Research Article

Improved Naive Bayes with optimal correlation factor for text classification

Jiangning Chen[1] · Zhibo Dai[1] · Juntao Duan[1] · Heinrich Matzinger[1] · Ionel Popescu[1]

Abstract
Naive Bayes (NB) estimator is widely-used in text classification problems. However, it does not perform well with small-size training datasets. Most previous literature focuses on either creating and modifying features or combing clustering to improve the performance of NB. We directly tackle the problem by constructing a new estimator, called Naive Bayes with correlation factor. We introduce a correlation factor to NB estimator that incorporates overall correlation among the different classes. This effectively exploits the idea of bootstrapping, which reuses data for all classes even if they only belong to one class. Moreover, we obtain a formula for the optimal correlation factor by balancing bias and variance of the estimator. Experimental results on real-world data show that our estimator achieves better accuracy compared with traditional Naive Bayes, yet at the same time maintaining the simplicity of NB.

Keywords Naive Bayes · Correlation factor · Text classification · Insufficient training set

1 Introduction

In recent years, rapid growth of text documents on the Internet and digital libraries has enhanced the importance of text classification, whose goal is to find the categories of each document given their contents. Text classification has many applications in natural language processing, such as topic detection [24], spam filtering [8, 11, 29], author identification [6], web page classification [25] and sentiment analysis [19]. Despite intensive research, it still remains an open problem today.

Although text classification can be realized with schemes having different settings, the fundamental scheme usually consists of two stages: feature generation and classification. In the classification step, there are two optional steps that would benefit the model: feature extraction and feature selection. Many research projects have been done on feature extraction and selection areas, such as some novel feature selection methods proposed by [21, 27, 30]. Other research projects [22] propose a simple heuristic solution of applying a hierarchical tree to assign components to classes, which performs better on large data sets.

For the second stage, classification, it has been studied from both supervised classification and unsupervised clustering. For supervised classification, if we assume all the categories follow independent multinomial distribution and each document is a sample generated by the distribution, a straight-forward idea would be applying some linear model to do classification, such as Support Vector Machine [4, 13], which is used to find the maximum-margin hyper-plane that divides the documents with different labels. Under these assumptions, another important method is Naive Bayes (NB) [7, 15, 26, 31], which uses scores based on the 'probabilities' of each document conditioned on the categories. NB classifier learns from training data to estimate the distribution of each category, then computes the conditional probability of each document

✉ Zhibo Dai, zdai37@gatech.edu; Jiangning Chen, jchen444@math.gatech.edu; Juntao Duan, jt.duan@gatech.edu; Heinrich Matzinger, matzi@math.gatech.edu; Ionel Popescu, ipopescu@math.gatech.edu | [1]School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30313, USA.

SN Applied Sciences (2019) 1:1129 | https://doi.org/10.1007/s42452-019-1153-5

Received: 31 May 2019 / Accepted: 23 August 2019 / Published online: 31 August 2019

given the class label by applying Bayes rule. The prediction of the class is done by choosing the highest posterior probability. When we take into account more factors, such as order of the sequence and meaning of words given a large enough data set, we can use deep learning models such as recurrent neural network [18, 28].

For unsupervised problems, [1] proposed SVD (Singular Value Decomposition) for dimension reduction, then use general clustering algorithm such as K-means. There also exist some algorithms based on EM algorithm, such as pLSA (Probabilistic latent semantic analysis) [10], which considers the probability of each co-occurrence of words and documents as a mixture of conditionally independent multinomial distributions. The parameters in pLSA can not be derived, therefore they used the standard EM algorithm for estimation. Using the same idea, but assuming that the topic distribution has sparse Dirichlet prior, [2] proposed LDA (Latent Dirichlet allocation). The sparse Dirichlet priors encode the intuition that documents cover only a small set of topics and topics frequently use only a small set of words. In practice, this results in a better disambiguation of words and a more precise assignment of documents to topics.

This paper focuses on the performance of Naive Bayes approach for text classification problems. Although it is a widely-used method, it requires plenty of well-labelled data for training purpose. Moreover, its conditional independence assumption is rarely held in reality. There are some researches on how to relax this restriction, such as the feature weighting approach [12, 33] and instance-weighting approach [32]. [3] proposed a method that finds better estimation of centroid, which helps improve the accuracy of Naive Bayes estimation. In order to tackle the situation where there does not exist enough labelled data for each class, we propose a novel estimation method. Different from other feature weighting approaches, our method not only relaxes the conditional independence assumption, but also improves the performance compared to traditional Naive Bayes with insufficient labelled training data for each class. The key part of our method is introducing a correlation factor, which would combine more feature information taken from different classes.

The remainder of this paper is organized as follows. Section 2 details the general settings for the whole paper. In Sect. 3, we give a detail review of Naive Bayes estimator. The error of the Naive Bayes estimator is discussed in Theorem 3.1. In Sect. 4, we derive our proposed method Naive Bayes with correlation factor (NBCF) (see 4.5). It improves traditional Naive Bayes in the situation where there is only limited available training data, which is a common situation in many real world text classification applications. Furthermore, in Theorem 4.1, we show the NBCF error is controlled by the correlation factor and the variation has a smaller order compared with Naive Bayes estimator. Section 5 is devoted to finding the optimal correlation factor. We also show the order of the error is less than traditional naive Bayes. In Sect. 6, we show the results of our simulations, which demonstrates the performance of our method presented in Sect. 4. Finally, Sect. 7 concludes our work and mentions possible future work.

2 General setting

Consider a classification problem with the sample (document) set S, and the class set C with k different classes:

$$C = \{C_1, C_2, \dots, C_k\}.$$

Assume we have totally v different words, thus for each document $d \in S$, we have:

$$d = \{x_1, x_2, \dots, x_v\}.$$

where x_i represents the number of occurrence for i-th word

Define $y = (y_1, y_2, \dots, y_k)$ as our label vector. For document d in class C_i, we have $y_i(d) = 1$. Notice that for a single label problem, we have: $\sum_{i=1}^{k} y_i = 1$.

