SCIENCE CHINA
Mathematics

• ARTICLES •

April 2018 Vol. 61 No. 4: 627–640
https://doi.org/10.1007/s11425-017-9209-6

Network-based naive Bayes model for social network

Danyang Huang[1], Guoyu Guan[2,*], Jing Zhou[1] & Hansheng Wang[3]

[1]*School of Statistics, Renmin University of China, Beijing* 100872*, China;*
[2]*KLAS of MOE, and School of Economics, Northeast Normal University, Changchun* 130034*, China;*
[3]*Guanghua School of Management, Peking University, Beijing* 100872*, China*

*Email: dyhuang@ruc.edu.cn, guangy*599*@nenu.edu.cn, zhoujing_*89*@*126*.com, hansheng@pku.edu.cn*

Received June 22, 2017; accepted November 18, 2017; published online December 29, 2017

Abstract Naive Bayes (NB) is one of the most popular classification methods. It is particularly useful when the dimension of the predictor is high and data are generated independently. In the meanwhile, social network data are becoming increasingly accessible, due to the fast development of various social network services and websites. By contrast, data generated by a social network are most likely to be dependent. The dependency is mainly determined by their social network relationships. Then, how to extend the classical NB method to social network data becomes a problem of great interest. To this end, we propose here a network-based naive Bayes (NNB) method, which generalizes the classical NB model to social network data. The key advantage of the NNB method is that it takes the network relationships into consideration. The computational efficiency makes the NNB method even feasible in large scale social networks. The statistical properties of the NNB model are theoretically investigated. Simulation studies have been conducted to demonstrate its finite sample performance. A real data example is also analyzed for illustration purpose.

Keywords classification, naive Bayes, Sina Weibo, social network data

MSC(2010) 62H30, 91D30

Citation: Huang D Y, Guan G Y, Zhou J, et al. Network-based naive Bayes model for social network. Sci China Math, 2018, 61: 627–640, https://doi.org/10.1007/s11425-017-9209-6

1 Introduction

Naive Bayes [13, 18] is one of the most popular statistical classification methods. Its spirit has been widely used and generalized even in recent literature with various applications, which include but are not limited to credit scoring for credit applicants (see [1]), classification of Chinese text documents (see [11]), risk prediction in genetic studies (see [20]), and feature augmentation (see [9]). The naive Bayes model is especially useful when the dimension of the predictor is high, which makes density estimation unattractive. The assumptions of the naive Bayes model are rather optimistic. However, in many cases, the flexible approach could perform better than far more sophisticated alternatives even under these assumptions (see [13, 23]).

The aforementioned methods and applications consider no social network relationship (e.g., friendship, kinship) among individuals. Development of Internet technology makes more and more social network services and websites become popular (e.g., Facebook, Twitter, Sina Weibo, WeChat). As a consequence,

[*] Corresponding author

social network data are becoming increasingly accessible, since relationships among individuals could be collected. Users of the social network services or websites could be in large scale. Social network data refer to a group of nodes or individuals, and the relationships between them (see [26]). Various statistical tools have been proposed to model the nodes and their relationships in social network. The relevant literature includes but is not limited to, the Erdös-Rényi model (see [8]), the p_1 model (see [14]), the stochastic block model (see [3, 6, 22, 25]), and the exponential random graph model (see [15, 16]). The abundant studies help us understand that the relationships between the nodes make them no longer independent. This may make the traditional classification methods inappropriate in this case.

In the field of machine learning, *collective classification* (see [7]) has been popularly used to deal with classification problems in network data. This approach makes inference about individual's class label based on the attributes of the individual's own and its connected friends. But there are still two main bottlenecks of the collective classification approach. First, despite the practical application in network data, its theoretical properties have not been clearly investigated. Second, the approximate inference relies on iterative algorithms, thus the computational cost could be high (see [19, 21]). As a consequence, the approach could hardly be applied in large scale social network data. In addition, another type of methods use the statistical metrics of network as extra input features of classifiers, then many exiting classification methods become feasible on network data (see [30]). For example, *autologistic* (see [24]) is a representative method, which treats some network metrics as extra predictors in logistic regression. However, both methods have not explored the generation mechanism of network structure.

This motivates us to propose a new approach, which is called the network-based naive Bayes (NNB) model for social network. Specifically, consider a social network with a finite number of nodes. Then, the objective of the NNB method is to classify each node into one of the predefined classes. However, different from the classical naive Bayes method, the NNB approach takes both the nodal attributes and the social network relationships into consideration. Empirical evidence in both simulations and a real Sina Weibo (a Twitter-type social media in China) example shows that the prediction accuracy outperforms that of the classical naive Bayes model. We show that the computational cost of prediction based on the NNB model is feasible even when the network size is in large scale. Furthermore, the statistical properties of the NNB model have been theoretically investigated.

The rest of the article is organized as follows. The NNB model is introduced in Section 2. Both the classification rule and the theoretical properties are established in this section. In Section 3, a number of simulation studies have been conducted to demonstrate the performance of the proposed approach. A real data example in Sina Weibo is also analyzed for empirical evidence. Concluding remarks are given in Section 4 and all technical proofs are left to the appendix.

2 Methodology

2.1 Network-based naive Bayes model

Let $\{Y_i, X_i\}$ be the observation collected from the i-th ($1 \leqslant i \leqslant n$) node of the network, where $Y_i \in \{1, \ldots, K\}$ is the class label of the i-th node, $X_i = (X_{i1}, \ldots, X_{id})^{\mathrm{T}} \in \{0,1\}^d$ is the associated d-dimensional binary predictor (i.e., attributes of node i), and n is the size of the social network. For convenience, we write $\mathbb{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}} \in \{1, \ldots, K\}^n$ and $\mathbb{X} = (X_1, \ldots, X_n)^{\mathrm{T}} \in \{0,1\}^{n \times d}$. Next, define $\alpha_k = \mathrm{P}(Y_i = k) > 0$ and $\sum_{k=1}^{K} \alpha_k = 1$. Then, we assume that $X_{i1}, \ldots, X_{id}$ are independent with each other given Y_i, and the conditional probability is given by

$$\mathrm{P}(X_i \,|\, Y_i = k) = \prod_{j=1}^{d} \mu_{kj}^{X_{ij}} (1 - \mu_{kj})^{1 - X_{ij}}, \tag{2.1}$$

where $\mu_{kj} = \mathrm{P}(X_{ij} = 1 \,|\, Y_i = k) \in (0, 1)$. This is a similar assumption with that of the classical naive Bayes model (see [13, 18]).

To describe the social network structure, define $\mathbb{A}^{(n)} = (a_{i_1 i_2}) \in \{0,1\}^{n \times n}$ as the adjacency matrix, where $a_{i_1 i_2} = 1$ if node i_1 follows node i_2, otherwise $a_{i_1 i_2} = 0$. Note that $\mathbb{A}^{(n)}$ could be asymmetric because node i_1 follows node i_2 does not mean that node i_2 follows node i_1. We follow the tradition and let $a_{ii} = 0$ for any $1 \leqslant i \leqslant n$. Conditional on $\mathbb{Y}$ and $\mathbb{X}$, assume different edges (i.e., $a_{i_1 i_2}$'s) are independent with each other and define the link probability as

$$\mathrm{P}(a_{i_1 i_2} = 1 \,|\, \mathbb{Y}, \mathbb{X}) = \mathrm{P}(a_{i_1 i_2} = 1 \,|\, Y_{i_1} = k_1, Y_{i_2} = k_2) = \pi_{k_1 k_2}, \tag{2.2}$$

where $k_1, k_2 \in \{1, \ldots, K\}$ and $\pi_{k_1 k_2} \in (0, 1)$. It is remarked that (2.2) assumes that $\mathbb{X}$ and $\mathbb{A}^{(n)}$ are independent given $\mathbb{Y}$, and the distribution of link $a_{i_1 i_2}$ only depends on class labels Y_{i_1} and Y_{i_2}. As a result, social network formation is only determined by the class labels. Both (2.1) and (2.2) constitute the network-based naive Bayes (NNB) model, which can be represented as,

$$\mathrm{P}(\mathbb{Y}, \mathbb{X}, \mathbb{A}^{(n)}) = \bigg\{ \prod_{i=1}^{n} \mathrm{P}(Y_i) \bigg\} \bigg\{ \prod_{i=1}^{n} \prod_{j=1}^{d} \mathrm{P}(X_{ij} \,|\, Y_i) \bigg\} \bigg\{ \prod_{i_1 \neq i_2} \mathrm{P}(a_{i_1 i_2} \,|\, Y_{i_1}, Y_{i_2}) \bigg\}. \tag{2.3}$$

