Learning Classi£ers from Imbalanced Data Based on Biased Minimax Probability Machine

Kaizhu Huang, Haiqin Yang, Irwin King, Michael R. Lyu
Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
{kzhuang, hqyang, king, lyu}@cse.cuhk.edu.hk

Abstract

We consider the problem of the binary classi£cation on imbalanced data, in which nearly all the instances are labelled as one class, while far fewer instances are labelled as the other class, usually the more important class. Traditional machine learning methods seeking an accurate performance over a full range of instances are not suitable to deal with this problem, since they tend to classify all the data into the majority, usually the less important class. Moreover, some current methods have tried to utilize some intermediate factors, e.g., the distribution of the training set, the decision thresholds or the cost matrices, to in¤uence the bias of the classi£cation. However, it remains uncertain whether these methods can improve the performance in a systematic way. In this paper, we propose a novel model named Biased Minimax Probability Machine. Different from previous methods, this model directly controls the worst-case real accuracy of classi£cation of the future data to build up biased classi£ers. Hence, it provides a rigorous treatment on imbalanced data. The experimental results on the novel model comparing with those of three competitive methods, i.e., the Naive Bayesian classi£er, the k-Nearest Neighbor method, and the decision tree method C4.5, demonstrate the superiority of our novel model.

1 Introduction

Learning classi£ers from imbalanced or skewed datasets is an important topic, arising very often in practice in classi£cation problems. In such problems, almost all the instances are labelled as one class, while far fewer instances are labelled as the other class, usually the more important class. It is obvious that traditional classi£ers seeking an accurate performance over a full range of instances are not suitable to deal with imbalanced learning tasks, since they tend to classify all the data into the majority class, which is usually the less important class.

To cope with imbalanced datasets, there are types of methods, such as the methods of sampling [7], the methods of moving the decision thresholds [9][10], and the methods of adjusting the cost-matrices[9]. The £rstschool of methods aims to reduce the data imbalance by "down-sampling" (removing) instances from the majority class or "up-sampling" (duplicating) the training instances from the minority class or both. The second school of methods tries to adapt the decision thresholds to impose bias on the minority class. Similarly, the third school of methods improves the prediction performance by adjusting the weight (cost) for each class.

A common problem for all the three families of methods is that they lack a rigorous and systematic treatment on imbalanced data. To adapt to the imbalanced learning, these methods adjust some intermediate factors, e.g., the prior probabilities (in the sampling methods), decision thresholds, and cost matrices, which are not directly related to the classi£cation accuracy and sometimes may cause problems. For example, down-sampling the data will lose information, while up-sampling will introduce noise. According to [10], one open question is that whether simply varying the skewness of the data distribution can improve predictive performance systematically. Furthermore, Breiman et al. [3] establishes the connection among the distribution of the training data, the prior probability of each class, the costs of misclassi£cation of each class, and the setup of the decision threshold. Changing one of these factors is equivalent to changing other factors. Thus, simply changing thresholds or adjust the weight for each class lacks the systematic foundation in the same sense as the sampling method.

In this paper, based on extending Mimimax Probability Machine (MPM) [8], a competitive model compared with a state-of-the-art classi£er, the Support Vector Machine,we propose a novel model named Biased Minimax Probability Machine (BMPM) to handle the tasks of learning from imbalanced data. Different from the sampling methods, BMPM does not remove or duplicate data. When compared with the methods of changing the thresholds or weights,

it constructs the classi£cation hyperplane by directly controlling the lower bound of the real accuracy of the future data. This distinguishes BMPM from the currently proposed methods and demonstrates its rigorous and systematic treatment on imbalanced data.

This paper is organized as follows. In the next section, we introduce the theory foundation of this paper.We then in Section 3 apply the BMPM to deal with the imbalanced learning tasks. Following that, we evaluate the BMPM model on a series of experiments in Section 4. Finally, we conclude this paper and present future work in Section 5.

2 Biased Minimax Probability Machine

In this section, we £rst introduce the model de£nition of BMPM. Next, we prove the solvability of BMPM. Following that, we propose an ef£cient algorithm to solve the corresponding optimization problem. We then in Section 2.4, discuss the kenerlization of the BMPM model to attack nonlinear classi£cation tasks.