For a test document d, our target is to predict:

$$\hat{y}(d) = f(d;\theta) = (f_1(d;\theta), f_2(d;\theta), \dots, f_k(d;\theta))$$

given training sample set S, where θ is the parameter matrix and $f_i(d;\theta)$ is the likelihood function of document d in class C_i. The detail proof of Theorems can be found in "Appendix".

3 Naive Bayes classifier in text classification problem

In this section we will discuss the properties of Naive Bayes estimator. The traditional Naive Bayes uses a simple probabilistic model to infer the most likely class of an unknown document. Let class C_i $(1 \le i \le k)$ with centroid $\theta_i = (\theta_{i_1}, \theta_{i_2}, \dots, \theta_{i_v})$ and θ_i satisfies: $\sum_{j=1}^{v} \theta_{i_j} = 1$. Assuming independence of the words, the most likely class for a document d is computed as:

$$\begin{aligned} label(d) &= argmax_i P(C_i) P(d|C_i) \\ &= argmax_i P(C_i) \prod_{j=1}^{v} (\theta_{i_j})^{x_j} \\ &= argmax_i \log P(C_i) + \sum_{j=1}^{v} x_j \log \theta_{i_j}. \end{aligned} \tag{3.1}$$

This gives the classification criteria once θ is estimated, namely finding the largest among

$$\log f_i(d;\theta) = \log P(C_i) + \sum_{j=1}^{v} x_j \log \theta_{i_j} \qquad 1 \le i \le k$$

Now we shall derive an maximum likelihood estimator for θ. For a class C_i, we have the standard likelihood function:

$$\begin{aligned} L(C_i, \theta) &= \prod_{d \in S} f_i(d;\theta)^{y_i(d)} \\ &= \prod_{d \in C_i} \prod_{j=1}^{v} \theta_{i_j}^{x_j} \end{aligned} \tag{3.2}$$

Take logarithm for both sides, we obtain the log-likelihood function:

$$\log L(C_i, \theta) = \sum_{d \in C_i} \sum_{j=1}^{v} x_j \log \theta_{i_j}. \tag{3.3}$$

We would like to solve optimization problem:

$$\begin{aligned} &\max \log L(C_i, \theta) \\ \text{subject to :} &\sum_{j=1}^{v} \theta_{i_j} = 1 \\ &\theta_{i_j} \ge 0 \end{aligned} \tag{3.4}$$

The problem (3.4) can be explicitly solved by Lagrange Multiplier, for class C_i, we have $\theta_i = \{\theta_{i_1}, \theta_{i_2}, \dots, \theta_{i_v}\}$, where:

$$\hat{\theta}_{i_j} = \frac{\sum_{d \in C_i} x_j}{\sum_{d \in C_i} \sum_{j=1}^{v} x_j}. \tag{3.5}$$

For estimator $\hat{\theta}$, we have following theorem.

Theorem 3.1 *Assume we have normalized length of each document, that is: $\sum_{j=1}^{v} x_j = m$ for all documents $d \in S$, the estimator [(3.5) satisfies following properties:*

(1) *$\hat{\theta}_{i_j}$ is unbiased.*
(2) *$E[|\hat{\theta}_{i_j} - \theta_{i_j}|^2] = \frac{\theta_{i_j}(1-\theta_{i_j})}{|C_i|m}$.*

The Naive Bayes with multinomial prior distribution has a strong assumption about the data: it assumes that words in documents are independent. However, this assumption clearly does not hold in real world text. There are many different kinds of dependence between words induced by semantic, pragmatic, and conversational structure of a text. Although it has its advantages in practice compared to some more sophisticated models, we propose a new method based on Naive Bayes model that has a better performance by introducing a correlation factor, especially for the situation where there is no sufficient data compared with large number of classes.

4 Naive Bayes with correlation factor

From Theorem 3.1, we can see that traditional Naive Bayes estimator $\hat{\theta}$ is an unbiased estimator with variance $O(\frac{\theta_{i_j}(1-\theta_{i_j})}{|C_i|m})$. Now we will try to find an estimator, and prove that it can perform better than traditional Naive Bayes estimator.

There are two main approaches to improve the Naive Bayes model: modifying the feature and modifying the model. Many researchers have proposed approaches to modify the document representation in order to better fit the assumption made by Naive Bayes. These include extracting more complex features such as syntactic or statistical phrases [20], extracting features using word clustering [5] and exploiting relations using lexical resources [9]. We propose an approach that modifies the probabilistic model. Therefore, our model should work well with other document representation modifications to achieve a better result.

Our basic idea is that, even for a single labeling problem, a document d usually contains words appearing in different classes, thus it should include some information from different classes. However, our label y in training set does not reflect that information because only one component of y is 1 and all others are 0. We would like to replace y by $y + t$ in Naive Bayes likelihood function 3.2 with some optimized t to get our new likelihood function L_1:

$$\begin{aligned} L_1(C_i, \theta) &= \prod_{d \in S} f_i(d;\theta)^{y_i(d)+t} \\ &= \prod_{d \in S} \left(\prod_{j=1}^{v} \theta_{i_j}^{x_j} \right)^{y_i(d)+t}. \end{aligned} \tag{4.1}$$

By introducing the correlation factor t, we include more information between the document and classes, which improves the classification accuracy.

Notice that to compute L_1 of a given class C_i in our estimator, instead of just using documents in C_1 as Naive Bayes estimator, we will use every $d \in S$.

