Statistics and Its Interface Volume 6 (2013) 223–229

Testing the statistical significance of an ultra-high-dimensional naïve Bayes classifier

Baiguo An, Hansheng Wang* and Jianhua Guo

The naïve Bayes approach is one of the most popular methods used for classification. Nevertheless, how to test its statistical significance under an ultra-high-dimensional (UHD) setup is not well understood. To fill this important theoretical gap, we propose a novel testing statistic with a standard normal asymptotic null distribution, even if the predictor dimension is considerably larger than the sample size. This makes the proposed method useful for UHD data analysis. Simulation studies are presented to demonstrate its finite sample performance and a text classification example is described for illustration.

AMS 2000 subject classifications: Primary 62H30.
Keywords and phrases: Binary Predictor, Hypothesis Testing, Naïve Bayes, Supervised Learning, Text Classification, Ultra-High-Dimensional Data.

1. INTRODUCTION

Supervised classification is widely used in real-world applications [8], including medical diagnoses [17], handwriting recognition [12], and web mining [11], among many others. Various supervised classification methods have been developed for such important applications. These include linear and quadratic discriminant analysis (LDA and QDA), logistic regression, nearest-neighbor methods, the naïve Bayes (NB) approach, support vector machines (SVMs), and many others. Among these, NB is ranked in the top ten most popular classification methods in practice [14, 18].

In theory, the NB method is applicable to situations with either continuous or binary predictors [2]. However, in this study we focus on the binary case because our work is mainly motivated by a text classification problem in which a text document is typically represented by a high-dimensional binary vector; see Example 4 in Section 3.3 for more details. Thus, we are interested in testing whether there exists at least one binary predictor related to the response category. Such a test is standard in classical linear regression analysis (i.e., the overall F-test) and is typically carried out before investigating the statistical significance of each individual predictor for good control of the family-wise error [15].

Conditional on the response category, the NB approach assumes that different predictors are mutually independent. Consequently, the statistical significance of a predictor can be tested by investigating its marginal relationship with the response variable. This idea is similar in approach to sure independent screening [3]. When the predictor is binary, this amounts to construction of a two-way contingency table and testing of its row and column independence [9]. Under a classical setup with a large sample size (denoted by n) and a fixed predictor dimension (denoted by p), this can be carried out via a standard chi-square test and has been widely used in practice [1, 16, 19].

However, in scientific research it is common to encounter situations in which the sample size n is very limited but the predictor dimension p can be very large. The limited sample size may be because of budget limitations, timing constraints, or other practical reasons. By contrast, the large predictor dimension p might be because of technological advances that enable researchers to collect a large amount of information quickly and economically. Thus, how to test the statistical significance of a NB method under an ultra-high-dimensional (UHD) setup is of interest. To fill this important theoretical gap, we propose a novel testing statistic with a standard normal asymptotic null distribution, even if the predictor dimension is much greater than the sample size. Thus, the proposed method is useful for UHD data analysis. Extensive simulation studies are presented to demonstrate its finite sample performance and a real text classification example is described for illustration.

The remainder of the article is organized as follows. Section 2 presents the model and the notation used. The test statistic and its theoretical properties are also described. Section 3 presents numerical studies, which include simulation studies and a real data example. Section 4 concludes with a short discussion. All the proofs are contained in the Appendix.

2. METHODOLOGY

2.1 Model and notation

Let (Y_i, X_i) be the observation for the ith subject ($1 \leq i \leq n$), where $Y_i \in \{0, 1\}$ is the response variable (or class label) and $X_i = (X_{i1}, \cdots, X_{ip})^\top \in \mathbb{R}^p$ is the associated p-dimensional binary predictor, i.e., $X_{ij} \in \{0, 1\}$ for every $1 \leq j \leq p$. We also assume that (Y_i, X_i) for $1 \leq i \leq n$ are independent and identically distributed. Furthermore, for a fixed i conditional on Y_i, the NB method requires different

*Corresponding author.

X_{ij}s to be mutually independent. Next, for every $k \in \{0,1\}$ and $l \in \{0,1\}$, we define $\pi_{kl}^j = P(Y_i = k, X_{ij} = l)$, $\pi_{k\cdot} = P(Y_i = k)$, $\pi_{\cdot l}^j = P(X_{ij} = l)$. If no predictor is related to the response, then the following null hypothesis should be correct.

$$\pi_{kl}^j = \pi_{k\cdot} \times \pi_{\cdot l}^j \quad \text{for every } k,\ l \text{ and } j. \tag{1}$$

For a given data set, these quantities can be empirically estimated by $\hat{\pi}_{kl}^j = n^{-1}\sum I(X_{ij} = l)I(Y_i = k)$, $\hat{\pi}_{k\cdot} = n^{-1}\sum I(Y_i = k)$, and $\hat{\pi}_{\cdot l}^j = n^{-1}\sum I(X_{ij} = l)$, respectively. It is then natural to consider the following test statistic

$$T_n = n\sum_{j=1}^{p}\sum_{k=0}^{1}\sum_{l=0}^{1}\frac{(\hat{\pi}_{kl}^j - \hat{\pi}_{k\cdot}\hat{\pi}_{\cdot l}^j)^2}{\hat{\pi}_{k\cdot}\hat{\pi}_{\cdot l}^j}. \tag{2}$$

Under the assumption that p is fixed but $n \to \infty$, T_n is asymptotically distributed as a chi-square distribution with p degrees of freedom [1]. Our numerical experience suggests that it is indeed a good approximation to the empirical distribution if the sample size is large and the predictor dimension is small. However, if the predictor dimension is large but the sample size is limited, a chi-square distribution is no longer a good approximation. This motivates us to develop a new test statistic that works better for cases with large p but limited n.

