Web-Scale Classification with Naive Bayes

Congle Zhang, Gui-Rong Xue, Yong Yu
Shanghai Jiao Tong University
Shanghai, 200240, China
{zhangcongle, grxue, yyu}@apex.sjtu.edu.cn

Hongyuan Zha
College of Computing Georgia Institute of Technology
Atlanta, GA - 30332
zha@cc.gatech.edu

ABSTRACT

Traditional Naive Bayes Classifier performs miserably on web-scale taxonomies. In this paper, we investigate the reasons behind such bad performance. We discover that the low performance are not completely caused by the intrinsic limitations of Naive Bayes, but mainly comes from two largely ignored problems: *contradiction pair problem* and *discriminative evidence cancelation problem*. We propose modifications that can alleviate the two problems while preserving the advantages of Naive Bayes. The experimental results show our modified Naive Bayes can significantly improve the performance on real web-scale taxonomies.

Categories and Subject Descriptors: I.2.6 [Artificial Intelligence]: Learning

General Terms: Algorithms, Experimentation

Keywords: Naive Bayes, Web-scale Taxonomy

1. INTRODUCTION

Text classification is widely used in information retrieval area and text mining area, especially in the web environment. Web-scale taxonomy classification entails high efficiency of the classifier, and Naive Bayes Classifier (NBC) naturally lends itself to such tasks because it is simple, fast, easy to implement and relatively effective. NBC has been studied for a long time in small datasets[2]. Unfortunately, when employing NBC on tasks with thousands classes, we found that it achieves extremely bad performance which we will discuss in detail later. One may think this poor performance must come from some intrinsic limitations of NBC (say, the independent assumption). And this gives the motivation in some works to try to use extra information (e.g., the hierarchy structure of the taxonomy)[3] or other complicated classification algorithms to handle the case of a large number of classes [1].

In this paper, we aim to maintain the advantage of NBC (e.g. simple algorithm, easy implementation and fast computation), and to achieve good performance at the same time. We discover that two largely ignored problems of NBC can severely hurt its classification performance. We call them *contradiction pair problem* and *discriminative evidence cancelation problem*, which will be discussed in Section 2. These two problems have the characteristic that they are rather benign for small number of classes but manifest themselves notably in the presence of a large number of classes.

To deal with the two problems, we propose two different modifications on NBC: Weight Manipulation Naive Bayes and Parametric Smoothing Naive Bayes. Our modified NBC can improve the accuracy from 9% to about 50% for a 1000-class case, and from 20% to about 70% for a 200-class case, while maintaining all the advantages of NBC. Moreover, each of them is designed to fix the two problems without making NBC much slower or significantly more difficult to implement.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

2. TWO PROBLEMS OF NAIVE BAYES

For word w_i and class c, standard NBC based on multinomial model and Laplacian smoothing yields maximum likelihood estimation of the class conditional probability $p^{ml}(w_i|c) = N_c^i/N_c$, and smoothed estimate $p^s(w_i|c) = \frac{N_c^i+1}{N_c+V}$. To classify a test document d, NBC assigns d by the label $\ell(d) = \arg\max_c\{\log p(c) + \sum_{w_i \in d} f_i \log p(w_i|c)\}$. Here N_c^i is the frequency count of the word w_i appearing in training set of class c, $N_c = \sum_i N_c^i$, f_i is the occurrence of w_i in the test document d.

The first problem is what we call the *contradiction pair problem*. It means, for a word and the classes, the smoothed estimates do not preserve the order of the maximum likelihood estimates (MLE), i.e. for a word w and two classes c_1, c_2, we may have MLEs with $p^{ml}(w|c_1) > p^{ml}(w|c_2)$ but the smoothed estimates with $p^s(w|c_2) < p^s(w|c_2)$. Figure 2 provides an illustration of this issue. Since NBC relies on smoothed estimates to make class predictions, the consequence of contradiction pair problem is that NBC loses some evidences carried by the maximum likelihood estimates.