Take logarithm for both sides of 4.1, we obtain the log-likelihood function:

$$\log L_1(C_i, \theta) = \sum_{d \in S} \left[(y_i(d) + t) \sum_{j=1}^{v} x_j \log \theta_{i_j} \right]. \tag{4.2}$$

Similar to Naive Bayes estimator, we would like to solve optimization problem:

$$\begin{aligned} &\max \log L_1(C_i, \theta) \\ \text{subject to}: &\sum_{j=1}^{v} \theta_{i_j} = 1 \\ &\theta_{i_j} \geq 0 \end{aligned} \tag{4.3}$$

Let:

$$G_i = 1 - \sum_{j=1}^{v} \theta_{i_j},$$

by Lagrange multiplier, we have:

$$\begin{cases} \dfrac{\partial \log(L_1)}{\partial \theta_{i_j}} + \lambda_i \dfrac{\partial G_i}{\partial \theta_{i_j}} = 0 \, \forall \, 1 \leq i \leq k \text{ and } \forall \, 1 \leq j \leq v \\ \displaystyle\sum_{j=1}^{v} \theta_{i_j} = 1, \, \forall \, 1 \leq i \leq k \end{cases}$$

plug in, we obtain:

$$\begin{cases} \displaystyle\sum_{d \in S} \frac{(y_i(d)+t)x_j}{\theta_{i_j}} - \lambda_i = 0, \, \forall \, 1 \leq i \leq k \text{ and } \forall \, 1 \leq j \leq v \\ \displaystyle\sum_{j=1}^{v} \theta_{i_j} = 1, \, \forall \, 1 \leq i \leq k \end{cases} \tag{4.4}$$

Solve (4.4), we got the solution of optimization problem (4.3):

$$\begin{aligned} \hat{\theta}_{i_j}^{L_1} &= \frac{\sum_{d \in S}(y_i(d)+t)x_j}{\sum_{j=1}^{v} \sum_{d \in S}(y_i(d)+t)x_j} \\ &= \frac{\sum_{d \in S}(y_i(d)+t)x_j}{m(|C_i| + t|S|)} \end{aligned} \tag{4.5}$$

For estimator $\hat{\theta}_{i_j}^{L_1}$, we have the following result:

Theorem 4.1 *Assume for each class, we have prior distributions $p_1, p_2, \ldots, p_k$ with $p_i = |C_i|/|S|$, and we have normalized length for each document, that is: $\sum_{j=1}^{v} x_j = m$. The estimator (4.5) satisfies following property:*

(1) $\hat{\theta}_{i_j}^{L_1}$ is biased, with: $|E[\hat{\theta}_{i_j}^{L_1}] - \theta_{i_j}| = O(t)$
(2) $E[|\hat{\theta}_{i_j}^{L_1} - E[\hat{\theta}_{i_j}^{L_1}]|^2] = O(\frac{1}{m|S|})$.

We can see that $E[|\hat{\theta}_{i_j}^{L_1} - E[\hat{\theta}_{i_j}^{L_1}]|^2]$ is in $O(\frac{1}{|S|})$, which means it converges faster than standard Naive Bayes $O(\frac{1}{|C_i|})$, however, since $E[|\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}|] \neq 0$, it is not an unbiased estimator.

5 Determine the correlation factor

In general statistical estimation theory, a biased estimator is acceptable, and sometimes even outperforms an unbiased estimator. A more important perspective is to find a suitable loss function to determine parameters. Li and Yang [17] introduces ways of choosing loss function for many famous models, the common idea is to sum a biased term and complexity penalty term for model parameters. Nigam et al. [23] uses the maximum entropy as the loss function for text classification problems.

In our problem, from 3.1 and 4.1, we know that traditional Naive Bayes estimator is unbiased. Our estimator is biased, but we want to find an optimal t to get smaller variance. In order to balance the trade-off between bias and variance, we would like to select a loss function which takes into account of both bias and variance.

In this task, we can use mean squared error as loss function. There is a well-known bias-variance decomposition for mean square error.

$$\begin{aligned} E\left[|\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}|^2\right] &= E\left[|\hat{\theta}_{i_j}^{L_1} - E\hat{\theta}_{i_j}^{L_1} + E\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}|^2\right] \\ &= E\left[|\hat{\theta}_{i_j}^{L_1} - E\hat{\theta}_{i_j}^{L_1}|^2\right] + \left[E\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}\right]^2 \\ &:= Var\left(\hat{\theta}_{i_j}^{L_1}\right) + Bias\left(\hat{\theta}_{i_j}^{L_1}\right)^2 \end{aligned}$$

In practice, we would like to minimize a general linear combination of bias and variance, namely,

$$L\left(\theta_{i_j}, c_1, c_2\right) = c_1 Bias\left(\hat{\theta}_{i_j}^{L_1}\right)^2 + c_2 Var\left(\hat{\theta}_{i_j}^{L_1}\right) \tag{5.1}$$

Theorem 5.1 *The minimum of the loss function 5.1 is achieved at*

$$t^* = \frac{c_2(1-p_i)\theta_{i_j}(1-\theta_{i_j})}{c_2 \sum_{l=1}^{k} p_l \theta_{l_j}(1-\theta_{l_j}) - c_2\theta_{i_j}(1-\theta_{i_j}) + c_1 m|S|(\sum_{l=1}^{k} p_l \theta_{l_j} - \theta_{i_j})^2} \tag{5.2}$$

We can see from (5.2) that the optimal correlation factor t^* should be a very small number close to $O(\frac{1}{m|S|})$. Therefore by Eq. A.2, we know the squared bias is

$$Bias\left(\hat{\theta}_{i_j}^{L_1}\right)^2 = O(t^2) = O\left(\frac{1}{m^2|S|^2}\right)$$

We have already shown in A.3 that the order of the variance

$$Var\left(\hat{\theta}^{L_1}_{i_j}\right) = O\left(\frac{1}{m|S|}\right)$$

Therefore in the case of expected square error $E[|\hat{\theta}^{L_1}_{i_j} - \theta_{i_j}|^2]$ $(c_1 = c_2 = 1)$ is dominated by the variance. Thus we have the following Corollary:

Theorem 5.2 *With any selection of* $t = O(\frac{1}{m|S|})$, *we have*

$$E\left[|\hat{\theta}^{L_1}_{i_j} - \theta_{i_j}|^2\right] = O\left(\frac{1}{m|S|}\right) \tag{5.3}$$

By Theorem 3.1, we know that for Naive Bayes, $E[|\hat{\theta}_{i_j} - \theta_{i_j}|^2] = O(\frac{1}{|C_i|})$, thus we can see that our estimator actually works better.