2.2 New test statistic

To solve the problem, we develop a new testing procedure. More specifically, we decompose the total sample size as $n = n_0 + n_1$, where n_k $(k = 0, 1)$ is the sample size associated with $Y_i = k$, that is, $n_k = \sum_{i=1}^{n} I(Y_i = k)$. We then define $\alpha_{kj} = P(X_{ij} = 1|Y_i = k)$ with $k \in \{0,1\}$. By definition, we know that $\alpha_{kj} = \pi_{k1}^j/\pi_{k\cdot}$. Then the null hypothesis (1) can be rewritten as

$$H_0 : \alpha_{0j} = \alpha_{1j} = \alpha_j \text{ for some } \alpha_j \text{ and every } j. \tag{3}$$

It is clear that α_{kj} can be estimated by $\hat{\alpha}_{kj} = n_k^{-1}\sum_{i=1}^{n} I(Y_i = k)X_{ij}$. We then consider the following rather simple test statistic:

$$T = \sum_{j=1}^{p}(\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^2. \tag{4}$$

Intuitively, if the null hypothesis (3) is indeed correct, the value of T is expected to be relatively small. Consequently, we can reject the null hypothesis (3) if the value of T is sufficiently large.

2.3 Mean-variance analysis

Although the basic idea of (4) is intuitive, it is not clear how large the value of T needs to be considered sufficiently large in practice. To address this issue, we conduct a mean-variance analysis for T.

Theorem 2.1. *Under the null hypothesis (3) and assuming $n \to \infty$, we have (1) $E(T) = (1/n_0 + 1/n_1)\sum_{j=1}^{p}\alpha_j(1-\alpha_j)$ and (2) $var(T) = 2(1/n_0 + 1/n_1)^2\sum_{j=1}^{p}\alpha_j^2(1-\alpha_j)^2 + o_p(pn^{-2})$.*

The proof is given in Appendix A. Note that $E(\cdot)$ denotes the expectation conditional on n_0 and n_1. The notation $\mathrm{var}(\cdot)$ is defined similarly. Theorem 2.1 leads to an important theoretical finding. Consider a highly simplified case with $n_0 = n_1 = n/2$ and $n/p \to 0$. Then, by Theorem 2.1 we should have that $E(T)$ is of order p/n and $\mathrm{var}(T)$ is of order p/n^2. As a result, the order of $E(T)/\mathrm{var}^{1/2}(T)$ should be $\sqrt{p} \to \infty$. Hence, it is impossible for $T/\mathrm{var}^{1/2}(T)$ to follow any non-degenerate probability distribution. The main reason is the bias of T. Consequently, before we can use T for hypothesis testing, appropriate bias correction is required.

2.4 Bias-corrected test statistic

Motivated by the theoretical findings in the previous subsection, we consider a bias-corrected test statistic given by

$$T_c = \sum_{j=1}^{p}\Big\{(\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^2 - \Big(\frac{1}{n_0} + \frac{1}{n_1}\Big)\Big(\frac{n}{n-1}\Big)\hat{\alpha}_j(1-\hat{\alpha}_j)\Big\}, \tag{5}$$

where $\hat{\alpha}_j = \sum X_{ij}/n$ is an estimator for α_j in the null hypothesis (3). We can easily verify that the new test statistic is exactly unbiased for 0 if the null hypothesis (3) is indeed correct. We next consider its asymptotic variance, which is given by the next theorem.

Theorem 2.2. *Under the null hypothesis (3) and assuming $n \to \infty$, we have (1) $ET_c = 0$ and (2) $var(T_c) = 2(1/n_0 + 1/n_1)^2\sum_{j=1}^{p}\alpha_j^2(1-\alpha_j)^2 + o_p(pn^{-2})$.*

The proof is given in Appendix B. By Theorem 2.2, we know that the bias-corrected test statistic T_c shares the same asymptotic variance as T. We can then consider use of $T_c/\mathrm{var}^{1/2}(T_c)$ as a test statistic because $\mathrm{var}(T_c)$ involves the unknown parameter α_j, which can be replaced by its estimator $\hat{\alpha}_j$. This leads to the final test statistic $T_f = T_c/\widehat{D}_p$, where $\widehat{D}_p^2 = 2(1/n_0+1/n_1)^2\sum_{j=1}^{p}\hat{\alpha}_j^2(1-\hat{\alpha}_j)^2$. The next theorem shows that such a test statistic has a standard normal asymptotic distribution if the null hypothesis (3) is indeed true.

Theorem 2.3. *Under the null hypothesis (3), assume that there exist $\alpha_{\min}$ and $\alpha_{\max}$ such that*

$$0 < \alpha_{\min} \le \alpha_j \le \alpha_{\max} < 1 \quad \textit{for every } j. \tag{6}$$

Further assume that $p/n^2 \to \infty$. We then have $T_f \to_d N(0,1)$ as $n \to \infty$, where "$\to_d$" denotes convergence in distribution.