Write $\theta = (\alpha^{\mathrm{T}}, \mathrm{vec}(\mu)^{\mathrm{T}}, \mathrm{vec}(\pi)^{\mathrm{T}})^{\mathrm{T}}$, $\alpha = (\alpha_1, \ldots, \alpha_K)^{\mathrm{T}} \in \mathbb{R}^K$, $\mu = (\mu_{kj}) \in \mathbb{R}^{K \times d}$ and $\pi = (\pi_{k_1 k_2}) \in \mathbb{R}^{K \times K}$, where $\mathrm{vec}(\cdot)$ represents the vectorization of a matrix. The likelihood function could be represented as

$$\begin{aligned} \mathcal{L}(\theta) = & \bigg[\prod_{i=1}^{n} \prod_{k=1}^{K} \alpha_k^{I(Y_i = k)} \bigg] \bigg[\prod_{i=1}^{n} \prod_{j=1}^{d} \prod_{k=1}^{K} \{ \mu_{kj}^{X_{ij}} (1 - \mu_{kj})^{1 - X_{ij}} \}^{I(Y_i = k)} \bigg] \\ & \times \bigg[\prod_{i_1 \neq i_2} \prod_{k_1, k_2} \{ (\pi_{k_1 k_2})^{a_{i_1 i_2}} (1 - \pi_{k_1 k_2})^{1 - a_{i_1 i_2}} \}^{I(Y_{i_1} = k_1, Y_{i_2} = k_2)} \bigg]. \end{aligned} \tag{2.4}$$

By maximizing the above likelihood function, we can get the maximum likelihood estimator which is denoted by $\hat{\theta} = (\hat{\alpha}^{\mathrm{T}}, \mathrm{vec}(\hat{\mu})^{\mathrm{T}}, \mathrm{vec}(\hat{\pi})^{\mathrm{T}})^{\mathrm{T}} = \mathrm{argmax}_{\theta} \mathrm{P}(\mathbb{Y}, \mathbb{X}, \mathbb{A}^{(n)} \,|\, \theta)$. More specifically,

$$\begin{aligned} \hat{\alpha}_k &= n^{-1} \sum_{i=1}^{n} I(Y_i = k), \\ \hat{\mu}_{kj} &= \bigg\{ \sum_{i=1}^{n} I(Y_i = k) \bigg\}^{-1} \sum_{i=1}^{n} X_{ij} I(Y_i = k), \\ \hat{\pi}_{k_1 k_2} &= \bigg\{ \sum_{i_1 \neq i_2} I(Y_{i_1} = k_1, Y_{i_2} = k_2) \bigg\}^{-1} \sum_{i_1 \neq i_2} a_{i_1 i_2} I(Y_{i_1} = k_1, Y_{i_2} = k_2). \end{aligned}$$

It is remarked that the model definitions (2.1) and (2.2) could exhibit in much more complex form. The definitions adopted here are for simplicity and feasibility even in large scale social network. As one can see, the numerators and denominators of these estimators are all counting statistics, which can be simply computed. The computational complexity is comparable with the network density (i.e., Density $= (n^2 - n)^{-1} \sum_{i \neq j} a_{ij}$). Thus the parameters could be easily estimated even in large scale social network. Furthermore, one can easily verify that the numerators and denominators of these estimators are also summations of independent variables. Then, by the law of large numbers, we have $|\hat{\alpha}_k - \alpha_k| = O_{\mathrm{P}}(n^{-1/2})$, $|\hat{\mu}_{kj} - \mu_{kj}| = O_{\mathrm{P}}(n^{-1/2})$, and $|\hat{\pi}_{k_1 k_2} - \pi_{k_1 k_2}| = O_{\mathrm{P}}(n^{-1/2})$ (see Lemma B.1 in Appendix B for more details). Note that, $\hat{\mu}$ does not depend on $\mathbb{A}^{(n)}$, and $\hat{\pi}$ does not depend on $\mathbb{X}$.

Remark 2.1. Model (2.3) assumes that predictor and network structure are equally important. To make it more flexible, a tuning parameter could be introduced to control the weight of the predictor and the network structure. Define $Z \in \{0, 1\}$ with probability $\mathrm{P}(Z = 1) = \omega$, which is marginally independent of $\mathbb{Y}$. Also assume $\mathrm{P}(\mathbb{X}, \mathbb{A}^{(n)} \,|\, \mathbb{Y}, Z = 1) = \mathrm{P}(\mathbb{A}^{(n)} \,|\, \mathbb{Y})$ and $\mathrm{P}(\mathbb{X}, \mathbb{A}^{(n)} \,|\, \mathbb{Y}, Z = 0) = \mathrm{P}(\mathbb{X} \,|\, \mathbb{Y})$. Thus $\mathrm{P}(\mathbb{X}, \mathbb{A}^{(n)}, \mathbb{Y}) = \mathrm{P}(\mathbb{Y}) \{ \omega \mathrm{P}(\mathbb{A}^{(n)} \,|\, \mathbb{Y}) + (1 - \omega) \mathrm{P}(\mathbb{X} \,|\, \mathbb{Y}) \}$. ω is the weight parameter which balances two sources of information, i.e., the network and the predictor. When $\omega = 0$, only the predictor information is included; while when $\omega = 1$, only the network structure is considered. When $0 < \omega < 1$, the posterior probability $\mathrm{P}(Y_{n+1} = t \,|\, \mathbb{Y}, \mathbb{X}, X_{n+1}, \mathbb{A}^{(n+1)})$ is proportional to

$$\mathrm{P}(Y_{n+1} = t) \{ \omega \mathrm{P}(\mathbb{A}^{(n+1)} \,|\, \mathbb{Y}, Y_{n+1} = t) + (1 - \omega) \mathrm{P}(\mathbb{X}, X_{n+1} \,|\, \mathbb{Y}, Y_{n+1} = t) \}.$$

Based on this posterior probability, a nonlinear prediction rule can be obtained. For simplicity, we omit the weight parameter and assume equal importance of information from network structure and predictor.

2.2 Classification rule

As is known to all, the most important task of classification is to predict the unknown class labels. The prediction for the NNB model is not that intuitive as the classical naive Bayes model. It is because nodes are no longer independent due to the incorporation of the social network structure. It is assumed that when the $(n+1)$-th node arrives, (1) the associated predictor X_{n+1}, and (2) the relationships between $(n+1)$-th node and other observed nodes $\{a_{i,n+1}, a_{n+1,i}\}_{1\leqslant i\leqslant n}$ could be collected. Then we intend to predict the unknown class label Y_{n+1} based on all collected data, which are, $\mathbb{Y}$, $\mathbb{X}$, X_{n+1} and $\mathbb{A}^{(n+1)}$.

After a simple calculation, the posterior probability can be represented as,

$$\begin{aligned}
&P(Y_{n+1}=t\,|\,\mathbb{Y},\mathbb{X},X_{n+1},\mathbb{A}^{(n+1)})\\
&\quad\propto \alpha_t\Bigg[\prod_{j=1}^{d}\mu_{tj}^{X_{n+1,j}}(1-\mu_{tj})^{1-X_{n+1,j}}\Bigg]\\
&\qquad\times\Bigg[\prod_{i=1}^{n}\prod_{k=1}^{K}\{(\pi_{kt})^{a_{i,n+1}}(1-\pi_{kt})^{1-a_{i,n+1}}(\pi_{tk})^{a_{n+1,i}}(1-\pi_{tk})^{1-a_{n+1,i}}\}^{I(Y_i=k)}\Bigg].
\end{aligned}\tag{2.5}$$

The symbol "$\propto$" means "be proportional to". In other words, the part which is not dependent on t is omitted. Detailed derivation of (2.5) is left to Appendix A. In the next step, we maximize the posterior probability (2.5) to get a predicted class label. Substituting the MLE of θ into (2.5), we can obtain the following classification rule,

$$\begin{aligned}
\hat{Y}_{n+1}&=\mathop{\arg\max}_{1\leqslant t\leqslant K}\hat{P}(Y_{n+1}=t\,|\,\mathbb{Y},\mathbb{X},X_{n+1},\mathbb{A}^{(n+1)})\\
&=\mathop{\arg\max}_{1\leqslant t\leqslant K}\Bigg[\hat{C}_t+\sum_{j=1}^{d}g(\hat{\mu}_{tj})X_{n+1,j}+\sum_{k=1}^{K}\{g(\hat{\pi}_{kt})U_{n+1,k}+g(\hat{\pi}_{tk})V_{n+1,k}\}\Bigg],
\end{aligned}\tag{2.6}$$

where $\hat{C}_t=\log\hat{\alpha}_t+\sum_{j=1}^{d}\log(1-\hat{\mu}_{tj})+n\sum_{k=1}^{K}\hat{\alpha}_k\log\left[(1-\hat{\pi}_{kt})(1-\hat{\pi}_{tk})\right]$ for $1\leqslant t\leqslant K$, $g(z)=\log\{z(1-z)^{-1}\}$, $U_{n+1,k}=\sum_{i=1}^{n}I(Y_i=k,a_{i,n+1}=1)$ and $V_{n+1,k}=\sum_{i=1}^{n}I(Y_i=k,a_{n+1,i}=1)$ for $1\leqslant k\leqslant K$. Note that, $\hat{C}_t$, $g(\hat{\mu}_{kj})$ and $g(\hat{\pi}_{kt})$ are all functions of parameter estimators, which have been computed in the training step.