6 Experiment

6.1 Simulation with different correlation factor

In the previous section we obtained that the order of t must be $O(\frac{1}{|S|})$. However, we will still need to determine how to choose the best correlation factor t. That we will have to tune the parameter by running t in some determined interval.

We applied our method on single labeled documents of 10 topics, which have almost the same sample size, in Reuters-21578 data [16], there are approximately 3000 documents in this sample set. For 20 news group data [14], it includes 20 groups and approximately 20,000 documents.

We take $t \in (0, 2)$ and use 10% of data for training and assess the trained model on the remaining test data.

In our simulation, we notice that when we choose correlation factor to be around 0.1, we get the best accuracy for our estimation. See Fig. 1a, b.

6.2 Compare with Naive Bayes

Next, we compare the result of traditional Naive Bayes estimator (3.5) $\hat{\theta}_{i_j}$and our estimator (4.5) $\hat{\theta}^{L_1}_{i_j}$. In this simulation, our correlation factor t is chosen to be 0.1 for Figs. 2, 3 and 4.

First, we run both algorithms on these two sample datasets. We know that when the sample size becomes large enough, our estimator is baised. But when the training set is small, our estimator should converge faster. Thus we first take the training size relatively small (10%). See Fig. 2a, b. According to the simulation, we can see our method is more accurate for most of the classes, and more accurate on average.

Then we test our estimator $\hat{\theta}^{L_1}$ with larger training set (90%). In our analysis above, we know that as datasets become large enough, our estimator converges to a biased estimator, so we expect a better result with traditional Naive Bayes estimator. See Fig. 3a, b. According to the simulation, we can see for 20 news group, traditional Naive Bayes performs better than our method, but our method

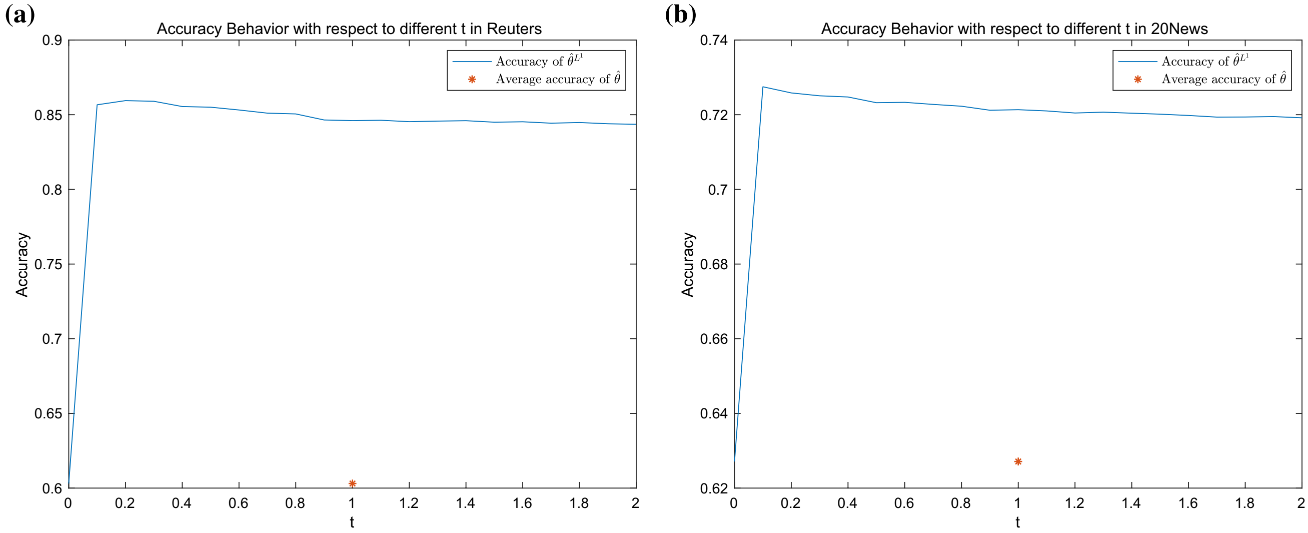

Fig. 1 We test accuracy behavior with respect to different correlation factors in Reuter-21578 (**a**) and 20 News group dataset (**b**). We take 10% of the data as training set. The y-axis is the accuracy and the x-axis is the correlation factor t