2.1 Model De£nition

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 classi£cation task, where $\mathbf{x}$, $\mathbf{y}$, $\overline{\mathbf{x}}$, $\overline{\mathbf{y}} \in \mathbb{R}^n$, and $\Sigma_{\mathbf{x}}$, $\Sigma_{\mathbf{y}} \in \mathbb{R}^{n\times n}$. For convenience, we also use $\mathbf{x}$ and $\mathbf{y}$ to represent the corresponding class of the $\mathbf{x}$ data and the $\mathbf{y}$ data respectively.

With given reliable $\{\overline{\mathbf{x}}, \Sigma_{\mathbf{x}}\}$, $\{\overline{\mathbf{y}}, \Sigma_{\mathbf{y}}\}$ for two classes of data, we try to £nd a hyperplane $\mathbf{a}^T\mathbf{z} = b$ ($\mathbf{a} \neq \mathbf{0}$, $\mathbf{z} \in \mathbb{R}^n$, $b \in \mathbb{R}$, superscript T denotes the transpose) 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. We formulate this objective as follows:

$$\max_{\alpha,\beta,b,\mathbf{a}\neq\mathbf{0}} \alpha \quad \text{s.t.} \quad \inf_{\mathbf{x}\in\{\overline{\mathbf{x}},\Sigma_{\mathbf{x}}\}} \Pr\{\mathbf{a}^T\mathbf{x} \geq b\} \geq \alpha \,, \quad (1)$$

$$\inf_{\mathbf{y}\in\{\overline{\mathbf{y}},\Sigma_{\mathbf{y}}\}} \Pr\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \beta \,, \quad (2)$$

$$\beta \geq \beta_0 \,, \quad (3)$$

where α represents the lower bound of the accuracy for the classi£cation, or the worst-case accuracy of future data points $\mathbf{x}$; likewise β. The parameter β_0 is a pre-speci£ed positive constant, which represents an acceptable accuracy level for the less important class $\mathbf{y}$.

The above formulation is derived from the MPM [8], which requires the probabilities of correct classi£cation for both classes to be an equal value α. Through this formulation, the BMPM model can handle the biased classi£cation in a direct way.First, this model provides a different treatment on different classes, i.e, the hyperplane $\mathbf{a}_*^T\mathbf{z} = b_*$ given by the solution of this optimization will favor the classi£cation of the important class $\mathbf{x}$ over the less important class $\mathbf{y}$. Second, given the reliable mean and covariance matrices, the derived decision hyperplane is directly associated with two real accuracy indicators of classi£cation of the future data, i.e., α and β, for each class. Third, this model inherits the distribution-free feature of MPM. With no assumption on data, the derived hyperplane seems to be more general and valid than generative classi£ers.Fourth, as shown shortly in this paper, either we can simply modify this BMPM optimization to automatically search the best β_0 in terms of some criteria popular in the machine learning literature, or slightly different from the current setting, we can quantitatively generate the trade-off curve between the accuracies on different classes and leave the task of choosing the best β_0 to the practitioners.

2.2 Solvability

First, by applying Lemma 1 in [8], we can obtain the following transformed optimization problem:

$$\max_{\alpha,\beta,b,\mathbf{a}\neq\mathbf{0}} \quad \alpha \quad \text{s.t.} \quad (4)$$

$$-b + \mathbf{a}^T\overline{\mathbf{x}} \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} \,, \quad (5)$$

$$b - \mathbf{a}^T\overline{\mathbf{y}} \geq \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \,, \quad (6)$$

$$\beta \geq \beta_0 \,, \quad (7)$$

where $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$, $\kappa(\beta) = \sqrt{\frac{\beta}{1-\beta}}$. Constraint (6) is the direct result of the Lemma.Similarly, by changing $\mathbf{a}^T\mathbf{x} \geq b$ to $\mathbf{a}^T(-\mathbf{x}) \leq -b$, (5) can be obtained from (1). From (5) and (6), we get

$$\mathbf{a}^T\overline{\mathbf{y}} + \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \leq b \leq \mathbf{a}^T\overline{\mathbf{x}} - \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} \,. \quad (8)$$

If we eliminate b from this inequality, we obtain

$$\mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}) \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} + \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \,. \quad (9)$$

Since the magnitude of $\mathbf{a}$ does not in¤uence the solution of (9), without loss of generality, we set $\mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}) = 1$. In addition, $\kappa(\alpha)$ increases monotonically with α, maximizing α is equivalent to maximizing $\kappa(\alpha)$. Thus, the problem can be further modi£ed to

$$\max_{\alpha,\beta,\mathbf{a}\neq\mathbf{0}} \quad \kappa(\alpha) \quad \text{s.t.} \quad (10)$$

$$1 \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} + \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \,, \quad (11)$$

$$\mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}) = 1 \,, \quad (12)$$

$$\kappa(\beta) \geq \kappa(\beta_0) \,, \quad (13)$$

where (13) is equivalent to (7) due to the monotonic property of the function κ.