It is remarked that (2.6) is a comprehensive classification rule which is constructed based on two important sources of information: (1) the second term in (2.6) is the information from the predictor $\mathbb{X}$ and X_{n+1}, which denote the attributes of all the collected nodes; (2) the third term in (2.6) is the information from the social network structure $\mathbb{A}^{(n+1)}$. Both the predictor and the social network structure play a critical role in the prediction of class labels.

As to the computational complexity in prediction, only $K(d+K)$ logarithmic operations need to be computed. In addition, $U_{n+1,k}$ and $V_{n+1,k}$ are counting statistics, such that $D_{n+1}^{\rm in}=\sum_{k=1}^{K}U_{n+1,k}$ is the in-degree and $D_{n+1}^{\rm out}=\sum_{k=1}^{K}V_{n+1,k}$ is the out-degree of the $(n+1)$-th node in the network structure $\mathbb{A}^{(n+1)}$. This means in the prediction step, we only need to compute $2K$ counting statistics, $2K+d$ multiplications and $2K+d+1$ summations. As a consequence, even in large scale social network, the cost of computation is practically feasible. The classification rule of (2.6) shows that the method of NNB is a linear classifier, which supports its practical usefulness.

2.3 Theoretical properties

We next investigate the theoretical properties of the proposed NNB model. Before the establishment of the theorem, the following technical conditions are needed.

(C1) There exists some positive constant $\nu \in (0, \min\{1/K, 1/3\})$, such that $\min_{1 \leqslant k \leqslant K}\{\alpha_k\} \geqslant \nu$, $\min_{1 \leqslant k \leqslant K, 1 \leqslant j \leqslant d}\{\mu_{kj}, 1-\mu_{kj}\} \geqslant \nu$, $\min_{1 \leqslant k \neq l \leqslant K, 1 \leqslant j \leqslant d}\{|\mu_{kj}-\mu_{lj}|\} \geqslant \nu$, $\min_{1 \leqslant k_1, k_2 \leqslant K}\{\pi_{k_1 k_2}\} \geqslant \nu n^{-\gamma}$, and $\min_{1 \leqslant k \leqslant K}\{\pi_{kk} - \max_{1 \leqslant k \neq l \leqslant K}\{\pi_{kl}, \pi_{lk}\}\} \geqslant \nu n^{-\gamma}$, where $\gamma \geqslant 0$.

It is remarked that in Condition (C1), first, all the parameters are assumed to be bounded away from both 0 and 1. This excludes those cases where one particular category's probability is extremely small or extremely large for $\mathbb{Y}$ and $\mathbb{X}$. However, the assumption on $\pi_{k_1 k_2}$ allows the network to be sparse. Second, the intra-class link probability π_{kk} is assumed to be lager than the extra-class link probability π_{kl} $(k \neq l)$. This is a straightforward assumption, because it is intuitive to assume nodes belonging to the same class tend to have a higher probability to follow each other. Then, we have the following theorem.

Theorem 2.1. *Assuming the technical conditions in* (C1)*, for fixed* $K \geqslant 2$*,* $d \propto n^{\lambda}$ *and* $\pi_{k_1 k_2} \propto n^{-\gamma}$ *subject to* (1) $0 \leqslant \gamma < 1/4$ *or* (2) $1/4 \leqslant \gamma < 1/2 < \lambda \leqslant 1$ *or* (3) $1/2 \leqslant \gamma < \lambda \leqslant 1$*, we have* $P(\hat{Y}_i = 1 \,|\, Y_i = 1) \to 1$ *as* $n \to \infty$*.*

From the above theorem, one could see that with the classification rule of NNB, one can precisely predict the class label with probability tending to 1 as n goes to infinity. To achieve the theoretical prediction accuracy, two constraints should be satisfied. First, when the size of the network n becomes larger, the dimension of predictors d is assumed to become larger in an appropriate speed. This is reasonable in real practice. For example, in the Sina Weibo platform, the self-created labels of users may be increasingly diversified as the network size n grows, which could be used as predictors. Secondly, the average in-degree $(\mathrm{E}\sum_{i=1}^{n} a_{ji} \propto n^{1-\gamma})$ and average out-degree $(\mathrm{E}\sum_{j=1}^{n} a_{ij} \propto n^{1-\gamma})$ of node i are assumed to increase in an appropriate rate to the network size n. This means that each user gradually makes new friends as the network size n becomes larger. It is remarked that when dimension d becomes larger, irrelevant predictors which independent of $\mathbb{Y}$ could be involved. Then feature selection procedure could help to exclude those irrelevant predictors before parameter estimation. We will illustrate this in simulation examples. In the next section, some numerical studies will be conducted to illustrate the finite sample performance of the NNB classification rule.

3 Numerical studies

3.1 Simulation examples

To illustrate the performance of the proposed NNB method, two competitors are included for comparison. They are, respectively, the classical naive Bayes classifier (NB), in which the network information will not be used; and the network classifier (NC), in which the predictor information will not be used. Analogous to (2.6), classification rules of NB and NC could be written in a similar form as,

$$\hat{Y}_{n+1}^{\mathrm{NB}} = \mathop{\arg\max}_{1 \leqslant t \leqslant K} \left\{ \log \hat{\alpha}_t + \sum_{j=1}^{d} \log(1-\hat{\mu}_{kj}) + \sum_{j=1}^{d} g(\hat{\mu}_{tj}) X_{n+1,j} \right\}, \tag{3.1}$$

$$\begin{aligned} \hat{Y}_{n+1}^{\mathrm{NC}} = \mathop{\arg\max}_{1 \leqslant t \leqslant K} \Bigg\{ & \log \hat{\alpha}_t + n \sum_{k=1}^{K} \hat{\alpha}_k \log[(1-\hat{\pi}_{kt})(1-\hat{\pi}_{tk})] \\ & + \sum_{k=1}^{K} [g(\hat{\pi}_{kt}) U_{n+1,k} + g(\hat{\pi}_{tk}) V_{n+1,k}] \Bigg\}, \end{aligned} \tag{3.2}$$

where the notations $U_{n+1,k}$, $V_{n+1,k}$ and function $g(\cdot)$ have been defined in the previous section.

To verify the classification ability of our proposed method, some irrelevant predictors are included. Thus, the L_0-regularization feature selection method (see [12]) specially proposed for naive Bayes model could be adopted. The conditional independence assumption on $\mathbb{X}$ and $\mathbb{A}^{(n)}$ given $\mathbb{Y}$ makes this feature selection method also appropriate for NNB. Practically, relevant predictors are selected on the training set, and predictions are made based on the selected ones on the testing set. In the rest of this article, the performance of NB and NNB with selected features will be reported.

Table 1 Detailed simulation results for balanced case ($K = 2, 3$) with $S = 500$ replications

			Density	RMSE		AME (%)	
K	d	n	($\times 10^2$)	($\times 10^2$)	NB	NC	NNB
2	50	200	3.00	4.38	9.04	12.15	3.75
		500	2.25	2.59	7.21	5.29	1.47
		1,000	1.50	1.96	6.74	3.12	0.84
	100	200	2.99	4.00	2.72	12.09	1.24
		500	2.25	2.54	1.93	5.32	0.46
		1,000	1.50	1.93	1.73	3.14	0.24
3	50	200	2.67	5.33	14.81	27.01	8.83
		500	2.00	3.29	11.04	15.55	3.98
		1,000	1.33	2.21	10.12	10.86	2.65
	100	200	2.66	5.28	5.95	26.78	3.79
		500	2.00	3.36	3.33	15.44	1.26
		1,000	1.33	2.33	2.89	10.93	0.79

Table 2 Detailed simulation results for unbalanced case ($K = 2$) with $S = 500$ replications

		Density	RMSE	Precision (%)			Recall (%)			F1 measure (%)		
d	n	($\times 10^2$)	($\times 10^2$)	NB	NC	NNB	NB	NC	NNB	NB	NC	NNB
50	200	3.64	7.32	96.5	96.2	99.0	54.4	54.7	71.3	68.3	68.7	82.1
	500	2.73	4.93	98.6	98.5	99.7	75.2	81.9	93.7	85.1	89.4	96.6
	1,000	1.82	2.85	98.8	99.1	99.8	80.4	89.6	97.3	88.6	94.1	98.5
100	200	3.63	6.65	99.2	96.4	99.7	67.2	54.6	76.1	79.2	68.6	85.8
	500	2.73	4.80	99.6	98.5	99.9	88.8	81.4	96.4	93.8	89.0	98.1
	1,000	1.82	3.04	99.7	99.2	99.9	93.8	89.4	99.0	96.6	94.0	99.5