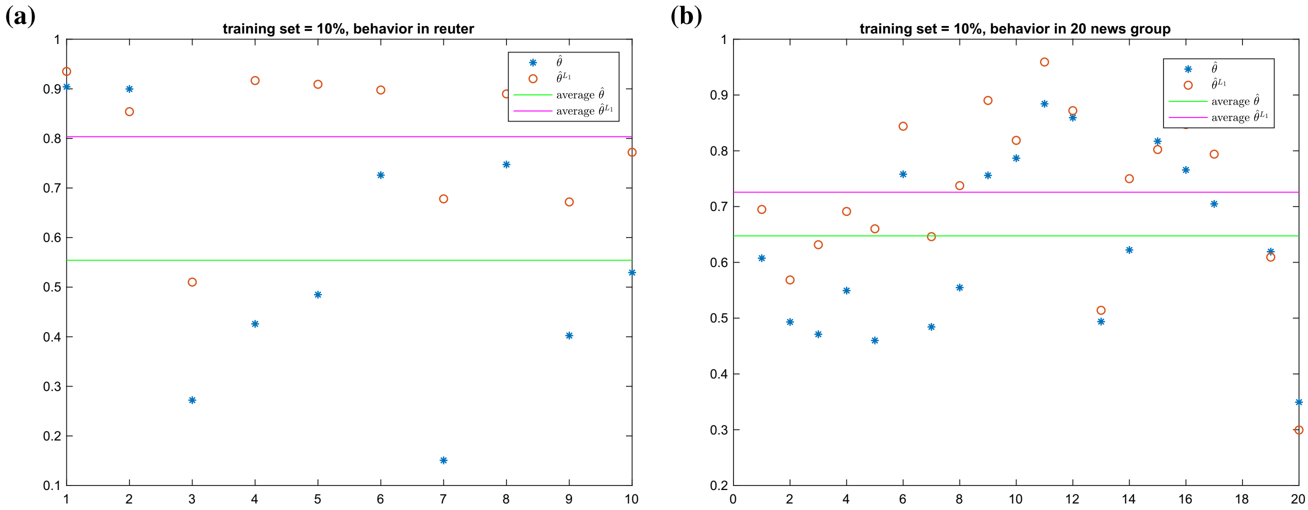

Fig. 2 We take 10 largest groups in Reuter-21578 dataset (**a**) and 20 news group dataset (**b**), and take 10% of the data as training set. The y-axis is the accuracy, and the x-axis is the class index

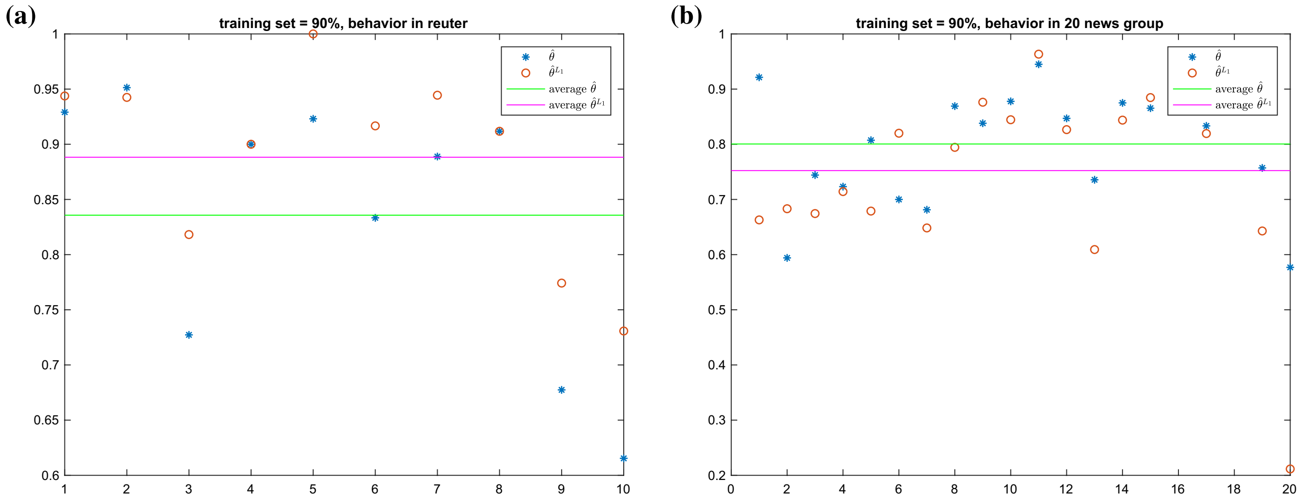

Fig. 3 We take 10 largest groups in Reuter-21578 dataset (**a**) and 20 news group dataset (**b**), and take 90% of the data as training set. The y-axis is the accuracy, and the x-axis is the class index

is still more accurate than Naive Bayes in Reuter's data. The reason might be that we have a huge unbalanced sample size in Reuter's data, 90% of the training set is still not large enough for many classes.

Finally, we apply the same training set with training size 10% and test the accuracy on the training set instead of the test set. We find the traditional Naive Bayes estimator actually achieves better results, which means it might have more over-fitting problems. This might be the reason why our method works better when the dataset is not too large: adding the correlation factor t helps us bring some uncertainty in training process, which helps avoid over-fitting. See Fig. 4a, b.

6.3 Robustness of t for prediction

For estimation purposes, t must satisfy $t = O(\frac{1}{|S|})$ by Theorem 5.2 in order to find the most accurate parameters. However, it turns out that for prediction purposes, there is a phase transition phenomenon. As long as $t \geq O(\frac{1}{|S|})$, the prediction power is not reduced even when we increase t to very large value (for example $t = 10^5$). In the simulation of finding best t in Fig. 1a, b, we see the testing error is only decreasing slightly as t increasing from 0.1 to 2. We summarize this fact as follows