Table 1. Example 1. Empirical size based on 1,000 simulation replicates

Sample size	Test statistic	Predictor dimension 100	200	500	1,000	5,000	40,000
50	T_f	0.053	0.039	0.054	0.047	0.043	0.0550
	T_n	0.068	0.058	0.075	0.125	0.291	0.8680
100	T_f	0.049	0.051	0.048	0.045	0.049	0.0510
	T_n	0.064	0.059	0.057	0.074	0.122	0.4110
200	T_f	0.052	0.051	0.062	0.060	0.051	0.0530
	T_n	0.054	0.067	0.070	0.072	0.087	0.1770

The proof is given in Appendix C. By Theorem 2.3 we know that the asymptotic null distribution of T_f is a standard normal distribution. As a result, we can reject the null hypothesis (3) if $|T_f| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$th percentile of a standard normal distribution.

Remark 1. Condition (6) is a very reasonable assumption. It can be violated if the frequency observed for some predictor is extremely high or low. In either case, such predictors carry little information about the response variable Y_i, and thus should be excluded from the formal analysis [10]. Otherwise, the asymptotic null distribution of T_f might be seriously distorted. Consider for example an extreme situation in which $\alpha_j = 1$ for every j. In this case, we would have both T_c and $\hat{D}_p$ equal to 0. Consequently, T_f becomes an invalid test statistic and the asymptotic normality theory as claimed in Theorem 2.3 is no longer applicable.

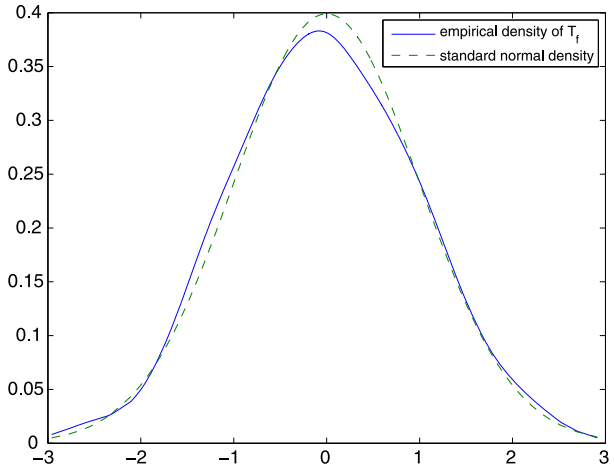

Figure 1. Empirical density of T_f compared to the standard normal density.

3. NUMERICAL STUDY

To demonstrate the finite sample performance of the proposed testing method, we present a number of numerical studies.

3.1 Size and power

Example 1. This example focuses on the empirical size of the proposed test statistic T_f. For comparison, the naïve test statistic T_n is also evaluated. The response Y_i is generated from a binary distribution with $P(Y_i = 1) = 0.5$. Regardless of the value of Y_i, X_{ij} is independently generated with $P(X_{ij} = 1) = \pi_{\cdot j}$, where $\pi_{\cdot j}$ is randomly simulated from a uniform distribution on $[0.3, 0.8]$. Because X_{ij}s are generated independently with respect to Y_i, we know that the null hypothesis (3) is true. The experiment is randomly replicated 1,000 times. The nominal level is fixed to 0.05 and results for the empirical sizes are listed in Table 1. The empirical size results for T_f are fairly close to the nominal level 0.05, which corroborates our asymptotic theory quite well. By contrast, those for the naïve test statistic T_n show large deviation from 0.05. This is particularly true for cases with small sample size and high dimension. To gain an intuitive understanding of the asymptotic distribution of T_f, the empirical density for $n = 200$ and $p = 5,000$ is compared with the standard normal distribution in Figure 1. The distributions are very close to each other, which confirms that the asymptotic null distribution of T_f is indeed a standard normal distribution.

Example 2. In this example, we numerically investigate the power of the proposed test statistic T_f. To this end, we fix an integer $p_0 = 10$ and for each $1 \le j \le p_0$ we simulate X_{ij} from a Bernoulli distribution with mean π_{kj}, given $Y_i = k$ with $k = 0$ or 1. Here π_{0j} is generated from a uniform distribution on $[0.2, 0.4]$, while π_{1j} is generated from a uniform distribution on $[0.6, 0.8]$. Next, for every $p_0 < j \le p$, X_{ij} is simulated from a Bernoulli distribution with parameter π_j, where π_j is generated from a uniform distribution on $[0.2, 0.8]$. Obviously, the null hypothesis (3) is violated for this case. Then, for each parameter specification for n and p, we replicate the experiment 1,000 times. Results for the empirical power with significance level $\alpha = 0.05$ are listed in Table 2. It is clear that for a given sample size, a larger predictor dimension leads to smaller power. This is expected because the number of relevant predictors is fixed to $p_0 = 10$. Thus, a larger predictor dimension introduces a greater number of irrelevant predictors, which decreases the power of the proposed test. Conversely, if the predictor dimension is fixed, a larger sample size leads to better power, as expected.