In the simulation study, we consider two cases of simulation examples: Case 1 with balanced classes, and Case 2 with unbalanced classes. The data of both two cases are simulated according to (2.3), in the following 3 steps. First, we independently generate the class labels $Y_1, \ldots, Y_n$ from $\{1, \ldots, K\}$ with probability $\mathrm{P}(Y_i = k) = \alpha_k$ for $1 \leqslant k \leqslant K$. Second, given each Y_i, the j-th binary predictor X_{ij} is generated from a Bernoulli distribution with probability $\mathrm{P}(X_{ij} = 1 \,|\, Y_i = k) = \mu_{kj}$ for $1 \leqslant k \leqslant K$ and $1 \leqslant j \leqslant d$. Here, we assume the first d_0 ($< d$) predictors are relevant to class labels, and the others are irrelevant, i.e., $\mu_{kj} = \mu_{0j}$ for $1 \leqslant k \leqslant K$ and $d_0 + 1 \leqslant j \leqslant d$. In addition, $\{\mu_{kj}\}_{1 \leqslant k \leqslant K, 1 \leqslant j \leqslant d_0}$ and $\{\mu_{0j}\}_{d_0+1 \leqslant j \leqslant d}$ are simulated from a uniform distribution on $[0.05, 0.95]$. Third, entries of adjacency matrix $\mathbb{A}^{(n)}$ are independently generated from a Bernoulli distribution with probability

$$\mathrm{P}(a_{i_1 i_2} = 1 \,|\, Y_{i_1} = k_1, Y_{i_2} = k_2) = \pi_{k_1 k_2}$$

for $1 \leqslant i_1 \neq i_2 \leqslant n$, and $a_{ii} = 0$ for $1 \leqslant i \leqslant n$.

To demonstrate the finite sample performance of the proposed method, various number of classes ($K = 2, 3$ for Case 1 and $K = 2$ for Case 2), the relevant predictor dimensions ($d_0 = 50, 100$), and sample sizes ($n = 200, 500, 1000$) are considered. Note that in all these cases, there are $d = 200$ predictors, which include d_0 relevant ones and $d - d_0$ irrelevant ones. For each fixed parameter setting, a total of $S = 500$ simulation replications are conducted. For each simulation replication, we adopt the leave-one-out manner, i.e., in i-th step, use i-th node for testing and the other $n-1$ nodes for training. Then, we use the notation $\hat{\theta}^{(s,i)}$ to represent the estimate obtained in the i-th step of the s-th simulation replication.

Next, the performance of the NNB method is measured from two aspects: (1) parameter estimation and (2) prediction accuracy. First, as to the performance of parameter estimation, we define the root mean squared error (RMSE), i.e.,

$$\mathrm{RMSE} = \left\{ S^{-1} n^{-1} m^{-1} \sum_{s=1}^{S} \sum_{i=1}^{n} \|\hat{\theta}^{(s,i)} - \theta\|^2 \right\}^{1/2},$$

where $m = K(1+K+d)$ is the dimension of θ. Moreover, the network density is computed as Density $= (n^2-n)^{-1}\sum_{i\neq j} a_{ij}$. Then RMSE ($\times 10^2$) and Density ($\times 10^2$) are reported in simulation results. Second, to evaluate the prediction accuracy, we adopt different criteria in the two cases, separately. Because the NNB approach is able to handle the problems with both $K=2$ and $K>2$. In Case 1, we adopt the general measurement of classification performance, which could be applied for both $K=2$ and $K>2$. While in Case 2, more detailed measurements will be adopted to illustrate the performance of NNB, which could be typically applied for $K=2$. The definitions of different criteria will be stated in the following cases, respectively.

Case 1 (Balanced classes). In this balanced case, we consider $K=2$ with class probabilities $\alpha = (1/2, 1/2)^{\mathrm{T}}$, and $K=3$ with class probabilities $\alpha = (1/3,1/3,1/3)^{\mathrm{T}}$. We then set the intra-class link probability as $\pi_{kk} = 0.04I(n=200) + 0.03I(n=500) + 0.02I(n=1000)$ for $1 \leqslant k \leqslant K$, and the extra-class link probability as $\pi_{kl} = 0.5\pi_{kk}$ for $k \neq l$. Note that, this setting could ensure: (1) the network density decreases as n grows; (2) the in- and out-degree of nodes increase as the network size n increases. Next, we use the mis-classification error to evaluate the prediction accuracies of all the three methods, which is, $\mathrm{ME}_s = n^{-1}\sum_{i=1}^n I(\hat{Y}_i^{(s)} \neq Y_i)$, where $\hat{Y}_i^{(s)}$ is the predicted class label for node i in the s-th replication. The average mis-classification error (AME $= S^{-1}\sum_s \mathrm{ME}_s$) of NB, NC and NNB over $S=500$ replications are reported as percentage in Table 1.

From Table 1, one could draw a conclusion that the NNB method always performs the best. This is as expected because the classification rule of NNB considers both information from the predictor and the network structure comprehensively. Second, when both K and d are fixed, the RMSE and AME values approach 0 quickly as n gets larger. This is because a larger sample size leads to more accurate estimates and predictions. Third, when both n and d are fixed, a larger class number K leads to larger RMSE and AME values, since when K gets larger, the sample sizes for each class becomes smaller and nodes have higher probability to be classified into incorrect classes, which leads to worse estimates and predictions. Lastly, for fixed K and n, a larger d leads to smaller RMSE and AME. This means the more relevant features are involved, the better estimates and predictors we could obtain.

Case 2 (Unbalanced classes). In this unbalanced case, we only consider the classification problem with $K=2$. The corresponding class probabilities are set to be $\alpha = (0.1, 0.9)^{\mathrm{T}}$, and the link probabilities π_{kl}s are set to be the same as in the balanced case. Unlike the balanced case, the mis-classification error is no longer a good criterion to evaluate the prediction accuracy. We adopt the frequently used evaluation criteria for the classification of unbalanced classes (see [17]). They are precision $P = tp/(tp+fp)$, recall $R = tp/(tp+fn)$ and F1 measure $F = 2PR/(P+R)$, where

$$tp = \left\{\sum_{i=1}^n I(Y_i = 1)\right\}^{-1} \sum_{i=1}^n I(\hat{Y}_i = 1, Y_i = 1)$$

represents the true positive rate,

$$fp = \left\{\sum_{i=1}^n I(Y_i = 2)\right\}^{-1} \sum_{i=1}^n I(\hat{Y}_i = 1, Y_i = 2)$$

represents the false positive rate, and

$$fn = \left\{\sum_{i=1}^n I(Y_i = 1)\right\}^{-1} \sum_{i=1}^n I(\hat{Y}_i = 2, Y_i = 1)$$

represents the false negative rate. Note that, class 1 is defined to be the positive class in this example. The average precision, recall, and F1 measure of all the three classification methods over $S=500$ replications are reported as percentage in Table 2.

From Table 2, one could see that the NNB method also always performs the best and the performance of RMSE is quite the same as in Table 1. Furthermore, for a fixed d, values of precision, recall and F1 measure approach 1 quickly, as n gets larger. Lastly, for fixed n, a larger d leads to larger precision, recall and F1 measure. This corroborates the conclusion in Theorem 2.1 and shows that the NNB method is also practically useful for unbalanced cases.

3.2 A Sina Weibo example

In this subsection, we analyze a real example and the data are collected from Sina Weibo, which is the largest Twitter-type social media in China. Sina Weibo allows different nodes or users to follow each other and share information between connected ones. In addition, Sina Weibo encourages users to create individual profiles. Each profile contains a list of self-created short labels (or keywords) for a typical user. These labels help to identify the user's most important characteristics. The information contained in labels could be diversified, which includes but is not limited to, a user's career status, life style and interests. As one can see, if labels are carefully prepared, they could be helpful for understanding the network structure. Since, intuitively, users with similar class labels are more likely to follow each other. This makes the NNB model combining the network structure and label information practically useful.

In this example, we collect the data starting from the official Weibo account of the MBA program of the Guanghua School of Management in Peking University. The number of users that follow the official account is 47,152 by the time we collect the data. As a matter of fact, due to the constraint imposed by Sina Weibo, only 5,000 followers could be sampled from the website. They are randomly selected by Sina Weibo but the random mechanism is confidential which is determined by Sina Weibo. Despite of the 446 incorrect accounts, the final sample contains $n = 5{,}000 - 446 = 4{,}554$ users.