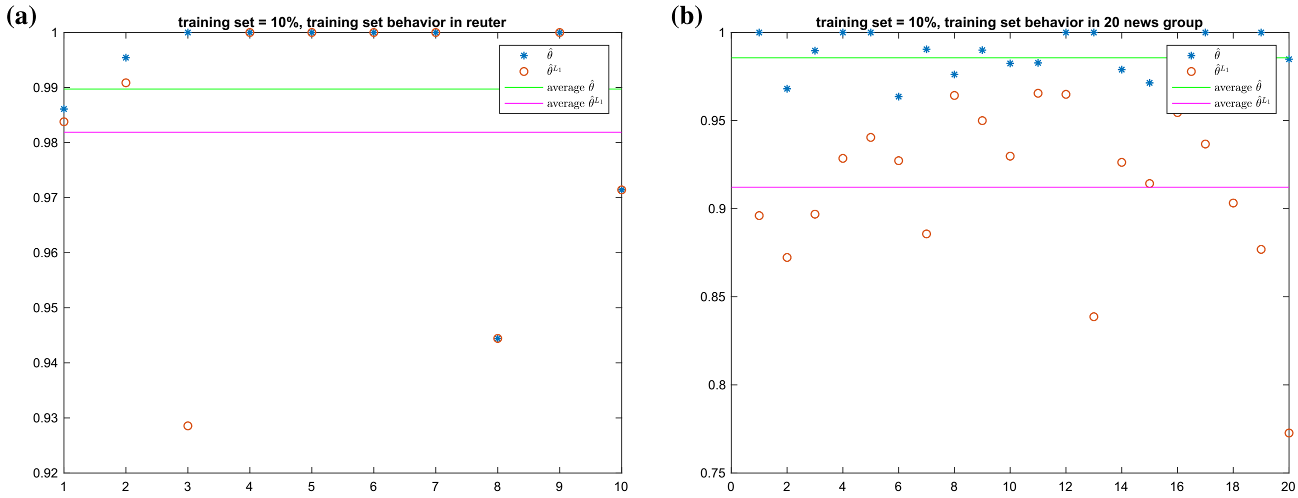

Fig. 4 We take 10 largest groups in Reuter-21578 dataset (**a**), and 20 news group dataset (**b**), and take 10% of the data as training set. We test the result on training set. The y-axis is the accuracy, and the x-axis is the class index

Proposition 6.1 *For prediction purpose, the correlation factor t can take value in the interval*

$$\frac{1}{|S|} \leq t \leq 1$$

The reason we restraint the upper bound to be 1 is that the effect of correlation factor should not exceed the effect of original class $y_i = 1$.

7 Conclusion

In this paper, we modified the traditional Naive Bayes estimator with a correlation factor to obtain a new estimator, NBCF, which is biased but has a smaller variance. We justified that our estimator has significant advantage by analyzing the mean square error. In simulation, we applied our method to real world text classification problems, and showed that it works better when the training data set is insufficient.

There are several important open problems related our estimator:

(1) Is our proposed estimator admissible for the square error loss? Even though we know it outperform Naive Bayes estimator, it might not be the optimal one.

(2) Will our estimator still work in other more independent datasets? We only test our result in Reuter's data [16] and 20 news group [14], these datasets are news from newspapers, which means they are highly correlated to each other.

(3) We can only use our method in single labeled dataset so far, it would be interesting to see if we can extend our result in partial labeled dataset or multi-labeled dataset.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix A: Proof of the theorems

A.1. Proof of Theorem 3.1

Proof With assumption $\sum_{j=1}^{v} x_j = m$, we can rewrite (3.5) as:

$$\hat{\theta}_{i_j} = \frac{\sum_{d \in C_i} x_j}{\sum_{d \in C_i} m} = \frac{\sum_{d \in C_i} x_j}{|C_i| m}.$$

Since $d = (x_1, x_2, \dots, x_v)$ is multinomial distribution in class C_i, we have: $E[x_j] = m\theta_{i_j}$, and $E[x_j^2] = m\theta_{i_j}(1 - \theta_{i_j} + m\theta_{i_j})$.

(1)
$$E\left[\hat{\theta}_{i_j}\right] = E\left[\frac{\sum_{d \in C_i} x_j}{|C_i| m}\right] = \frac{\sum_{d \in C_i} E[x_j]}{|C_i| m}$$
$$= \frac{\sum_{d \in C_i} m\theta_{i_j}}{|C_i| m} = \theta_{i_j}.$$

Thus $\hat{\theta}_{i_j}$ is unbiased.

(2) By (1), we have:

$$E\left[|\hat{\theta}_{i_j} - \theta_{i_j}|^2\right] = E[\hat{\theta}_{i_j}^2] - 2\theta_{i_j} E\left[\hat{\theta}_{i_j}\right] + \theta_{i_j}^2 = E\left[\hat{\theta}_{i_j}^2\right] - \theta_{i_j}^2.$$

Then notice

$$\begin{aligned}\hat{\theta}_{i_j}^2 &= \frac{\left(\sum_{d\in C_i} x_j\right)^2}{|C_i|^2 m^2} \\ &= \frac{\sum_{d\in C_i} x_j^2 + \sum_{d\neq d'\in C_i} x_j^d x_j^{d'}}{|C_i|^2 m^2},\end{aligned} \tag{A.1}$$

where $d = (x_1^d, x_2^d, \dots, x_v^d)$.

Since:

$$\begin{aligned}E\left[\frac{\sum_{d\in C_i} x_j^2}{|C_i|^2 m^2}\right] &= \frac{|C_i| m\theta_{i_j}\left(1 - \theta_{i_j} + m\theta_{i_j}\right)}{|C_i|^2 m^2} \\ &= \frac{\theta_{i_j}\left(1 - \theta_{i_j} + m\theta_{i_j}\right)}{|C_i| m},\end{aligned}$$

and

$$E\left[\frac{\sum_{d\neq d'\in C_i} x_j^d x_j^{d'}}{|C_i|^2 m^2}\right] = \frac{|C_i|\left(|C_i| - 1\right) m^2 \theta_{i_j}^2}{|C_i|^2 m^2} = \frac{\left(|C_i| - 1\right)\theta_{i_j}^2}{|C_i|}.$$

Plugging them into (A.1) obtains:

$$E\left[\hat{\theta}_{i_j}^2\right] = \frac{\theta_{i_j}\left(1 - \theta_{i_j}\right)}{|C_i| m} + \theta_{i_j}^2,$$

thus: $E\left[|\hat{\theta}_{i_j} - \theta_{i_j}|^2\right] = \frac{\theta_{i_j}\left(1 - \theta_{i_j}\right)}{|C_i| m}$.