Lemma 1 *The maximum value of $\kappa(\alpha)$ under the constraints of (11), (12), and (13) is achieved when the right hand side of (11) is strictly equal to* 1.

Proof: Assume the maximum is achieved when $1 > \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} + \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}$. A new solution constructed by increasing $\kappa(\alpha)$ with a small positive amount and maintaining $\kappa(\beta)$ and $\mathbf{a}$ unchanged will satisfy the constraints and will be a better solution. ■

Since $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{y}}$ can be considered as positive de£nite matrices,[1] we obtain $\kappa(\alpha) = \frac{1-\kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}}$. It is a linear function with respect to $\kappa(\beta)$. Since $\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}$ is a positive term, this optimization function is maximized when $\kappa(\beta)$ is set to its lower bound $\kappa(\beta_0)$. The BMPM optimization problem is changed to:

$$\max_{\mathbf{a}\neq\mathbf{0}} \frac{1-\kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}} \quad \text{s.t.} \quad \mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}}) = 1\ . \quad (14)$$

Further the above formulation (14) can be written as the so-called Fractional Programming (FP) problem [11],

$$\max_{\mathbf{a}\neq\mathbf{0}} \frac{f(\mathbf{a})}{g(\mathbf{a})}\ , \quad \text{s.t.} \quad \mathbf{a}\in A = \{\mathbf{a}|\mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}}) = 1\}\ , \quad (15)$$

where $f(\mathbf{a}) = 1-\kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}$, $g(\mathbf{a}) = \sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}$. In the following, we propose Lemma 2 to show that this FP problem is solvable.

Lemma 2 *The Fractional Programming problem (15) is a strictly quasiconcave problem and is thus solvable.*

Proof: It is easy to see that the domain A is a convex set on $\mathbb{R}^n$, $f(\mathbf{a})$ and $g(\mathbf{a})$ are differentiable on A. Moreover, since $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{y}}$ can be both considered as positive de£nite matrices, $f(\mathbf{a})$ is a concave function on A and $g(\mathbf{a})$ is a convex function on A. Then $\frac{f(\mathbf{a})}{g(\mathbf{a})}$ is a concave-convex FP or a pseudoconcave problem. Hence it is strictly quasiconcave on A according to [11]. Therefore, every local maximum is a global maximum [11]. In other words, this Fractional Programming problem is solvable. ■

2.3 Practical Solving Method

Many methods can be used to solve the FP problem. For example, a conjugate gradient method can solve this problem in n (the dimension of the data points) steps if the initial point is suitably assigned [1]. In each step, the computational cost to calculate the conjugate gradient is $O(n^2)$. Thus this method has a worst-case $O(n^3)$ time complexity. Adding the time cost to estimate $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$, $\Sigma_{\mathbf{x}}$, $\Sigma_{\mathbf{y}}$, the total cost is $O(n^3+Nn^2)$, where N is the number of the data points. This computational cost is the same order to the Minimax Probability Machine [8] and the linear Support Vector Machine.

In this paper, we use the Rosen Gradient projection method [1] to solve the concave-convex FP problem, which is proven to converge to a local maximum with a worse-case linear convergence rate [1]. More importantly, the local maximum will be exactly the global maximum in this problem.

From Lemma 1, we can see that the inequalities in (8) will become equalities at the maximum point. The optimal b will thus be obtained by

$$b_* = \mathbf{a}_*^T\overline{\mathbf{x}} - \kappa(\alpha^*)\sqrt{\mathbf{a}_*^T\Sigma_{\mathbf{x}}\mathbf{a}_*} = \mathbf{a}_*^T\overline{\mathbf{y}} + \kappa(\beta_0)\sqrt{\mathbf{a}_*^T\Sigma_{\mathbf{y}}\mathbf{a}_*}$$

where $\mathbf{a}_*$ and α^* are obtained by solving the FP problem.

2.4 Extension into Nonlinear Classi£cations

As the classi£er derived from above BMPM is given in a linear con£guration, to handle more general cases, namely, the nonlinear classi£cation tasks, we need to develop methods to extend the linear BMPM. Fortunately, as shown in [8], the kenerlization trick can be used to map the n-dimensional data points into a high-dimensional feature space $\mathbb{R}^f$, where a linear classi£er corresponds to a nonlinear hyperplane in the original space. It is easy to verify the kernelization procedure similar to [8] can be applied to BMPM as well. To save space, we omit the kernelization in this paper and refer the interested readers to [8, 5].