In this dataset, the relationships of users could be represented by an adjacency matrix $\mathbb{A}^{(n)} = (a_{ij})$, where $a_{ij} = 1$ if the i-th user follows the j-th user, and $a_{ij} = 0$ otherwise. The network density is 0.0041, which indicates a fairly sparse network. This dataset also contains a total of 39 self-created labels. Specially, class labels are considered based on the labels which are the names of four famous business schools in China. Our objective here is to discriminate whether a user is from one of these four business schools or not, if other information is collected. Accordingly, the binary response $Y_i = 1$ if the i-th user carries at least one of above four labels of school names, and $Y_i = 2$ otherwise. Besides the above labels, other 35 self-created labels are viewed as binary predictors, i.e., the predictor $X_{ij} = 1$ if the i-th user have the j-th class label, and $X_{ij} = 0$ otherwise. To sum up, there are $\sum_i^n I(Y_i = 1) = 502$ users from class 1 (positive class), and the other 4,052 users from class 2 (negative class). Next, the class probabilities are estimated as $\hat{\alpha} = (0.1102, 0.8898)^{\mathrm{T}}$, which states that the classes are unbalanced. Then the link probabilities are estimated as $(\hat{\pi}_{11}, \hat{\pi}_{12}, \hat{\pi}_{21}, \hat{\pi}_{22}) = (0.0438, 0.0118, 0.0061, 0.0023)$, which show that users in class 1 are more likely to follow each other.

To illustrate the performance of the proposed method in this real dataset, the leave-one-out manner is adopted. For the sake of comparison, other popular classification methods, such as support vector machine SVM (see [28]), random forest RF (see [4]), adaptive boosting AdaB (see [5]) and autologistic AL (see [24]) are also considered. For the sake of fairness, out- and in-degrees in two classes of each node (i.e., $\sum_j I(Y_j = 1, a_{ij} = 1)$, $\sum_j I(Y_j = 2, a_{ij} = 1)$, $\sum_j I(Y_j = 1, a_{ji} = 1)$ and $\sum_j I(Y_j = 2, a_{ji} = 1)$) are treated as extra predictors for these competitors.

The L_0-regularization feature selection method (see [12]) is adopted both for NB and NNB. There are 12 predictors selected in average. The simulation results are shown in Table 3. We can find that NNB always performs better than the other methods in both recall and F1 measure. But NNB is not uniformly optimal, since it is worse than some competitors in precision. Therefore, we would continue to optimize this approach in the future.

Table 3 Results for the Sina Weibo example

Evaluation criteria	Classification methods						
	SVM	RF	AdaB	AL	NB	NC	NNB
Precision (%)	75.37	93.64	93.80	91.32	81.64	83.73	83.79
Recall (%)	22.51	41.04	38.45	52.99	10.76	58.17	58.57
F1 measure (%)	34.67	57.06	54.54	67.06	19.01	68.65	68.94

4 Concluding remarks

The novel network naive Bayes model for social network data has been proposed in this study, which incorporates social network structure into the classical naive Bayes model. After deriving the MLE of parameters, the computation of the classification rule is rather easy and fast, which is feasible even in large scale social network. Statistical properties of the proposed NNB have been theoretically established. Numerical studies have been conducted to empirically show the superior performance of the proposed NNB method.

To conclude this article, we present four other interesting topics for future research. First, the inspiration of NNB could be applied to other Bayesian classifiers, such as tree-augmented naive Bayes (see [10]), lazy Bayesian rules (see [31]), averaged one-dependence estimators (see [27]) and weighted naive Bayes (see [29]). Second, the assumptions on social network we considered are simple and intuitive in this study, such as the conditional independence assumption of $\mathbb{X}$ and $\mathbb{A}^{(n)}$ given $\mathbb{Y}$. However, when the dimension d is fairly large, the conditional independence may not be satisfied. To get more accurate prediction, these predictors relative with the network should be identified, which is another intriguing research topic. Third, the statistical properties of estimators when we adopt the weight parameter ω need to be established, and the corresponding prediction rule should be linearly approximated to make it feasible in the future work. Fourth, Laplacian support vector machine [2] is practically useful in classifying the observations with adjacent relationship. It learns the adjacent relationship from $\mathbb{X}$ and constructs a Laplacian matrix. While the network structure $\mathbb{A}^{(n)}$ naturally defines the adjacent relationship in network data, how to construct a more reasonable Laplacian matrix L and get better prediction is worth studying in the future.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 11701560, 11501093, 11631003, 11690012, 71532001 and 11525101), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (Grant No. 16XNLF01), the Beijing Municipal Social Science Foundation (Grant No. 17GLC051), Fund for Building World-Class Universities (Disciplines) of Renmin University of China, the Fundamental Research Funds for the Central Universities (Grant Nos. 130028613, 130028729 and 2412017FZ030), China's National Key Research Special Program (Grant No. 2016YFC0207700) and Center for Statistical Science at Peking University.

References

1 Antonakis A C, Sfakianakis M E. Assessing naïve Bayes as a method for screening credit applicants. J Appl Stat, 2009, 36: 537–545

2 Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res, 2006, 7: 2399–2434

3 Bickel P J, Chen A. A nonparametric view of network models and Newman-Girvan and other modularities. Proc Natl Acad Sci USA, 2009, 106: 21068–21073

4 Breiman L. Random forest. Mach Learn, 2001, 45: 5–32

5 Buhlmann P, Yu B. Boosting with the L_2 loss: Regression and classification. J Amer Statist Assoc, 2003, 98: 324–340

6 Choi D, Wolfe P, Airoldi E. Stochastic blockmodels with a growing number of classes. Biometrika, 2012, 99: 273–284

7 Craven M, McCallum A, PiPasquo D, et al. Learning to extract symbolic knowledge from the World Wide Web. In: Proceedings of the 15th National Conference on Artificial Intelligence. World Wide Web Internet and Web Information Systems, vol. 118. Menlo Park: Amer Assoc Artif Intell, 1998, 509–516

8 Erdős P, Rényi A. On the evolution of random graphs. Magyar Tud Akad Mat Kutató Int Közl, 1960, 5: 17–61

9 Fan J, Feng Y, Jiang J, et al. Feature augmentation via nonparametrics and selection (FANS) in high-dimensional classification. J Amer Statist Assoc, 2016, 111: 275–287

10 Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn, 1997, 29: 131–163

11 Guan G, Guo J, Wang H. Varying naive Bayes models with applications to classification of Chinese text documents. J Bus Econom Statist, 2014, 32: 445–456

12 Guan G, Shan N, Guo J. Feature screening for ultrahigh dimensional binary data. Stat Interface, 2018, 11: 41–50

13 Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York: Springer, 2001

14 Holland P W, Leinhardt S. An exponential family of probability distributions for directed graphs. J Amer Statist

Assoc, 1981, 76: 33–50

15 Hunter D R, Handcock M S. Inference in curved exponential family models for networks. J Comput Graph Statist, 2006, 15: 565–583

16 Hunter D R, Handcock M S, Butts C T, et al. Ergm: A package to fit, simulate and diagnose exponential-family models for networks. J Statist Softw, 2008, 24: 1–29

17 Lewis D D. Evaluating and optimizing autonomous text classification systems. In: International Acm Sigir Conference on Research and Development in Information Retrieval. New York: ACM, 1995, 246–254

18 Lewis D D. Naive Bayes at forty: The independence assumption in information retrieval. In: Proceedings of ECML-98, 10th European Conference on Machine Learning. London: Springer-Verlag, 1998, 4–15

19 Macskassy S A, Provost F. Classification in networked data: A toolkit and a univariate case study. J Mach Learn Res, 2007, 8: 935–983

20 Minnier J, Yuan M, Liu J S, et al. Risk classification with an adaptive naive Bayes kernel machine model. J Amer Statist Assoc, 2015, 110: 393–404

21 Neville J, Jensen D. Iterative classification in relational data. In: Proceedings of American Association for Artificial Intelligence Workshop on Learning Statistical Models from Relational Data. Palo Alto: AAAI Press, 2000, 42–49

22 Nowicki K, Snijders T A B. Estimation and prediction for stochastic block structures. J Amer Statist Assoc, 2001, 96: 1077–1087

23 Ozuysal M, Calonder M, Lepetit V, et al. Fast keypoint recognition using random ferns. IEEE Trans Pattern Anal Mach Intell, 2010, 32: 448–461

24 Robins G, Pattison P, Elliott P. Network models for social influence processes. Psychometrika, 2001, 66: 161–189

25 Wang Y J, Wong G Y. Stochastic blockmodels for directed graphs. J Amer Statist Assoc, 1987, 82: 8–19

26 Wasserman S, Faust K. Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press, 1994

27 Webb G I, Boughton J R, Wang Z. Not so naive Bayes: Aggregating one-dependence estimators. Mach Learn, 2005, 58: 5–24

28 Wu Y, Liu Y. Robust truncated-hinge-loss support vector machines. J Amer Statist Assoc, 2007, 102: 974–983

29 Zaidi N A, Cerquides J, Carman M, et al. Alleviating naive Bayes attribute independence assumption by attribute weighting. J Mach Learn Res, 2013, 14: 1947–1988