□

A.2. Proof of Theorem 4.1

Proof

(1) With assumption $\sum_{j=1}^{v} x_j = m$, we have:

$$\begin{aligned}E[\hat{\theta}_{i_j}^{L_1}] &= \frac{\sum_{d\in S}(y_i(d) + t)E[x_j]}{m(t|S| + |C_i|)} \\ &= \frac{\sum_{d\in S} tE[x_j] + \sum_{x\in C_i} E[x_j]}{m(t|S| + |C_i|)} \\ &= \frac{t\sum_{l=1}^{k} |C_l|\theta_{l_j} + \theta_{i_j}|C_i|}{t|S| + |C_i|} \\ &= \frac{t|S|\sum_{l=1}^{k} p_l\theta_{l_j} + \theta_{i_j}|C_i|}{t|S| + |C_i|}\end{aligned}$$

Thus:

$$\begin{aligned}|E[\hat{\theta}_{i_j}^{L_1}] - \theta_{i_j}| &= \frac{t|S||\sum_{l=1}^{k} p_l\theta_{l_j} - \theta_{i_j}|}{t|S| + |C_i|} \\ &= \frac{|\sum_{l=1}^{k} p_l\theta_{l_j} - \theta_{i_j}|}{1 + p_i/t} \\ &= O(t).\end{aligned} \tag{A.2}$$

This shows our estimator is biased. The error is controlled by t. When t converges to 0, our estimator converges to the unbiased Naive Bayes estimator. We can also derive a lower bound for the square error:

$$\begin{aligned}E\left[|\hat{\theta}_{i_j}^{L_1} - \theta_{i_j}|^2\right] &\geq \left(E\left[\hat{\theta}_{i_j}^{L_1}\right] - \theta_{i_j}\right)^2 \\ &= \frac{|\sum_{l=1}^{k} p_l\theta_{l_j} - \theta_{i_j}|^2}{(1 + p_i/t)^2}\end{aligned}$$

(2) For variance part, since

$$\hat{\theta}_{i_j}^{L_1} = \frac{\sum_{d\in S}(y_i(d) + t)x_j}{m(|C_i| + t|S|)},$$

we have:

$$\begin{aligned}&E[|\hat{\theta}_{i_j}^{L_1} - E[\hat{\theta}_{i_j}^{L_1}]|^2] \\ &= E\left[\left|\frac{\sum_{d\in S}(y_i(d) + t)(x_j - E[x_j])}{m(|C_i| + t|S|)}\right|^2\right] \\ &= \frac{\sum_{d\in S}(y_i(d) + t)^2 E(|x_j - E[x_j]|^2)}{m^2(|C_i| + t|S|)^2} \\ &= \frac{\sum_{d\in C_i}(1 + t)^2 m\theta_{i_j}(1 - \theta_{i_j}) + \sum_{d\in C_l, l\neq i} t^2 m\theta_{l_j}(1 - \theta_{l_j})}{m^2(|C_i| + t|S|)^2} \\ &= \frac{|C_i|(1 + 2t)\theta_{i_j}(1 - \theta_{i_j}) + \sum_{l=1}^{k} |C_l| t^2\theta_{l_j}(1 - \theta_{l_j})}{m(|C_i| + t|S|)^2} \\ &= \frac{|S|p_i(1 + 2t)\theta_{i_j}(1 - \theta_{i_j}) + |S|\sum_{l=1}^{k} p_l t^2\theta_{l_j}(1 - \theta_{l_j})}{m(|S|p_i + t|S|)^2} \\ &= \frac{p_i(1 + 2t)\theta_{i_j}(1 - \theta_{i_j}) + \sum_{l=1}^{k} p_l t^2\theta_{l_j}(1 - \theta_{l_j})}{m|S|(p_i + t)^2} \\ &= O\left(\frac{1}{m|S|}\right)\end{aligned} \tag{A.3}$$

□

A.3. Proof of Theorem 5.1

Proof First let us fix some notations for constants do not involve t to simplify the derivation. Let $\Theta_{i_j} := \theta_{i_j}(1-\theta_{i_j})$, $A = \sum_{l=1}^{k} p_l\theta_{l_j}(1-\theta_{l_j})$ and $B = (\sum_{l=1}^{k} p_l\theta_{l_j} - \theta_{i_j})^2$. As shown in Eq. A.2, the squared bias is

$$Bias\left(\hat{\theta}^{L_1}_{i_j}\right)^2 = \left[E\hat{\theta}^{L_1}_{i_j} - \theta_{i_j}\right]^2 = \frac{m|S|t^2\left(\sum_{l=1}^{k} p_l\theta_{l_j} - \theta_{i_j}\right)^2}{m|S|(p_i+t)^2} = \frac{t^2 m|S|B}{m|S|(p_i+t)^2}$$

From A.3, the variance is

$$Var\left(\hat{\theta}^{L_1}_{i_j}\right) = E\left[|\hat{\theta}^{L_1}_{i_j} - E\hat{\theta}^{L_1}_{i_j}|^2\right] = \frac{p_i(1+2t)\theta_{i_j}\left(1-\theta_{i_j}\right) + t^2\sum_{l=1}^{k} p_l\theta_{l_j}(1-\theta_{l_j})}{m|S|(p_i+t)^2} = \frac{p_i(1+2t)\Theta_{i_j} + t^2 A}{m|S|(p_i+t)^2}$$

Therefore,

$$L\left(\theta_{i_j}, c_1, c_2\right) = \frac{c_2 p_i(1+2t)\Theta_{i_j} + t^2(c_2 A + c_1 m|S|B)}{m|S|(p_i+t)^2}$$