Table 2. Example 2. Empirical powers based on 1,000 simulation replicates

Sample size	Predictor dimension					
	500	1,000	1,500	2,000	2,500	3,000
50	0.775	0.721	0.393	0.376	0.290	0.211
100	0.997	0.997	0.973	0.742	0.742	0.730
200	1.000	1.000	1.000	0.994	0.996	0.996

Table 3. Example 3. Number of important variables missed, number of irrelevant variables selected, and prediction accuracy for various sample sizes and significance levels α

Sample Size	Performance Measure	α Specification					
		0.01	0.05	0.10	0.50	0.75	1
50	Missed variables	14	12	12	8	7	0
	Irrelevant variables	0	1	2	3	3	980
	Prediction accuracy	0.857	0.882	0.864	0.896	0.911	0.656
100	Missed variables	10	9	8	6	6	0
	Irrelevant variables	0	2	2	6	8	980
	Prediction accuracy	0.933	0.923	0.936	0.931	0.922	0.793
200	Missed variables	8	6	6	4	4	0
	Irrelevant variables	0	0	0	1	2	980
	Prediction accuracy	0.963	0.969	0.969	0.974	0.970	0.906

3.2 Variable selection and prediction accuracy

Example 3. In this example, we demonstrate some potential applications of our method for variable selection. We also evaluate its effect on the prediction accuracy. We fix $p = 1,000$ in this experiment. For a fixed sample size n, we simulate each Y_i from a Bernoulli distribution with $P(Y_i = 1) = 0.5$. Given the value of $Y_i = k$, we simulate X_{ij} for $1 \le j \le p_0 = 20$ from a Bernoulli distribution with $P(X_{ij} = 1) = \pi_{kj}$. Here π_{0j} is generated from a uniform distribution on $[0.2, 0.4]$, while π_{1j} is generated from a uniform distribution on $[0.6, 0.8]$. For $j > p_0 = 20$, X_{ij} is always simulated from a Bernoulli distribution with $P(X_{ij} = 1) = 0.5$, regardless of the value of Y_i. As a result, only the first $p_0 = 20$ predictors are relevant to the response prediction. This gives us the training data.

For variable selection, we first compute a marginal chi-square test statistic for each predictor as $\chi_j^2 = n\sum_{k=0}^{1}\sum_{l=0}^{1}(\hat{\pi}_{kl}^j - \hat{\pi}_{k\cdot}\hat{\pi}_{\cdot l}^j)^2/(\hat{\pi}_{k\cdot}\hat{\pi}_{\cdot l}^j)$. Using a sure independent screening method [3], we then rank the importance of each predictor in decreasing χ_j^2 order. This gives us a solution path, denoted as $j_1, j_2, \ldots, j_p$, where $\{j_1, j_2, \ldots, j_p\} = \{1, 2, \ldots, p\}$. If the first k $(1 \le k \le p)$ predictors are sufficient for response prediction, the test statistic computed based on the remaining predictors should follow the standard normal distribution asymptotically. As a result, the resulting p-value is expected to be nonsignificant. By contrast, if the first k predictors are insufficient for response prediction, the resulting p-value computed based on the remaining predictors should be significant. Thus, the best model can be specified as $\mathcal{M} = \{j_1, j_2, \ldots, j_k\}$, where $(k+1)$ is the first integer for which the test statistic T_f is nonsignificant.

Depending on the choice of α, the resulting models can differ and their out-of-sample forecasting accuracy might be different. To gain an intuitive idea of this, the out-of-sample forecasting accuracy was evaluated for testing data with a sample size of $1,000$. The results are given in Table 3. It is evident that the model size increases with α. As a result, the number of important variables missed by our method decreases but the number of irrelevant variables selected increases. If the ultimate objective is forecasting, it seems that $\alpha = 0.50$ is a reasonable choice. Such a phenomenon is not surprising, because the asymptotic null distribution of the empirical p-value is uniform on $[0, 1]$. As a result, for a p-value less than 0.50, it is more likely that some relevant variables will not be included.

3.3 Text mining example

Example 4. We apply the method to a Chinese text data set for 1,119 text documents. Each document represents one appeal phone call made to the Mayor's public hotline (MPH). The appeals are classified into two categories according to the functional department involved, giving sample sizes of 550 and 569. For a more detailed description about MPH, one can refer to [4]. We then represent each document by a UHD binary vector $X_i = (x_{i1}, \ldots, x_{ip})$ with $p = 15,759$, where $X_{ij} = 1$ if the jth keyword exists in the ith document, and 0 otherwise [13]. A total of $n_0 = n_1 = 100$ documents are randomly selected from each category as the training data, while the rest are reserved for testing. We follow the same technique as in the previous simulation study

to select important variables based on the training data set. A NB classifier is then constructed and its prediction accuracy is evaluated for the test data. For reliable evaluation, the experiment was randomly replicated 1,000 times. The resulting average prediction accuracy is 0.9201. In comparison, classification and regression tree [8] and linear support vector machine [10] methods gave prediction accuracy of 0.9292 and 0.9181, respectively. These values are similar to the NB accuracy. However, the NB method is much simpler, both computationally and theoretically.

4. DISCUSSION

We proposed a novel procedure for testing the statistical significance of a UHD NB model. Our numerical experience suggests that the proposed method is particularly useful for UHD data. However, it is not clear if the proposed test statistic is optimal under certain alternative hypotheses. To the best of our knowledge, well-established optimality results for high-dimensional tests are extremely limited in literature, although some pioneer results for regression models have been published [5–7]. Further research into this issue is required.

APPENDIX A. PROOF OF THEOREM 2.1.

We consider first Theorem 2.1 (1). We can easily verify that $E\hat{\alpha}_{kj} = \alpha_j$ and $E\hat{\alpha}_{kj}^2 = n_k^{-1}\alpha_j(1-\alpha_j) + \alpha_j^2$. Consequently, we have $E(\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^2 = E\hat{\alpha}_{0j}^2 + E\hat{\alpha}_{1j}^2 - 2E\hat{\alpha}_{0j}E\hat{\alpha}_{1j} = (1/n_0+1/n_1)\alpha_j(1-\alpha_j)$. This leads to $E(T) = \sum_{j=1}^p E(\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^2 = (1/n_0 + 1/n_1)\sum_{j=1}^p \alpha_j(1-\alpha_j)$.

