Efficient Training on Biased Minimax Probability Machine for Imbalanced Text Classification

Xiang Peng
Department of Computer Science & Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
xpeng@cse.cuhk.edu.hk

Irwin King
Department of Computer Science & Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
king@cse.cuhk.edu.hk

ABSTRACT

The Biased Minimax Probability Machine (BMPM) constructs a classifier which deals with the imbalanced learning tasks. In this paper, we propose a Second Order Cone Programming (SOCP) based algorithm to train the model. We outline the theoretical derivatives of the biased classification model, and address the text classification tasks where negative training documents significantly outnumber the positive ones using the proposed strategy. We evaluated the learning scheme in comparison with traditional solutions on three different datasets. Empirical results have shown that our method is more effective and robust to handle imbalanced text classification problems.

Categories and Subject Descriptors

1.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier Design and Evaluation*; H.2.8 [**Database Management**]: Database Application—*Data Mining*

General Terms

Algorithm, Management, Experimentation

Keywords

Biased Classification, Second Order Cone Programming, Biased Minimax Probability Machine, Text Classification

1. INTRODUCTION

With the rapid growth of text information on the World Wide Web (WWW), text classification has become one of the most important topic in both the community of research and engineering [1]. However there are two major problems with current algorithms involving in text classification task. One key challenge is that almost all the algorithms treat the problem as a balanced classification task and they do not consider the imbalanced dataset matter, which means the number of negative documents is larger than the number of positive ones. Take the task of learning which news articles are of interest to a particular person reading Google News for example. The articles which the person interested may be just a small portion in the whole text database. Methods that filter and present only the ones that user finds interesting are highly desirable. Currently, most researchers treat this problem as a strict binary classification problem while ignore the fact that the number of uninterested documents is extremely larger than the interested ones. How to make the returned document set as accuracy as possible is a crucial problem. At the same time, in order to build a reliable classifier for text classification, we need to train the model with huge number of predefined documents, which is usually a very time consuming process. Thus, how to reduce the time required for training a reliable text classifier is a crucial obstacle for large scale text classification. This is particularly challenging for text classification of WWW documents given its nature of large volume.

In this paper, we apply the model of Biased Minimax Probability Machine (BMPM) to the problem of imbalanced text classification, and propose a new training algorithm to tackle the complexity and accuracy issues in BMPM learning task. This model is transformed into a Second Order Cone Programming (SOCP) problem. Under this new proposed framework, the imbalanced text classification problem could be modelled and solved efficiently.

The rest of this paper is organized as follows. Section II introduces the concept of BMPM. Section III presents an effective learning algorithm based on SOCP for effective training with BMPM. Section IV presents the results of our empirical study.

2. BIASED MINIMAX PROBABILITY MACHINE

We assume two random vectors $\mathbf{x}$ and $\mathbf{y}$ represent two classes of data with mean and covariance matrices as $\{\overline{\mathbf{x}}, \Sigma_{\mathbf{x}}\}$ and $\{\overline{\mathbf{y}}, \Sigma_{\mathbf{y}}\}$, respectively in a two-category classification task, where $\mathbf{x}$, $\mathbf{y}$, $\overline{\mathbf{x}}$, $\overline{\mathbf{y}} \in \mathbb{R}^n$, and $\Sigma_{\mathbf{x}}$, $\Sigma_{\mathbf{y}} \in \mathbb{R}^{n\times n}$. Biased Minimax Probability Machine (BMPM) attempts to determine the hyperplane $\mathbf{a}^T\mathbf{z} = b$ with $\mathbf{a}^T\mathbf{z} > b$ being considered as class $\mathbf{x}$ and $\mathbf{a}^T\mathbf{z} < b$ being judged as class $\mathbf{y}$ to separate the important class of data $\mathbf{x}$ with a maximal probability while keeping the accuracy of less important class of data $\mathbf{y}$ acceptable.[1] We formulate this objective as follows:

$$\begin{array}{ll} \max\limits_{\alpha,\beta,b,\mathbf{a}\neq\mathbf{0}} & \alpha \\ s.t. & \inf\limits_{\mathbf{x}\sim(\overline{\mathbf{x}},\Sigma_{\mathbf{x}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x} \geq b\} \geq \alpha, \\ & \inf\limits_{\mathbf{y}\sim(\overline{\mathbf{y}},\Sigma_{\mathbf{y}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \beta, \\ & \beta \geq \beta_0, \end{array} \tag{1}$$

where α and β represent the lower bounds of the accuracy for

[1]The reader may refer to [2] for a more detailed and complete description.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

	$BMPM_{SOCP}$	$BMPM_{FP}$	MPM	SVM	kNN
α	**81.42 ± 0.22** ↑	80.35 ± 0.13 ↑	76.30 ± 0.28	-	-
β	**70.00 ± 0.00**	70.00 ± 0.00	76.30 ± 0.34	-	-
TSA_x	**83.10 ± 0.60** ↑	81.07 ± 0.63 ↑	74.91 ± 0.61	73.23 ± 1.59	71.60 ± 0.38
TSA_y	**72.61 ± 0.84**	74.48 ± 0.69	75.20 ± 0.62	74.60 ± 0.47	69.40 ± 0.60
TSA	**77.85 ± 0.04**	77.70 ± 0.21	75.05 ± 0.37	73.90 ± 0.44	70.50 ± 0.55

Table 1: Lower Bound α and Test-Set Accuracy on the Reuter-21578 dataset (%)

future data classification, namely, the worst-case accuracy. Meanwhile, β_0 is a pre-specified positive constant which represents an acceptable accuracy for the less important class.