Then we should optimize t to minimize the loss $L(\theta_{i_j}, c_1, c_2)$. Taking derivative with respect to t and setting it to be 0. We find

$$\left[c_2 p_i\Theta_{i_j} + t(c_2 A + c_1 m|S|B)\right](p_i+t) - \left[c_2 p_i(1+2t)\Theta_{i_j} + t^2(c_2 A + c_1 m|S|B)\right] = 0.$$

That simplifies to

$$c_2(p_i - 1 - t)\Theta_{i_j} + t(c_2 A + c_1 m|S|B) = 0$$

which shows

$$t = \frac{c_2(1-p_i)\Theta_{i_j}}{c_2(A - \Theta_{i_j}) + c_1 m|S|B}.$$

Plug in original parameters, we obtain

$$t = \frac{c_2(1-p_i)\theta_{i_j}(1-\theta_{i_j})}{c_2\sum_{l=1}^{k} p_l\theta_{l_j}(1-\theta_{l_j}) - c_2\theta_{i_j}(1-\theta_{i_j}) + c_1 m|S|\left(\sum_{l=1}^{k} p_l\theta_{l_j} - \theta_{i_j}\right)^2}$$

□

References

1. Albright R (2004) Taming text with the SVD. SAS Institute Inc, Cary
2. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3(Jan):993–1022
3. Chen J, Matzinger H, Zhai H, Zhou M (2018) Centroid estimation based on symmetric KL divergence for multinomial text classification problem. In: 2018 17th IEEE international conference on machine learning and applications (ICMLA). IEEE, pp 1174–1177
4. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
5. Dhillon IS, Mallela S, Kumar R (2003) A divisive information-theoretic feature clustering algorithm for text classification. J Mach Learn Res 3(Mar):1265–1287
6. Feng X, Qin H, Shi Q, Zhang Y, Zhou F, Haochen W, Ding S, Niu Z, Yan L, Shen P (2014) Chrysin attenuates inflammation by regulating m1/m2 status via activating pparγ. Biochem Pharmacol 89(4):503–514
7. Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Mach Learn 29(2–3):131–163
8. Günal S, Ergin S, Bilginer GM, Gerek Ö (2006) On feature extraction for spam e-mail detection. In: International workshop on multimedia content representation, classification and security. Springer, pp 635–642
9. Hidalgo JMG, Rodriguez MdB (1997) Integrating a lexical database and a training collection for text categorization. arXiv preprint arXiv:cmp-lg/9709004
10. Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of the fifteenth conference on uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., pp 289–296
11. Idris I, Selamat A, Omatu S (2014) Hybrid email spam detection model with negative selection algorithm and differential evolution. Eng Appl Artif Intell 28:97–110
12. Jiang L, Li C, Wang S, Zhang L (2016) Deep feature weighting for Naive Bayes and its application to text classification. Eng Appl Artif Intell 52:26–39
13. Joachims T (1998) Text categorization with support vector machines: learning with many relevant features. In: European conference on machine learning. Springer, pp 137–142
14. Lang K 20 newsgroups data set. http://qwone.com/~jason/20Newsgroups/
15. Langley P, Iba W, Thompson K et al (1992) An analysis of Bayesian classifiers. Aaai 90:223–228
16. Lewis DD Reuters-21578. http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
17. Li F, Yang Y (2003) A loss function analysis for classification methods in text categorization. In: Proceedings of the 20th international conference on machine learning (ICML-03), pp 472–479
18. Liu P, Qiu X, Huang X (2016) Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101
19. Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 5(4):1093–1113

20. Mladenic D, Grobelnik M (1998) Word sequences as features in text-learning. In: Proceedings of the 17th electrotechnical and computer science conference (ERK98). Citeseer
21. Narayanan V, Arora I, Bhatia A (2013) Fast and accurate sentiment classification using an enhanced naive bayes model. In: International conference on intelligent data engineering and automated learning. Springer, pp 194–201
22. Nguyen N, Yamada K, Suzuki I, Unehara M (2018) Hierarchical scheme for assigning components in multinomial naive bayes text classifier. In: 2018 Joint 10th international conference on soft computing and intelligent systems (SCIS) and 19th international symposium on advanced intelligent systems (ISIS), pp 335–340
23. Nigam K, Lafferty J, McCallum A (1999) Using maximum entropy for text classification. In: IJCAI-99 workshop on machine learning for information filtering, vol 1, pp 61–67
24. Rill S, Reinel D, Scheidt J, Zicari RV (2014) Politwi: early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. Knowl-Based Syst 69:24–33
25. Saraç E, Özel SA (2014) An ant colony optimization based feature selection for web page classification. Sci World J. https://doi.org/10.1155/2014/649260
26. Su J, Shirab JS, Matwin S (2011) Large scale text classification using semi-supervised multinomial Naive Bayes. In: Proceedings of the 28th international conference on machine learning (ICML-11). Citeseer, pp 97–104
27. Tang B, Kay S, He H (2016) Toward optimal feature selection in Naive Bayes for text categorization. IEEE Trans Knowl Data Eng 28(9):2508–2521
28. Tang D, Qin B, Liu T (2015) Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 1422–1432
29. Uysal AK, Gunal S, Ergin S, Sora Gunal E (2013) The impact of feature extraction and selection on sms spam filtering. Elektron Elektrotech 19(5):67–72
30. Uysal AK (2016) An improved global feature selection scheme for text classification. Expert Syst Appl 43:82–92
31. Venkatesh, RKV (2018) Classification and optimization scheme for text data using machine learning nave bayes classifier. In: 2018 IEEE world symposium on communication engineering (WSCE), pp 33–36
32. Wu Y (2018) A new instance-weighting Naive Bayes text classifiers. In: 2018 IEEE international conference of intelligent robotic and control engineering (IRCE), pp 198–202
33. Zhang L, Jiang L, Li C, Kong G (2016) Two feature weighting approaches for Naive Bayes text classifiers. Knowl-Based Syst 100:137–144

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.