3 BMPM for Imbalanced Learning

In this section, we £rst review four standard imbalanced learning criteria, which are widely used in previous literatures. We then, based on two of them, apply BMPM to the imbalanced learning tasks.

3.1 Four Criteria

In general, four criteria are used to evaluate the performance of classi£ers in learning from imbalanced data. They are (1) Minimum Cost criterion (MC), (2) the criterion of Maximum Geometry Mean (MGM) of the accuracy on the majority class and the minority class, (3) the criterion of the Maximum Sum (MS) of the accuracy on the majority class and the minority class, and (4) the criterion of Receiver Operating Characteristic (ROC) analysis.

[1] In practice, we can always add a small positive amount to the diagonal elements of these two matrices and make them positive de£nite.

The MC criterion [2] minimizes the cost measured by $Cost = F_p \cdot C_{F_p} + F_n \cdot C_{F_n}$, where F_p is the number of the false positive, C_{F_p} is the cost of a false positive, F_n is the number of false negative, and C_{F_n} is the cost of a false negative. However, the cost of misclassi£cation is generally unknown in real cases, this restricts the usage of this measure. The criterion of MGM maximizes the geometric mean of the accuracy [6], but it contains a nonlinear form, which is not easy to be automatically optimized. Comparatively, MS maximizing the sum of the accuracy on the positive class and the negative class (or maximizing the difference between the true-positive and false-positive probability) [4], is a linear form. The ROC analysis originated in signal detection theory has been introduced to evaluate the performance in learning from imbalanced data [12] [9]. This criterion plots a so-called ROC curve to visualize the tradeoff between the false-positive rate and the true-positive rate and leaves the task of the selection of a speci£c tradeoff to the real practitioners. It has been suggested that the area beneath an ROC curve can be used as a measure of accuracy in many applications [12]. Thus, a good classi£er in learning from imbalanced data should have a larger area under the ROC curve.

Based on the above review, in this paper, we will focus on using the criterion of MS and the ROC curve analysis to evaluate the imbalanced learning.

3.2 BMPM for MS

When using BMPM for the criterion of MS, we can modify the formulation of BMPM as follows:

$$\max_{\alpha,\beta,b,\mathbf{a}\neq\mathbf{0}} \quad \alpha+\beta \quad \text{s.t.} \tag{16}$$

$$\inf_{\mathbf{x}\in\{\overline{\mathbf{x}},\Sigma_{\mathbf{x}}\}} \Pr\{\mathbf{a}^T\mathbf{x}\geq b\}\geq\alpha\ , \tag{17}$$

$$\inf_{\mathbf{y}\in\{\overline{\mathbf{y}},\Sigma_{\mathbf{y}}\}} \Pr\{\mathbf{a}^T\mathbf{y}\leq b\}\geq\beta\ , \tag{18}$$

The above formulation directly maximizes the sum of the lower bounds of the accuracies so as to maximize the sum of the accuracies. In comparison, to achieve the maximum sum of the accuracies, other approaches, e.g., the methods of sampling or the methods of adapting the weights often have to search the best sampling proportion or the best weights by trials, which are in general very time-consuming.

It is interesting that a similar modi£cation can be made when the cost for each class is known. Maximizing a weighted worst-case accuracy, i.e., $C_{\mathbf{x}}\alpha+C_{\mathbf{y}}\beta$ instead, will be easily derived in this case, where $C_{\mathbf{x}}, C_{\mathbf{y}}$ are the costs for $\mathbf{x}$ and $\mathbf{y}$ respectively.

Similar to BMPM and applying Lemma 1, we can transform (16) as follows:

$$\max_{\alpha,\beta,\mathbf{a}\neq\mathbf{0}} \quad \alpha+\beta \quad \text{s.t.} \tag{19}$$

$$1=\kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}+\kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}\ , \tag{20}$$

$$\mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}})=1. \tag{21}$$

It can be further transformed as:

$$\max_{\beta,\mathbf{a}\neq\mathbf{0}} \quad \frac{\kappa^2(\alpha)}{\kappa^2(\alpha)+1}+\beta \quad \text{s.t.} \tag{22}$$

$$\mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}})=1, \tag{23}$$

where $\kappa(\alpha)=\frac{1-\kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}}$.