We next consider Theorem 2.1 (2). We can verify that $E\hat{\alpha}_{kj}^4 = n_k^{-3}\alpha_j\{(n_k-1)(n_k-2)(n_k-3)\alpha_j^3 + 6(n_k-1)(n_k-2)\alpha_j^2 + 7(n_k-1)\alpha_j + 1\}$, $E(\hat{\alpha}_{0j}^2\hat{\alpha}_{1j}^2) = n_0^{-1}n_1^{-1}\alpha_j^2\{(n_0-1)\alpha_j+1\}\{(n_1-1)\alpha_j+1\}$, and for $k \neq k'$, $E(\hat{\alpha}_{kj}^3\hat{\alpha}_{k'j}) = n_k^{-2}\alpha_j^2\{(n_k-1)(n_k-2)\alpha_j^2 + 3(n_k-1)\alpha_j + 1\}$. We then have $E(\hat{\alpha}_{0j}-\hat{\alpha}_{1j})^4 = \{3(n_0^{-1}+n_1^{-1})^2 - 6(n_0^{-3}+n_1^{-3})\}(\alpha_j^4 - 2\alpha_j^3 + \alpha_j^2) + (n_0^{-3}+n_1^{-3})(\alpha_j - \alpha_j^2)$. This leads to

$$\begin{aligned}
\mathrm{var}(T) &= \sum_{j=1}^p \mathrm{var}(\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^2 \\
&= \sum_{j=1}^p E(\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^4 - \sum_{j=1}^p \left\{E(\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^2\right\}^2 \\
&= \left\{2(n_0^{-1}+n_1^{-1})^2 - 6(n_0^{-3}+n_1^{-3})\right\}\sum_{j=1}^p \alpha_j^2(1-\alpha_j)^2 \\
&\quad + (n_0^{-3}+n_1^{-3})\sum_{j=1}^p \alpha_j(1-\alpha_j).
\end{aligned}$$

By the law of large numbers, we have $n_k/n \to_p \pi_k$. for $k = 0, 1$. This, together with the fact that $\alpha_j^2(1-\alpha_j)^2 < 1$, implies that $(n_0^{-3}+n_1^{-3})\sum_{j=1}^p \alpha_j^2(1-\alpha_j)^2 \le p(n_0^{-3}+n_1^{-3}) = pn^{-2}o_p(1)$. Consequently, $(n_0^{-3}+n_1^{-3})\sum_{j=1}^p \alpha_j^2(1-\alpha_j)^2 = o_p(pn^{-2})$. Similarly, it is also true that $(n_0^{-3}+n_1^{-3})\sum_{j=1}^p \alpha_j(1-\alpha_j) = o_p(pn^{-2})$. Hence, we have $\mathrm{var}(T) = 2(1/n_0+1/n_1)^2\sum_{j=1}^p \alpha_j^2(1-\alpha_j)^2 + o_p(pn^{-2})$. This completes the proof.

APPENDIX B. PROOF OF THEOREM 2.2.

We consider first Theorem 2.2(1). We know that $E(\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^2 = (1/n_0+1/n_1)\alpha_j(1-\alpha_j)$ for every $1 \le j \le p$. Hence, it suffices to show that $E\{\hat{\alpha}_j(1-\hat{\alpha}_j)\} = (n-1)n^{-1}\alpha_j(1-\alpha_j)$. Under the null hypothesis (3), we know that $E(\hat{\alpha}_j) = \alpha_j$ and $E(\hat{\alpha}_j^2) = \alpha_j^2 + n^{-1}\alpha_j(1-\alpha_j)$. This proves $E\{\hat{\alpha}_j(1-\hat{\alpha}_j)\} = (n-1)n^{-1}\alpha_j(1-\alpha_j)$.

We next consider Theorem 2.2 (2). Denote $(\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^2 - (1/n_0 + 1/n_1)n(n-1)^{-1}\hat{\alpha}_j(1-\hat{\alpha}_j)$ by T_j. We then have $T_c = \sum_{j=1}^p T_j$ and $\mathrm{var}(T_c) = \sum_{j=1}^p \mathrm{var}(T_j) = \sum_{j=1}^p E(T_j^2)$. We next investigate $E(T_j^2)$ as

$$\begin{aligned}
T_j^2 &= (\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^4 + A^2(\hat{\alpha}_j^4 - 2\hat{\alpha}_j^3 + \hat{\alpha}_j^2) \\
&\quad - 2A(\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^2(\hat{\alpha}_j - \hat{\alpha}_j^2),
\end{aligned} \tag{7}$$