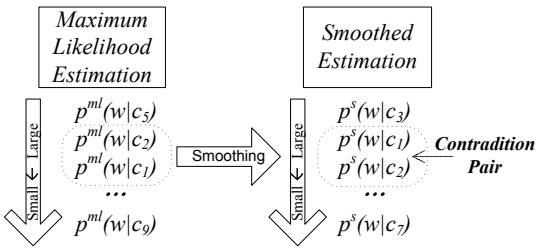


Figure 1: Illustration of contradiction pair problem.

We call the second problem *discriminative evidence cancelation problem* which can be illustrated using a simple example in Figure 2. Assume a document with three words $d = \{w_1, w_2, w_3\}$ and class c is one of the candidate classes

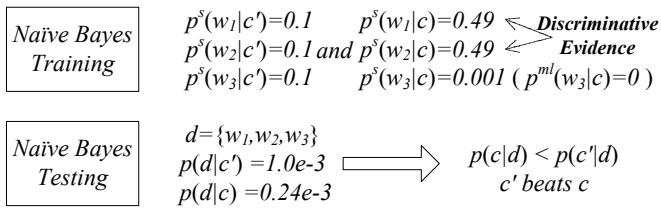


Figure 2: Illustration of discriminative evidence cancelation problem where we assumed $p(c) = p(c')$. Besides, there is $p(c|d) = p(d|c)p(c)/p(d)$

to label d, suppose $\{(w_1, c), (w_2, c)\}$ are *discriminative evidence*. The smoothed estimation $p^s(w_i|c)$ is discriminative evidence when $p^s(w_1|c)$ is far larger than the average of $p^s(w_1|c_j)$, i.e. $p^s(w_1|c) >> Avg_j\{p^s(w_1|c_j)\}$. We assume the same is true for $p^s(w_2|c)$. In this case, for d, it is natural to prefer c over c' unless $p^s(w_3|c')$ is discriminative evidence for class c'. Unfortunately, standard NBC does not do this: when $p(w_3|c)$ is very small, class c will not be able to compete against c' if $p^s(w_1|c'), p^s(w_2|c'), p^s(w_3|c')$ are only moderately large. That is to say, the effects of discriminative evidences are overwhelmed by low probability estimates of other words in NBC.

Contradiction pair problem and discriminative evidence cancelation problem manifest themselves notably in the presence of a large number of classes. It is because: (i)More classes mean more chances of contradiction pairs. Average to every prediction, the number of possible contradiction pairs is linear to the number of class. (ii) there is a large amount of zero-frequency in word-class frequency table. They may get small and unreliable values in smoothing and then easily overwhelm those discriminative evidence. (iii)the values of N_c may vary more greatly, which largely randomize the order of $p^s(w|c_j)$. The above claims can be verified by some careful calculations.

3. MODIFICATIONS

We propose two different modifications in this section. The first is *Weight Manipulation Naive Bayes* (WMNB). We regard estimations $p(w_i|c)$ as the weight instead of probability, which means they are free from the constraint $\sum_i p(w_i|c) = 1$. For word w_i and c_u, we use z^i_u to denote the weight. Our idea is that we could directly use maximum likelihood estimations as the weight for non-zero frequencies of w_i in c_u, and use moderate small weight for zero frequencies. The modified training process of Naive Bayes is presented as:

$$z^i_u = \begin{cases} \log p^{ml}(w_i|c_u) = \log N^i_u - \log N_u & \text{if } N^i_u \neq 0 \\ \gamma / \sum_{j:N^j_u=0} 1 & \text{otherwise} \end{cases} \quad (1)$$

where $\gamma < 0$ is a constant value, $\sum_{j:N^j_u=0} 1$ counts the number of zero-freqency word for class c. To classify $d = \{f_1, f_2 \dots f_V\}$, the modified testing process of Naive Bayes is: $\ell(d) = \arg\max_u\{\log p(c_u) + \sum_{w_i \in d} f_i \cdot z^i_u\}$.