30 Zanin M, Papo D, Sousa P A, et al. Combining complex networks and data mining: Why and how. Phys Rep, 2016, 635: 1–44

31 Zheng Z, Webb G I. Lazy learning of Bayesian rules. Mach Learn, 2000, 41: 53–84

Appendix A Detailed derivation of (2.5)

In the following derivation, the chain rule and Bayes' rule in probability theory can be used. Then the posterior probability can be derived as

$$
\begin{aligned}
&\mathrm{P}(Y_{n+1}=t\,|\,\mathbb{Y},\mathbb{X},X_{n+1},\mathbb{A}^{(n+1)})\\
&\quad=\{\mathrm{P}(\mathbb{Y},\mathbb{X},X_{n+1},\mathbb{A}^{(n+1)})\}^{-1}\mathrm{P}(Y_{n+1}=t,\mathbb{Y},\mathbb{X},X_{n+1},\mathbb{A}^{(n+1)})\\
&\quad=\{\mathrm{P}(\mathbb{Y},\mathbb{X},X_{n+1},\mathbb{A}^{(n+1)})\}^{-1}\alpha_t\Bigg[\prod_{i=1}^{n}\prod_{k=1}^{K}\alpha_k^{I(Y_i=k)}\Bigg]\\
&\qquad\times\Bigg[\prod_{i=1}^{n}\prod_{j=1}^{d}\prod_{k=1}^{K}\{\mu_{kj}^{X_{ij}}(1-\mu_{kj})^{1-X_{ij}}\}^{I(Y_i=k)}\Bigg]\Bigg[\prod_{j=1}^{d}\mu_{tj}^{X_{n+1,j}}(1-\mu_{tj})^{1-X_{n+1,j}}\Bigg]\\
&\qquad\times\Bigg[\prod_{1\leqslant i_1\neq i_2\leqslant n}\ \prod_{1\leqslant k_1,k_2\leqslant K}\{(\pi_{k_1k_2})^{a_{i_1i_2}}(1-\pi_{k_1k_2})^{1-a_{i_1i_2}}\}^{I(Y_{i_1}=k_1,Y_{i_2}=k_2)}\Bigg]\\
&\qquad\times\Bigg[\prod_{i=1}^{n}\prod_{k=1}^{K}\{(\pi_{kt})^{a_{i,n+1}}(1-\pi_{kt})^{1-a_{i,n+1}}(\pi_{tk})^{a_{n+1,i}}(1-\pi_{tk})^{1-a_{n+1,i}}\}^{I(Y_i=k)}\Bigg]\\
&\quad=C\alpha_t\Bigg[\prod_{j=1}^{d}\mu_{tj}^{X_{n+1,j}}(1-\mu_{tj})^{1-X_{n+1,j}}\Bigg]\\
&\qquad\times\Bigg[\prod_{i=1}^{n}\prod_{k=1}^{K}\{(\pi_{kt})^{a_{i,n+1}}(1-\pi_{kt})^{1-a_{i,n+1}}(\pi_{tk})^{a_{n+1,i}}(1-\pi_{tk})^{1-a_{n+1,i}}\}^{I(Y_i=k)}\Bigg],
\end{aligned}
\tag{A.1}
$$

where C is a constant independent on t, i.e.,

$$\{\mathrm{P}(\mathbb{Y}, \mathbb{X}, X_{n+1}, \mathbb{A}^{(n+1)})\}^{-1}\Bigg[\prod_{i=1}^{n}\prod_{k=1}^{K}\alpha_k^{I(Y_i=k)}\Bigg]\Bigg[\prod_{i=1}^{n}\prod_{j=1}^{d}\prod_{k=1}^{K}\{\mu_{kj}^{X_{ij}}(1-\mu_{kj})^{1-X_{ij}}\}^{I(Y_i=k)}\Bigg]$$
$$\times\Bigg[\prod_{1\leqslant i_1\neq i_2\leqslant n}\prod_{1\leqslant k_1,k_2\leqslant K}\{(\pi_{k_1k_2})^{a_{i_1i_2}}(1-\pi_{k_1k_2})^{1-a_{i_1i_2}}\}^{I(Y_{i_1}=k_1,Y_{i_2}=k_2)}\Bigg].$$

Appendix B A technical lemma

Lemma B.1. *Assume* (C1), *for any* $1 \leqslant k_1, k_2 \leqslant K$ *and* $\gamma \geqslant 0$, *we have* $|\hat{\pi}_{k_1k_2} - \pi_{k_1k_2}| = O_{\mathrm{P}}(n^{-1/2})$. *Given* $Y_{n+1} = 1$, *for any* $1 \leqslant k \leqslant K$, *we have* $|\tilde{\pi}_{k1} - \pi_{k1}| = O_{\mathrm{P}}(n^{-1/2})$ *and* $|\tilde{\pi}_{1k} - \pi_{1k}| = O_{\mathrm{P}}(n^{-1/2})$, *where*

$$\tilde{\pi}_{k1} = \Bigg\{\sum_{i=1}^{n} I(Y_i = k)\Bigg\}^{-1}\sum_{i=1}^{n} a_{i,n+1} I(Y_i = k)$$

and

$$\tilde{\pi}_{1k} = \Bigg\{\sum_{i=1}^{n} I(Y_i = k)\Bigg\}^{-1}\sum_{i=1}^{n} a_{n+1,i} I(Y_i = k).$$

Proof. We first derive the order of $|\hat{\pi}_{k_1k_2} - \pi_{k_1k_2}|$. Here, we only consider the case of $k_1 \neq k_2$ in this proof. Then one can prove the case of $k_1 = k_2$ in a similar way. For fixed k_1 and k_2 subject to $k_1 \neq k_2$, we define $Z_{i_1i_2} = a_{i_1i_2} I(Y_{i_1} = k_1, Y_{i_2} = k_2)$ and $W_{i_1i_2} = I(Y_{i_1} = k_1, Y_{i_2} = k_2)$ for convince. Thus we can write $\hat{\pi}_{k_1k_2} = \bar{W}^{-1}\bar{Z}$, where

$$\bar{Z} = n^{-1}(n-1)^{-1}\sum_{i_1\neq i_2} Z_{i_1i_2}$$

and

$$\bar{W} = n^{-1}(n-1)^{-1}\sum_{i_1\neq i_2} W_{i_1i_2}.$$

Then we have $\mathrm{E}\bar{Z} = \alpha_{k_1}\alpha_{k_2}\pi_{k_1k_2}$ and

$$\mathrm{var}(\bar{Z}) = n^{-2}(n-1)^{-2}\sum_{i\neq j}\sum_{k\neq l}\mathrm{cov}(Z_{ij}, Z_{kl}).$$

The variance contains four cases of terms as follows:

Case 1. For $i \neq j \neq k \neq l$, $\mathrm{cov}(Z_{ij}, Z_{kl}) = \mathrm{E}(Z_{ij}Z_{kl}) - \mathrm{E}Z_{ij}\mathrm{E}Z_{kl} = 0$.

Case 2. For $i \neq j$, $k \neq l$, $i = k$ and $j \neq l$, $\mathrm{cov}(Z_{ij}, Z_{kl}) = \pi_{k_1k_2}^2\alpha_{k_1}(1-\alpha_{k_1})\alpha_{k_2}^2$.

Case 3. For $i \neq j$, $k \neq l$, $i \neq k$ and $j = l$, $\mathrm{cov}(Z_{ij}, Z_{kl}) = \pi_{k_1k_2}^2\alpha_{k_1}^2\alpha_{k_2}(1-\alpha_{k_2})$.

Case 4. For $i \neq j$, $k \neq l$, $i = k$ and $j = l$, $\mathrm{cov}(Z_{ij}, Z_{kl}) = \pi_{k_1k_2}\alpha_{k_1}\alpha_{k_2} - \pi_{k_1k_2}^2\alpha_{k_1}^2\alpha_{k_2}^2$.