The optimization of (22) corresponds to £nding an optimal β^*, making $f(\beta^*)=\alpha(\beta^*)+\beta^*$ maximal, where $\alpha(\beta^*)=\frac{\kappa^2(\alpha)}{\kappa^2(\alpha)+1}$. Therefore, if we £x β to a speci£c value, the optimization will be equivalent to maximizing $\alpha(\beta)$ and further equivalent to maximizing $\kappa(\alpha)$, which is exactly the BMPM problem. We then change β and repeat the BMPM optimization procedure until an optimal β^* is found, such that $f(\beta^*)$ is maximized. The above procedure is also the so-called line search problem. Many methods can be used to solve it.In this paper, we use the Quadratic Interpolation (QI) method, which is shown to converge superlinearly [1].

3.3 BMPM for ROC Analysis

It is straightforward to apply the BMPM model to plot the ROC curve, since the lower bounds α and β directly and quantitatively control the accuracies for two classes. We only need to adapt the acceptable level for β, namely β_0, from 0 to 1, to obtain a sequence of trade-off between the accuracy of the positive class and the negative class. This demonstrates one of advantages of BMPM over the other methods by adapting the weights or thresholds.

4 Experimental Results

In this section, we evaluate the performance of BMPM, in both the linear (BMPML) and Gaussian (BMPMG) kernel setting, on two real-world imbalanced datasets, namely the Recidivism dataset and the Rooftop dataset in comparison with three competitive classi£ers: the Naive Bayesian (NB) classi£er,the k-Nearest Neighbor (k-NN) method, and the decision tree classi£er C4.5.To adapt to the imbalanced learning, these three methods are modi£ed by changing either the priority distribution or the cost matrices according to the methods introduced in [9]. For the k-NN methods, k is set to the odd number from 1 to 21, but only the best three results are presented for brevity. The width parameter

for the Gaussian kernel is tuned via cross validation methods.

The Recidivism dataset was obtained from a cohort of releasees of the North Carolina prison system in a time period from July 1, 1977 to June 30, 1978. There are totally $4,618$ individuals in this dataset, including a training set with $1,540$ individuals and a test set with $3,078$ individuals. In the training set, 570 (27.5%) individuals were recidivists and 970 (72.5%) were not. In the test set, $1,151$ individuals were recidivists and $1,927$ were not. Although this dataset is not skewed as severely as other reported dataset such as the Rooftop dataset used in the following, it is enough to use this dataset to evaluate the performance of the imbalanced learning.

We £rst present the experimental results based on the MS criterion in the 2-4 columns of Table 1. It is clearly observed that BMPML and BMPMG outperform other methods. Next we present the experimental results based on the ROC analysis. By setting the thresholds or costs with trials for the NB, the k-NN, and C4.5, the ROC curves are generated with good shapes as evenly distributed along their length as possible. As discussed in [9], although this generation method may increase the running time for some methods, e.g., the k-NN, it works well in C4.5 and the NB and is suf£cient to evaluate the performance of imbalanced learning. For the BMPM model, since the lower bounds β_0 serves as the accuracy indicators, we simply vary it from 0 to 1 to generate the corresponding ROC curve. The ROC curves are shown in Figure 1(a). As seen in this £gure, the performances of BMPML and BMPMG are again superior to those of other methods. In addition, in real applications, not all the portions of the ROC curve are of great interest. Usually, those with a small false positive rate and a high true positive rate should be more of interest and importance. We thus especially show the portion of the ROC curve in the range when the false positive rate FP $\in [0,\ 0.5]$ and the true positive rate TP $\in [0.5,\ 1]$. As shown in Figure 1(b), in this critical portion of the ROC curve, the superiority of the BMPL and BMPMG is more obvious than the whole ROC curve analysis. This again demonstrates our model's advantages over other methods. To quantitatively demonstrate the difference, we show the areas beneath the ROC curves approximated by using the trapezoid rule in the 6 column of Table 1. The BMPML and BMPMG show a consistent superiority to the NB, the best of the other three methods.

The Rooftop datasetconsists of $17,048$ overhead images, in which 781 images are labelled as positive examples while $17,048$ images are labelled as negative examples [9]. It is clearly observed that this is a severely skewed dataset.