WMNB can alleviate contradiction pair problem: if we have $p^{ml}(w_i|c_u) - p^{ml}(w_i|c_v) > 0$, there are two possibilities: (i) $p^{ml}(w_i|c_u) > p^{ml}(w_i|c_v) > 0$: then we have $z^i_u - z^i_v = \log p^{ml}(w_i|c_u) - \log p^{ml}(w_i|c_v) > 0$ (ii) $p^{ml}(w_i|c_u) > p^{ml}(w_i|c_v) = 0$: then since γ is the constant value to turn, we can always take one small enough γ that meets $z^i_u > z^i_v$. For discriminative evidence cancelation problem, WMNB can alleviate it by controlling the sum of those small weights.

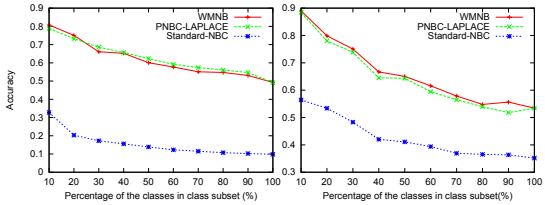
We call the second modification *Parametric Smoothing Naive Bayes*. It tries to modify traditional smoothing methods to alleviate the two problems in this paper. Besides Laplacian method employed by standard Naive Bayes, we can also borrow other smoothing methods from language model. We modify four well-known smoothing methods and lead to four variants: PNBC-Laplace, PNBC-Absolute, PNBC-Linear and PNBC-WittenBell. The basic idea is that NBC should be able to control the extent of reducing the probability of non-zero-frequency words. For example, in PNBC-Laplace, we take $p^s(w_i|c) = \frac{N^i_c + \alpha}{N_c + \sum_w \alpha}$, with a moderate samll α decided by cross validation, the smoothed estimations are close to maximum likelihood estimations and avoid discriminative evidence cancelation problem as well.

Table 1: Performance Comparison on three datasets

	20 NG	ODP	YahooQA
Standard NB	0.8771	0.0946	0.3637
WMNB	0.9151	0.4915	0.5241
PNBC-Laplace	0.9136	0.4925	0.5069
PNBC-Absolute	0.9109	0.4986	0.5023
PNBC-Linear	0.9120	0.4940	0.5031
PNBC-Wittenbell	0.9104	0.4987	0.5023

4. EVALUATION

Table 1 shows the overall performance of WMNB and the variants of PNBC on a 1,048 classes and 185,728 documents Open Directory Project(ODP) dataset; 505 classes and 1,910,741 documents Yahoo Question Answer (YahooQA) dataset; and 20 Newgroups as well. We use standard NBC as the comparisons. It can be seen clearly that our modified Naive Bayes results in remarkable improvements on ODP and YahooQA by alleviating the contradiction pair problem and discriminative evidence cancelation problem. Moreover, the performances of our modified algorithms are similar to each other. In other words, not a specific kind of variants, but their shared underlying principles proposed before count. Figure 3 plots the performance curve for standard NBC, WMNB and PNBC-Laplace on YahooQA and ODP datasets with different number of classes. We can see that WMNB and PNBC-Laplace beat Standard NBC at every points. These evaluations proves our modified Naive Bayes algorithms are effective on web-scale taxonomies.



(a) ODP (b) YahooQA

Figure 3: Standard NB, WMNB and PNBC-Laplace with different number of classes

5. REFERENCES

[1] L. Cai and T. Hofmann. Hierarchical document categorization with support vector machines. In CIKM 2004, pages 78-87, 2004.

[2] J. Rennie, L. Shih, J. Teevan, and D. Karger. Tackling the poor assumptions of naive bayes text classifiers. In ICML 2003, pages 285-295, 2003.

[3] G.-R. Xue, D. Xing, Q. Yang, and Y. Yu. Deep classification in large-scale text hierarchies. In SIGIR 2008, pages 100-107, 2008.