We can find that, for Cases 2 and 3, there are a total of $n(n-1)(n-2)$ terms in $\mathrm{var}(\bar{Z})$. For Case 4, there are a total of $n(n-1)$ terms in $\mathrm{var}(\bar{Z})$. Then we have $\mathrm{var}(\bar{Z}) = n^{-1}(n-1)^{-1}(n-2)\pi_{k_1k_2}^2\alpha_{k_1}\alpha_{k_2}(\alpha_{k_1} + \alpha_{k_2} - 2\alpha_{k_1}\alpha_{k_2}) + n^{-1}(n-1)^{-1}\pi_{k_1k_2}\alpha_{k_1}\alpha_{k_2}(1 - \pi_{k_1k_2}\alpha_{k_1}\alpha_{k_2})$. By the assumption $\pi_{k_1k_2} \propto n^{-\gamma}$, we have $\mathrm{var}(\bar{Z}) = O(n^{-\min\{2\gamma+1,2+\gamma\}})$. Consequently,

$$|\bar{Z} - \alpha_{k_1}\alpha_{k_2}\pi_{k_1k_2}| = O_{\mathrm{P}}(n^{-\min\{\gamma+1/2,1+\gamma/2\}}).$$

Similarly, we have $\mathrm{E}\bar{W} = \alpha_{k_1}\alpha_{k_2}$ and $\mathrm{var}(\bar{W}) = n^{-1}(n-1)^{-1}(n-2)\alpha_{k_1}\alpha_{k_2}(\alpha_{k_1} + \alpha_{k_2} - 2\alpha_{k_1}\alpha_{k_2}) + n^{-1}(n-1)^{-1}\alpha_{k_1}\alpha_{k_2}(1-\alpha_{k_1}\alpha_{k_2}) = O(n^{-1})$. Then we have $|\bar{W} - \mathrm{E}\bar{W}| = O_{\mathrm{P}}(n^{-1/2})$. Therefore, for $k_1 \neq k_2$, it is not hard to verify that $|\hat{\pi}_{k_1k_2} - \pi_{k_1k_2}| = |\bar{W}^{-1}\bar{Z} - \pi_{k_1k_2}| = O_{\mathrm{P}}(\max\{n^{-\min\{\gamma+1/2,1+\gamma/2\}}, n^{-1/2}\}) = O_{\mathrm{P}}(n^{-\min\{\gamma+1/2,1+\gamma/2,1/2\}}) = O_{\mathrm{P}}(n^{-1/2})$ when $\gamma \geqslant 0$. We can also have $|\hat{\pi}_{kk} - \pi_{kk}| = O_{\mathrm{P}}(n^{-1/2})$ for any $1 \leqslant k \leqslant K$ in the same way, but proof details are omitted. Thus, for any $1 \leqslant k_1, k_2 \leqslant K$, $|\hat{\pi}_{k_1k_2} - \pi_{k_1k_2}| = O_{\mathrm{P}}(n^{-1/2})$.

Next, for a fixed k, analogous to the above discussion, we consider the order of $|\tilde{\pi}_{k1}-\pi_{k1}|$ when given $Y_{n+1}=1$. Define $Z_i^*=a_{i,n+1}I(Y_i=k)$ and $W_i^*=I(Y_i=k)$. Then $\tilde{\pi}_{01}=(\bar{W}^*)^{-1}\bar{Z}^*$, where $\bar{Z}^*=n^{-1}\sum_{i=1}^n Z_i^*$ and $\bar{W}^*=n^{-1}\sum_{i=1}^n W_i^*$. Then we have $\mathrm{E}(\bar{Z}^*\,|\,Y_{n+1}=1)=\pi_{k1}\alpha_k$, $\mathrm{var}(\bar{Z}^*\,|\,Y_{n+1}=1)=n^{-1}(\pi_{k1}\alpha_k-\pi_{k1}^2\alpha_k^2)=O(n^{-\gamma-1})$ when $\gamma\geqslant 0$. Hence, given $Y_{n+1}=1$, $|\bar{Z}^*-\mathrm{E}(\bar{Z}^*|Y_{n+1}=1)|=O_{\mathrm{P}}(n^{-1/2-\gamma/2})$.

Similarly, we have $\mathrm{E}(\bar{W}^*\,|\,Y_{n+1}=k)=\alpha_k$, $\mathrm{var}(\bar{W}^*\,|\,Y_{n+1}=k)=n^{-1}(\alpha_k-\alpha_k^2)=O(n^{-1})$. Then, given $Y_{n+1}=1$, we have $|\bar{W}^*-\mathrm{E}(\bar{W}^*\,|\,Y_{n+1}=k)|=O_{\mathrm{P}}(n^{-1/2})$. Therefore, given $Y_{n+1}=1$, $|\tilde{\pi}_{k1}-\pi_{k1}|=|(\bar{W}^*)^{-1}\bar{Z}^*-\pi_{k1}|=O_{\mathrm{P}}(n^{-\min\{1/2+\gamma/2,1/2\}})=O_{\mathrm{P}}(n^{-1/2})$ when $\gamma\geqslant 0$. In the same way, given $Y_{n+1}=1$, we can also prove that $|\tilde{\pi}_{1k}-\pi_{1k}|=O_{\mathrm{P}}(n^{-1/2})$, because their mathematical expressions are similar. □

Appendix C Proof of Theorem 2.1

Let $Q_t=\hat{\mathrm{P}}(Y_{n+1}=t\,|\,\mathbb{Y},\mathbb{X},X_{n+1},\mathbb{A}^{(n+1)})$. Then we have

$$\mathrm{P}(\hat{Y}_{n+1}=1\,|\,Y_{n+1}=1)=\mathrm{P}\bigg(\bigcap_{t\geqslant 2}(Q_1>Q_t)\,\bigg|\,Y_{n+1}=1\bigg)\geqslant 1-\sum_{t\geqslant 2}\mathrm{P}(Q_t\geqslant Q_1\,|\,Y_{n+1}=1).$$

Thus, in order to prove $\mathrm{P}(\hat{Y}_{n+1}=1\,|\,Y_{n+1}=1)\to 1$ as $n\to\infty$, we only need to prove $\mathrm{P}(Q_1-Q_t\geqslant 0\,|\,Y_{n+1}=1)\to 1$ as $n\to\infty$, for any $t\geqslant 2$. To this end, after some simple mathematical derivations, we can write $Q_1-Q_t=F_1+F_2+F_3+F_4$, where $F_1=\log(\hat{\alpha}_t^{-1}\hat{\alpha}_1)$ and

$$F_2=\sum_{j=1}^d\bigg\{X_{n+1,j}\log\frac{\hat{\mu}_{1j}}{\hat{\mu}_{tj}}+(1-X_{n+1,j})\log\frac{1-\hat{\mu}_{1j}}{1-\hat{\mu}_{tj}}\bigg\},$$

$$F_3=n\sum_{k=1}^K\hat{\alpha}_k\bigg\{\tilde{\pi}_{k1}\log\frac{\hat{\pi}_{k1}}{\hat{\pi}_{kt}}+(1-\tilde{\pi}_{k1})\log\frac{1-\hat{\pi}_{k1}}{1-\hat{\pi}_{kt}}\bigg\},$$

$$F_4=n\sum_{k=1}^K\hat{\alpha}_k\bigg\{\tilde{\pi}_{1k}\log\frac{\hat{\pi}_{1k}}{\hat{\pi}_{tk}}+(1-\tilde{\pi}_{1k})\log\frac{1-\hat{\pi}_{1k}}{1-\hat{\pi}_{tk}}\bigg\}.$$

In what follows the above four terms will be carefully evaluated separately.

Step 1. Firstly, we need to prove that F_1 is bounded with probability one, as given $Y_{n+1}=1$. By Condition (C1), we immediately know that $|\log(\alpha_t^{-1}\alpha_1)|<-\log\nu$. Then by $|\hat{\alpha}_k-\alpha_k|=O_{\mathrm{P}}(n^{-1/2})$, we have

$$|\log\hat{\alpha}_k-\log\alpha_k|=|\log\{1+\alpha_k^{-1}(\hat{\alpha}_k-\alpha_k)\}|=O_{\mathrm{P}}(n^{-1/2}),$$

for $1\leqslant k\leqslant K$. Thus, we can immediately obtain that

$$\begin{aligned}|F_1|&\leqslant|\log(\alpha_t^{-1}\alpha_1)|+|\log(\hat{\alpha}_t^{-1}\hat{\alpha}_1)-\log(\alpha_t^{-1}\alpha_1)|\\&\leqslant-\log\nu+\big|\log\hat{\alpha}_1-\log\alpha_1\big|+\big|\log\hat{\alpha}_t-\log\alpha_t\big|=O_{\mathrm{P}}(n^{-1/2}).\end{aligned}$$

Step 2. We next prove F_2 tends to infinity with probability one, as given $Y_{n+1}=1$. By Taylor's expansion with Lagrange remainder term at point (μ_{1j},μ_{tj}), there exist a number $\tilde{\mu}_{1j}$ between $\hat{\mu}_{1j}$ and μ_{1j}, and a number $\tilde{\mu}_{tj}$ between $\hat{\mu}_{tj}$ and μ_{tj}, such that $F_2=F_{21}+F_{22}$, where

$$F_{21}=\sum_{j=1}^d\bigg\{X_{n+1,j}\log\frac{\mu_{1j}}{\mu_{tj}}+(1-X_{n+1,j})\log\frac{1-\mu_{1j}}{1-\mu_{tj}}\bigg\},$$

$$F_{22}=\sum_{j=1}^d\bigg\{\frac{X_{n+1,j}-\tilde{\mu}_{1j}}{\tilde{\mu}_{1j}(1-\tilde{\mu}_{1j})}(\hat{\mu}_{1j}-\mu_{1j})-\frac{X_{n+1,j}-\tilde{\mu}_{tj}}{\tilde{\mu}_{tj}(1-\tilde{\mu}_{tj})}(\hat{\mu}_{tj}-\mu_{tj})\bigg\}.$$