where $A = (1/n_0 + 1/n_1)n(n-1)^{-1}$. We then evaluate the above three terms separately. For the first term, following similar techniques as in the proof of Theorem 2.1, we have $E(\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^4 = \{3(n_0^{-1}+n_1^{-1})^2 - 6(n_0^{-3}+n_1^{-3})\}\alpha_j^2(1-\alpha_j)^2 + (n_0^{-3}+n_1^{-3})\alpha_j(1-\alpha_j)$. For the second term, we have $E(\hat{\alpha}_j^4 - 2\hat{\alpha}_j^3 + \hat{\alpha}_j^2) = n^{-3}(n-1)(n-2)(n-3)\alpha_j^2(1-\alpha_j)^2 + (n-1)^2n^{-3}\alpha_j(1-\alpha_j)$. We then have $A^2E(\hat{\alpha}_j^4 - 2\hat{\alpha}_j^3 + \hat{\alpha}_j^2) = (n_0^{-1}+n_1^{-1})^2\{(n-2)(n-3)n^{-1}(n-1)^{-1}\alpha_j^2(1-\alpha_j)^2 + n^{-1}\alpha_j(1-\alpha_j)\}$. We next consider the third term. Because $\hat{\alpha}_j = (n_0\hat{\alpha}_{0j} + n_1\hat{\alpha}_{1j})/n$, we have

$$\begin{aligned}
&(\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^2(\hat{\alpha}_j - \hat{\alpha}_j^2) \\
&\quad = n^{-1}n_0(\hat{\alpha}_{0j}^3 + \hat{\alpha}_{0j}\hat{\alpha}_{1j}^2 - 2\hat{\alpha}_{0j}^2\hat{\alpha}_{1j}) \\
&\qquad + n^{-1}n_1(\hat{\alpha}_{1j}\hat{\alpha}_{0j}^2 + \hat{\alpha}_{1j}^3 - 2\hat{\alpha}_{0j}\hat{\alpha}_{1j}^2) \\
&\qquad - n^{-2}\hat{\alpha}_{0j}^2(n_0^2\hat{\alpha}_{0j}^2 + n_1^2\hat{\alpha}_{1j}^2 + 2n_0n_1\hat{\alpha}_{0j}\hat{\alpha}_{1j}) \\
&\qquad - n^{-2}\hat{\alpha}_{1j}^2(n_0^2\hat{\alpha}_{0j}^2 + n_1^2\hat{\alpha}_{1j}^2 + 2n_0n_1\hat{\alpha}_{0j}\hat{\alpha}_{1j}) \\
&\qquad + 2n^{-2}\hat{\alpha}_{0j}\hat{\alpha}_{1j}(n_0^2\hat{\alpha}_{0j}^2 + n_1^2\hat{\alpha}_{1j}^2 + 2n_0n_1\hat{\alpha}_{0j}\hat{\alpha}_{1j}) \\
&\quad = L_1 + L_2 - L_3 - L_4 + 2L_5.
\end{aligned}$$

The quantities L_1 to L_5 then need to be investigated separately as follows. $E(L_1) = \alpha_j^3\{-n^{-1}n_0(n_0^{-1}+n_1^{-1}) + 2n^{-1}n_0^{-1}\} + \alpha_j^2\{n^{-1}n_0(n_0^{-1}+n_1^{-1}) - 3n^{-1}n_0^{-1}\} + \alpha_j n^{-1}n_0^{-1}$. Similarly, $E(L_2) = \alpha_j^3\{-n^{-1}n_1(n_0^{-1}+n_1^{-1}) + 2n^{-1}n_1^{-1}\} + \alpha_j^2\{n^{-1}n_1(n_0^{-1}+n_1^{-1}) - 3n^{-1}n_1^{-1}\} + \alpha_j n^{-1}n_1^{-1}$. Next, we have $E(L_3) = \alpha_j^4\{n^{-2}n_0^{-1}(n-2)(n-3)(n_0-1)\} + \alpha_j^3\{5n^{-2}(n-2) + n^{-2}n_0^{-1}(n-2)(n-6)\} + \alpha_j^2\{4n^{-2} + n^{-2}n_0^{-1}(3n-7)\} + \alpha_j n^{-2}n_0^{-1}$. We also have $E(L_4) = \alpha_j^4\{n^{-2}n_1^{-1}(n-2)(n-3)(n_1-1)\} + \alpha_j^3\{5n^{-2}(n-2) + n^{-2}n_1^{-1}(n-2)(n-6)\} + \alpha_j^2\{4n^{-2} + n^{-2}n_1^{-1}(3n-7)\} +$

$\alpha_j n^{-2} n_1^{-1}$. Finally, we can verify that $E(L_5) = \alpha_j^4 n^{-2}(n-2)(n-3) + \alpha_j^3 5n^{-2}(n-2) + \alpha_j^2 4n^{-2}$. Combining the above results, we obtain

$$\begin{aligned}
&E\{(\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^2(\hat{\alpha}_j - \hat{\alpha}_j^2)\} \\
&\quad = E(L_1) + E(L_2) - \{E(L_3) + E(L_4) - 2E(L_5)\} \\
&\quad = (n_0^{-1} + n_1^{-1})\{n^{-2}(n-2)(n-3)\alpha_j^2(1-\alpha_j)^2 \\
&\qquad + n^{-2}(n-1)\alpha_j(1-\alpha_j)\}.
\end{aligned}$$

The above conclusions together with (7) lead to

$$\begin{aligned}
E(T_j^2) &= E(\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^4 + A^2 E(\hat{\alpha}_j^4 - 2\hat{\alpha}_j^3 + \hat{\alpha}_j^2) \\
&\quad - 2AE((\hat{\alpha}_{0j} - \hat{\alpha}_{1j})^2(\hat{\alpha}_j - \hat{\alpha}_j^2)) \\
&= \big\{(2 + n^{-1}(n-1)^{-1}(4n-6))(n_0^{-1} + n_1^{-1})^2 \\
&\quad - 6(n_0^{-3} + n_1^{-3})\big\}\alpha_j^2(1-\alpha_j)^2 \\
&\quad + \big\{(n_0^{-3} + n_1^{-3}) - n^{-1}(n_0^{-1} + n_1^{-1})^2\big\}\alpha_j(1-\alpha_j).
\end{aligned}$$

