[a]서울시립대학교 통계학과, [b]경기대학교 응용정보통계학과

(2015년 1월 20일 접수, 2015년 3월 3일 수정, 2015년 3월 10일 채택)

요 약

단순 베이즈 분류(Naïve Bayes classification)는 출력변수가 주어졌을 때 입력변수들이 조건부 독립이라는 가정에 기반한다. 단순 베이즈 가정은 비현실적이지만 고차원의 확률 추정 문제를 일련의 일차원 확률 추정 문제로 단순화 시킨다는 장점이 있으며, 특히 스팸 메일 필터링, 추천 시스템(recommendation system) 등 방대한 데이터를 다루는 분야야에서 흔히 사용된다. 본 논문에서는 입력변수와 출력변수간의 카이제곱 통계량에 기반한 변수선택법을 제안한다. 이 방법은 단순 베이즈 분류의 장점인 데이터 처리 및 계산의 단순성을 유지하면서도 설명력이 있는 변수를 선택할 수 있으며 SNP(single nucleotide polymorphism)에 의한 질병의 분류 등의 초고차원 혹은 빅데이터에서 유용할 것으로 기대된다.

주요용어: 빅 데이터, 카이제곱 통계량, 단순 베이즈 가정, SNP.

이 논문은 2014년도 서울시립대학교 교내학술연구비에 의하여 지원되었음.

[1]교신저자: (130-743) 서울특별시 동대문구 전농동 90, 서울시립대학교 통계학과. E-mail: park463@uos.ac.kr