These two terms will be considered separately. For convenience, we denote

$$T_{n+1,j}=X_{n+1,j}\log(\mu_{tj}^{-1}\mu_{1j})+(1-X_{n+1,j})\log\{(1-\mu_{tj})^{-1}(1-\mu_{1j})\}.$$

Then by [12, Lemma 3] and $E(X_{n+1,j}\,|\,Y_{n+1}=1)=\mu_{1j}$, we immediately know that

$$E(T_{n+1,j}\,|\,Y_{n+1}=1):=\eta_j=\mu_{1j}\log\frac{\mu_{1j}}{\mu_{tj}}+(1-\mu_{1j})\log\frac{1-\mu_{1j}}{1-\mu_{tj}}\geqslant\nu\log\frac{1+\nu}{1-\nu}, \tag{C.1}$$

$$|T_{n+1,j}-\eta_j|=\left|(X_{n+1,j}-\mu_{1j})\log\frac{\mu_{1j}(1-\mu_{tj})}{\mu_{tj}(1-\mu_{1j})}\right|\leqslant 2\log\frac{1-\nu}{\nu}. \tag{C.2}$$

Furthermore, by $\mathrm{var}(X_{n+1,j}\,|\,Y_{n+1}=1)=\mu_{1j}(1-\mu_{1j})$, we have

$$\frac{1}{d}\sum_{j=1}^{d}\mathrm{var}(T_{n+1,j}-\eta_j\,|\,Y_{n+1}=1)=\frac{1}{d}\sum_{j=1}^{d}\mu_{1j}(1-\mu_{1j})\bigg[\log\frac{\mu_{1j}(1-\mu_{tj})}{\mu_{tj}(1-\mu_{1j})}\bigg]^2\leqslant\bigg(\log\frac{1-\nu}{\nu}\bigg)^2. \tag{C.3}$$

For convenience, we denote $\xi=0.5\nu\log\{(1-\nu)^{-1}(1+\nu)\}$ and $\kappa=\log\{\nu^{-1}(1-\nu)\}$. Based on (C.1)–(C.3), we have

$$\begin{aligned}\mathrm{P}(F_{21}\geqslant d\xi\,|\,Y_{n+1}=1)&\geqslant P\bigg(F_{21}\geqslant\sum_{j=1}^{d}\eta_j-d\xi\,\bigg|\,Y_{n+1}=1\bigg)\\&=P\bigg(\frac{1}{d}\sum_{j=1}^{d}(T_{n+1,j}-\eta_j)\geqslant-\xi\,\bigg|\,Y_{n+1}=1\bigg)\geqslant 1-\exp\bigg(-\frac{d\xi^2}{2\kappa^2+4\kappa\xi/3}\bigg),\end{aligned}$$

where the last sign of inequality is obtained by Bernstein's inequality. It shows that $\mathrm{P}(d^{-1}F_{21}\geqslant\xi\,|\,Y_{n+1}=1)\to 1$, as $d\to+\infty$.

We next consider the term F_{22}. By Condition (C1), $|\hat{\mu}_{kj}-\mu_{kj}|=O_P(n^{-1/2})$ and $|\tilde{\mu}_{kj}-\mu_{kj}|\leqslant|\hat{\mu}_{kj}-\mu_{kj}|$, we have

$$\begin{aligned}|\tilde{\mu}_{kj}(1-\tilde{\mu}_{kj})|^{-1}&\leqslant\{\mu_{kj}(1-\mu_{kj})-|\tilde{\mu}_{kj}(1-\tilde{\mu}_{kj})-\mu_{kj}(1-\mu_{kj})|\}^{-1}\\&\leqslant\{\mu_{kj}(1-\mu_{kj})-|\tilde{\mu}_{kj}-\mu_{kj}|(|1-2\mu_{kj}|+|\tilde{\mu}_{kj}-\mu_{kj}|)\}^{-1}\\&\leqslant\{\nu(1-\nu)-2|\hat{\mu}_{kj}-\mu_{kj}|\}^{-1}=O_{\mathrm{P}}(1).\end{aligned}$$

Together with $|X_{n+1,j}-\tilde{\mu}_{1j}|\leqslant 1$, we have $|F_{22}|=O_{\mathrm{P}}(dn^{-1/2})=o_{\mathrm{P}}(d)$, as given $Y_{n+1}=1$. Hence, we can obtain that $\mathrm{P}(d^{-1}F_2\geqslant\xi/2\,|\,Y_{n+1}=1)\to 1$, as $d\to+\infty$. As a result, F_2 tends to infinity at the speed of d ($\propto n^{\lambda}$).

Step 3. We then consider F_3 in this step. Write $F_3=F_3^*+(F_3-F_3^*)$, where

$$F_3^*=n\sum_{k=1}^{K}\alpha_k[\pi_{k1}\log(\pi_{kt}^{-1}\pi_{k1})+(1-\pi_{k1})\log\{(1-\pi_{kt})^{-1}(1-\pi_{k1})\}].$$

Based on Condition (C1) and [12, Lemma 3], we immediately know that

$$F_3^*\geqslant\nu n^{1-\gamma}\log\{(1-\nu n^{-\gamma})^{-1}(1+\nu n^{-\gamma})\}=\nu n^{1-\gamma}\log\{1+2\nu n^{-\gamma}(1-\nu n^{-\gamma})^{-1}\}\approx 2\nu^2 n^{1-2\gamma}.$$

As a result, F_3^* tends to infinity at the speed of $n^{1-2\gamma}$ when $\gamma<1/2$.

Next, we can see that $|F_3-F_3^*|$ can be bounded by

$$\begin{aligned}&n\sum_{k=1}^{K}\{|\hat{\alpha}_k\tilde{\pi}_{k1}\log\hat{\pi}_{k1}-\alpha_k\pi_{k1}\log\pi_{k1}|+|\hat{\alpha}_k\tilde{\pi}_{k1}\log\hat{\pi}_{kt}-\alpha_k\pi_{k1}\log\pi_{kt}|\\&\quad+|\hat{\alpha}_k\log(1-\hat{\pi}_{k1})-\alpha_k\log(1-\pi_{k1})|+|\hat{\alpha}_k\tilde{\pi}_{k1}\log(1-\hat{\pi}_{k1})-\alpha_k\pi_{k1}\log(1-\pi_{k1})|\\&\quad+|\hat{\alpha}_k\log(1-\hat{\pi}_{kt})-\alpha_k\log(1-\pi_{kt})|+|\hat{\alpha}_k\tilde{\pi}_{k1}\log(1-\hat{\pi}_{kt})-\alpha_k\pi_{k1}\log(1-\pi_{kt})|\}.\end{aligned}$$

In order to compute the upper bound of $|F_3-F_3^*|$, the following results are needed. Under the condition $\pi_{kt}\propto n^{-\gamma}$ and $|\hat{\pi}_{kt}-\pi_{kt}|=O_P(n^{-1/2})$ by Lemma B.1, we have $|\log(1-\pi_{kt})|=O(n^{-\gamma})$ and

$$|\log\hat{\pi}_{kt}-\log\pi_{kt}|=|\log\{1+\pi_{kt}^{-1}(\hat{\pi}_{kt}-\pi_{kt})\}|=O_{\mathrm{P}}(n^{\gamma-1/2}),$$

$$|\log(1-\hat{\pi}_{kt})-\log(1-\pi_{kt})| = |\log\{1-(1-\pi_{kt})^{-1}(\hat{\pi}_{kt}-\pi_{kt})\}| = O_{\mathrm{P}}(n^{-1/2}).$$

We also have $|\hat{\alpha}_k-\alpha_k| = O_P(n^{-1/2})$ and $|\tilde{\pi}_{k1}-\pi_{k1}| = O_P(n^{-1/2})$ by Lemma B.1. Based on the above equations, we have $|F_3-F_3^*| = O_P(n^{\max\{\gamma,1/2-\gamma,(\log\log n)/\log n+1/2\}})$. Similarly, the order of F_4 is the same as F_3, which can be discussed in the same way.

Consequently, by the condition of Theorem 2.1, (1) $0 \leqslant \gamma < 1/4$ or (2) $1/4 \leqslant \gamma < 1/2 < \lambda \leqslant 1$ or (3) $1/2 \leqslant \gamma < \lambda \leqslant 1$, we have

$$\max\{\lambda, 1-2\gamma\} > \max\{\gamma, 1/2-\gamma, (\log\log n)/\log n + 1/2\},$$

as $n \to +\infty$. Hence, the speed of F_2 or F_3^* can dominate the speed of $|F_3-F_3^*|$, which implies $\mathrm{P}(Q_1-Q_t \geqslant 0\,|\,Y_{n+1}=1) \to 1$ as $n \to \infty$. Thus, the conclusion of Theorem 2.1 is obtained and the proof is completed.