Recall that $n_k/n \to_p \pi_{k\cdot}$, so we have $E(T_j^2) = \{2(n_0^{-1} + n_1^{-1})^2 + O_p(n^{-3})\}\alpha_j^2(1-\alpha_j)^2 + O_p(n^{-3})\alpha_j(1-\alpha_j)$. As a result, we obtain $\mathrm{var}(T_c) = \{2(n_0^{-1} + n_1^{-1})^2 + O_p(n^{-3})\}\sum_{j=1}^p \alpha_j^2(1-\alpha_j)^2 + O_p(n^{-3})\sum_{j=1}^p \alpha_j(1-\alpha_j)$. Because $\alpha_j(1-\alpha_j) < 1$, it is obvious that $O(n^{-3})\sum_{j=1}^p \alpha_j^2(1-\alpha_j)^2 = o(pn^{-2})$ and $O(n^{-3})\sum_{j=1}^p \alpha_j(1-\alpha_j) = o(pn^{-2})$. Consequently, $\mathrm{var}(T_c) = 2(n_0^{-1} + n_1^{-1})^2\sum_{j=1}^p \alpha_j^2(1-\alpha_j)^2 + o_p(pn^{-2})$. This completes the proof.

APPENDIX C. PROOF OF THEOREM 2.3.

Denote $\mathrm{var}(T_c)$ by A_p^2 and let $B_p^2 = 2(1/n_0 + 1/n_1)^2\sum_{j=1}^p \alpha_j^2(1-\alpha_j)^2$. Then, by Theorem 2.2, we have $A_p^2 = B_p^2 + o_p(pn^{-2})$. Theorem 2.3 can be proved in three steps.

First step We first prove that $A_p^2/B_p^2 \to_p 1$. Since there exist $\alpha_{\min}, \alpha_{\max}$ such that $0 < \alpha_{\min} \le \alpha_j \le \alpha_{\max} < 1$ for every j, there exists a constant M such that $0 < M \le \alpha_j(1-\alpha_j) \le 1/4$ for every j. Consequently,

$$2M^2 p\Delta^2 \le B_p^2 = 2\Delta^2\sum_{j=1}^p \alpha_j^2(1-\alpha_j)^2 \le p\Delta^2/8,$$

where $\Delta = 1/n_0 + 1/n_1$. Thus, $|po(n^{-2})/B_p^2| \le |o(n^{-2})|/(2M^2\Delta^2) \to 0$. Combining the fact that

$$A_p^2/B_p^2 = 1 + po_p(n^{-2})/B_p^2,$$

we can obtain $A_p^2/B_p^2 \to_p 1$.

Second step In this step, we verify that $T_c/A_p \to_d N(0,1)$. Because

$$T_j = \big(\hat{\alpha}_{0j} - \hat{\alpha}_{1j}\big)^2 - \Delta\frac{n}{n-1}\hat{\alpha}_j(1-\hat{\alpha}_j),$$

it is easy to show that for every j, $|T_j| \le 1 + \Delta n/(n-1) \triangleq K$. By the strong law of large numbers, with probability 1, we have $n_k/n \to \pi_{k\cdot}$ for $k = 0, 1$. Hence, $K = 1 + o(n^{-1/2})$ with probability 1. Using the proof of Theorem 2.2, we also have that $A_p^2 = 2\Delta^2\sum_{j=1}^p \alpha_j^2(1-\alpha_j)^2 + pn^{-2}o(1) \ge pn^{-2}(2M + o(1))$ with probability 1. Furthermore, $p/n^2 \to \infty$, so it is true that $A_p^2 \to \infty$ with probability 1. This implies that, with probability 1, $K/A_p \to 0$. As a result, as long as n is large enough, we have that, for arbitrary $\tau > 0$ and every j, $P(|T_j| \le \tau A_p) = 1$. Consequently,

$$\lim_{n\to\infty}\frac{1}{A_p^2}\sum_{j=1}^p \int_{|x|>\tau A_p} x^2 dF_j(x) = 0,$$

where F_j is the distribution function of T_j given n_0, n_1. Consequently, the Lindeberg condition is satisfied. Then, by the central limit theorem, we have $T_c/A_p \to_d N(0,1)$.

Third step Finally, we demonstrate that $\widehat{D}_p^2/B_p^2 \to_p 1$. Let $z_j = \sqrt{n}(\hat{\alpha}_j - \alpha_j)$; then $\hat{\alpha}_j = n^{-1/2}z_j + \alpha_j$, $1 - \hat{\alpha}_j = (1 - \alpha_j) - n^{-1/2}z_j$ and $E(z_j) = 0, \mathrm{var}(z_j) = \alpha_j(1-\alpha_j) < 1$. Furthermore, we can obtain