3. EFFICIENT BMPM TRAINING

3.1 Motivation

Most of recent studies on BMPM are usually based on the Fractional Programming problem (we name it $BMPM_{FP}$) which could be solved by Rosen Gradient method. However the problem reformulation has some crucial assumption when doing the transformation which would lead to failure of the model solution. Another issue is that when applying the Fractional Programming based $BMPM_{FP}$ into large real-world classification problems, it would be very sensitive to data dimension and very time consuming.

3.2 Proposed Strategy

Our main result is stated below.

THEOREM 1. *If $\overline{\mathbf{x}} = \overline{\mathbf{y}}$, then the minimax probability decision problem (1) does not have a meaningful solution: the optimal worst-case misclassification probability that we obtain is 1 - α^* =1. Otherwise, an optimal hyperplane $H(\mathbf{a}^*, b^*)$ exists, and can be determined by solving the convex optimization problem:*

$$\begin{aligned} \min_{t,\mathbf{a}\neq 0} \quad & t - \mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}) \\ s.t. \quad & \| \Sigma_{\mathbf{x}}^{\frac{1}{2}} \mathbf{a} \| \leq 1, \\ & \| \Sigma_{\mathbf{y}}^{\frac{1}{2}} \mathbf{a} \| \leq \sqrt{\tfrac{1-\beta_0}{\beta_0}} t, \end{aligned} \tag{2}$$

and setting b to the value

$$b^* = \mathbf{a}^{*T}\overline{\mathbf{y}} + \sqrt{\frac{\beta_0}{1-\beta_0}}\sqrt{\mathbf{a}^{*T}\Sigma_{\mathbf{y}}\mathbf{a}^*} = \mathbf{a}^{*T}\overline{\mathbf{x}} - \sqrt{\frac{\alpha^*}{1-\alpha^*}}\sqrt{\mathbf{a}^{*T}\Sigma_{\mathbf{x}}\mathbf{a}^*},$$

where $\mathbf{a}^*$ is an optimal solution of (1), and $t \in \mathbb{R}$ is a new optimization variable. Furthermore, if either $\Sigma_{\mathbf{x}}$ or $\Sigma_{\mathbf{y}}$ is positive definite, the optimal hyperplane is unique.

Lemma 1. The Second Order Cone Programming problem with linear objective function and norm constraints is a convex optimization problem and thus can be solved efficiently.

We omit the details of the proofs due to space limitations.

4. EXPERIMENTAL RESULTS

We evaluated our proposed biased learning algorithm in comparison to the state-of-the-art approaches by conducting empirical comparisons on three standard datasets for text document classification: Reuters-21578 dataset, 20-Newsgroup data collection and Enron Corpus. For all three datasets, the same data pre-processing and feature selection procedures are applied. Due to space limitations, we only present our results on Reuters-21578 dataset.

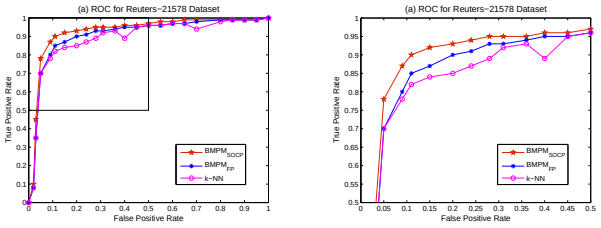

Figure 1: ROC curves on Reuters-21578 dataset: Full Range (Left), Crucial Part (Right)

Applying BMPM-based technique in text classification is a very straightforward task, where we just need to assume the interested documents to be the more important class ($\mathbf{x}$) in the biased classification framework while assume the uninterested ones to be the less important class ($\mathbf{y}$). For experimental setting up, we employ Receiver Operating Characteristic (ROC) analysis and Test Set Accuracy (TSA) as the performance measurements. The involved traditional algorithms are the Support Vector Machine (SVM), k-Nearest Neighbor (kNN) and Minimax Probability Machine (MPM).

Table 1 shows the experimental results of TSA performance evaluation, where we can see that our two BMPM models achieve better performances than the other algorithms in most of the cases while the $BMPM_{SOCP}$ generally outperforms the $BMPM_{FP}$ method.

Furthermore, It is observed from the ROC curves in Figure 1 that most parts of the ROC curve of BMPMs are above the corresponding curve of kNN along with the $BMPM_{SOCP}$ curve is above the one of $BMPM_{FP}$, which demonstrate the superiority of the BMPM models and our proposed $BMPM_{SOCP}$ algorithm.

5. ACKNOWLEDGMENTS

The work described in this paper is supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4235/04E) and is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

6. REFERENCES

[1] S. C. Hoi, R. Jin, and M. Lyu. Large-scale text categorization by batch mode active learning. In *Proc. of World Wide Web Conference*, pages 633–642, 2006.

[2] K. Huang, H. Yang, I. King, M. Lyu, and L. Chan. Minimum error minimax probability machines. *Journal of Machine Learning Research*, 5:1253–1286, 2004.