$$\hat{\alpha}_j^2(1-\hat{\alpha}_j)^2 = \alpha_j^2(1-\alpha_j)^2 + e_j,$$

where $e_j = n^{-2}z_j^4 + (4\alpha_j - 2)n^{-3/2}z_j^3 + (1 - 6\alpha_j(1-\alpha_j))n^{-1}z_j^2 + 2\alpha_j(1-\alpha_j)(1-2\alpha_j)n^{-1/2}z_j$. For arbitrary ϵ, let $S^2 = 2/\epsilon$. By Tchebycheff's inequality, we have $P(|z_j| \le S) \ge 1 - \mathrm{var}(z_j)/S^2 > 1 - 1/S^2 > 1 - \epsilon$, for every$1 \le j \le p$. For fixed $M_0 = S^4 + S^3 + S^2 + 2S$, we have $P(|\sqrt{n}e_j| \le M_0) \ge P(|z_j^4| + |z_j^3| + |z_j^2| + 2|z_j| \le M_0) \ge P(|z_j| \le S) > 1 - \epsilon$, which implies that $\sqrt{n}e_j = O_p(1)$ for every j. Consequently, it is true that $\sum_{j=1}^p e_j = po_p(1)$. Hence, $\widehat{D}_p^2 = 2(1/n_0 + 1/n_1)^2\sum_{j=1}^p\{\alpha_j^2(1-\alpha_j)^2 + e_j\} = B_p^2 + pn^{-2}o_p(1)$. From the first step, we know that $pn^{-2}o_p(1)/B_p^2 \to_p 0$, and hence $\widehat{D}_p^2/B_p^2 \to_p 1$. Combining the results of these three steps, we have

$$T_f = \frac{T_c}{\widehat{D}_p} = \frac{T_c}{A_p}\frac{A_p}{B_p}\frac{B_p}{\widehat{D}_p} \to_d N(0,1).$$

This completes the proof.

ACKNOWLEDGEMENTS

We are very grateful to the editor, an associate editor, and two anonymous referees for their helpful and constructive comments, which led to a much improved manuscript. Research by Baiguo An and Jianhua Guo was supported in part by the National Natural Science Foundation of China (Nos. 10871038, 10926186, 11025102 and 11101182) and a Jilin Project (No. 20100401). Research by Hansheng Wang was supported in part by the Fox Ying Tong Education Foundation, the National Natural Science Foundation of China

(No. 11131002), Fundamental Research Funds for the Central Universities, Research Funds of Renmin University of China, and the Center for Statistical Science at Peking University.

Received 2 January 2012

REFERENCES

[1] Agresti, A. (1990). *Categorical Data Analysis*, John Wiley, New York. MR1044993

[2] Bouckaert, R. (2004). Naïve Bayes classifiers that perform well with continuous variables. In: *Proc. of the 17th Australian Conference on AI (AI 04), Lecture Notes AI.*

[3] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B.* **70** 849–911. MR2530322

[4] Feng, G., Guo, J., Jing, B. and Hao, L. (2012). A Bayesian feature selection paradigm for text classification. *Information Processing and Management.* **48** 283–302.

[5] Goeman, J., Geer, S., Kort, F. and Houwelingen, J. (2004). A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics.* **20** 93–99.

[6] Goeman, J., Geer, S. and Houwelingen, J. (2006). Testing against a high-dimensional alternative. *Journal of the Royal Statistical Society, Series B.* **68** 477–493. MR2278336

[7] Goeman, J., Houwelingen, H. and Finos, L. (2011). Testing against a high-dimensional alternative in generalized linear models: Asymptotic type I error control. *Biometrika.* **98** 381–390. MR2806435

[8] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer, New York. MR1851606

[9] Higgins, J. E. and Koch, G. G. (1977). Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey. *International Statistical Review.* **45** 51–62.

[10] Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*, Springer.

[11] Kosala, R. and Blockeel, H. (2000). Web mining research: A survey. *ACM SIGKDD Explorations Newsletter.* **2** 1–15.

[12] Kussul, E. and Baidyk, T. (2004). Improved method of handwritten digit recognition tested on MNIST database. *Image and Vision Computing.* **22** 971–981.

[13] Lee, D. L., Chuang, H. and Seamons, K. (1997). Document ranking and the vector-space model. *Software, IEEE.* **14** 67–75.

[14] Lewis, D. D. (1998). Naïve Bayes at forty: The independence assumption in information retrieval. In: *Proc. of ECML-98, 10th European Conference on Machine Learning.*, pp. 4–15.

[15] Milliken, G. A. and Johnson, D. E. (1993). *Analysis of Messy Data, Volume I: Designed Experiments*, Chapman and Hall, New York.

[16] Schneider, K. M. (2005). Techniques for improving the performance of naïve Bayes for text classification. *Lecture Notes in Computer Science.* **3406** 682–693.

[17] Stamey, T., Kablin, J., McNeal, J., Johnston, I., Freiha, F., Redwine, E. and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II. Radical prostatectomy treated patients. *Journal of Urology.* **16** 1076–1083.

[18] Wu, X. and Kumar, V. (2008). The top ten algorithms in data mining. *Knowledge and Information Systems.* **14** 1–37.

[19] Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In: *Proc. of the Fourteenth International Conference on Machine Learning.* 412–420.

Baiguo An
Key Laboratory for Applied Statistics
of the Ministry of Education
School of Mathematics and Statistics
Northeast Normal University
E-mail address: anbg200@gmail.com

Hansheng Wang
Guanghua School of Management
Peking University
E-mail address: hansheng@gsm.pku.edu.cn

Jianhua Guo
Key Laboratory for Applied Statistics
of the Ministry of Education
School of Mathematics and Statistics
Northeast Normal University
E-mail address: jhguo@nenu.edu.cn