We randomly split the rooftop data into a training set with 60% data and a test set with 40% data. We construct the classi£ers ten times with the same hold-out proportion and use the average of the results as the performance metric. The results are summarized in 7-12 columns of Table 1 and Figure 1(c). As is clearly observed, for both criteria, the BMPM methods demonstrate its superiority to other methods, since they have higher sums of the accuracies and larger areas under the ROC curves. Similar to what we do in Recivisim dataset, we also plot the more critical proportion of the ROC curve in Figure 1(d). The predominance of the BMPML and the BMPMG are clearly observed. To evaluate the performance more reliably, we perform a signi£cance test based on both LabMRMC [9] and a T-test. The analysis shows that the accuracies of BMPML and BMPMG are signi£cantly different from those of other methods at $p \leq 0.05$, both in terms of the MS criterion and the ROC curve criterion. Note that in the above, BMPM already includes MPM in the case of $\alpha = \beta$. Since the ROC curve plots all the results when β is changed from 0 to 1, the result of MPM is thus implicitly contained in our experiments.

5 Conclusion and Future Work

In this paper, we propose a novel model named Biased Minimax Probability Machine to deal with the task of learning from imbalanced datasets. Given the reliable estimation of the mean and covariance of data, this model constructs the classi£cation boundary by directly controlling the lower bound of the real accuracy and thus provides a systematical and rigorous treatment on skewed data. We prove the solvability, and propose ef£cient algorithms to solve the optimization problem of BMPM. Moreover, we evaluate our novel model on two real world datasets in terms of two criteria. In both criteria, the performance is shown to be the best when compared with other competitive methods such as the Naive Bayesian classi£er, the k-Nearest Neighbor method, and the decision tree classi£er, C4.5.

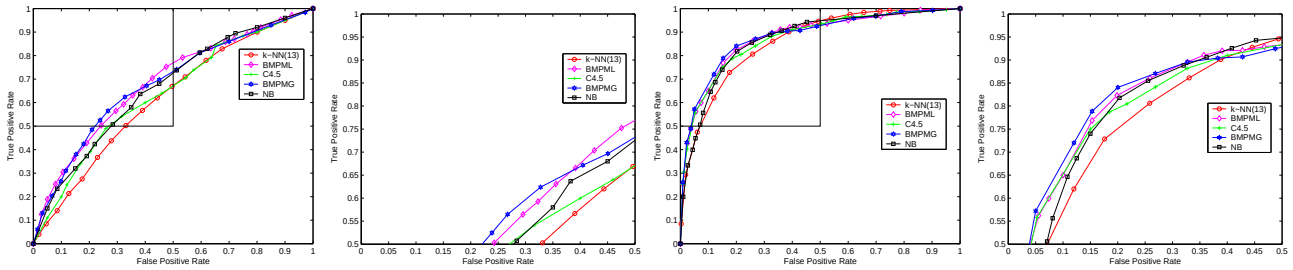
Some important issues need to be checked as our future work. Firstly, how to estimate the means and covariances accurately and robustly? Secondly, are there other more ef£cient methods to solve the Fractional Programming optimization problem? Can some decomposable techniques be applied in the Gram matrix and thus speed up the least-squares training? Finally, how to extend the scheme to the multi-category tasks is also one of our research topics in the near future.

Acknowledgment

The work described in this paper was fully supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4182/03E and Project No. CUHK4351/02).

Table 1: Evaluation on Recidivism and Rooftop using the MS criterion and ROC analysis

Dataset	Recidivism					Rooftop					
Criterion	MS (%)			ROC (%)		MS (%)				ROC (%)	
Method	TN	TP	(TN+TP) / 2	Method	Area	Method	TN	TP	(TN+TP) / 2	Method	Area
NB	61.8	63.8	62.7	NB	66.5	NB	79.7 ± 0.4	81.8 ± 0.8	80.7 ± 0.7	NB	86.8 ± 0.6
k-NN(9)	62.6	54.6	58.6	k-NN(11)	61.6	kNN(7)	75.1 ± 0.6	80.7 ± 0.6	77.9 ± 0.5	kNN(9)	86.0 ± 0.9
k-NN(11)	62.4	55.4	58.9	k-NN(13)	61.9	kNN(13)	74.1 ± 0.5	81.4 ± 0.8	77.7 ± 0.6	kNN(11)	85.7 ± 0.6
k-NN(13)	55.7	62.0	58.9	k-NN(17)	61.5	kNN(15)	74.3 ± 0.7	82.1 ± 0.7	78.2 ± 0.7	kNN(15)	85.8 ± 0.6
C4.5	74.1	49.0	61.5	C4.5	63.8	C4.5	81.8 ± 0.4	79.4 ± 0.6	80.6 ± 0.5	C4.5	87.4 ± 0.6
BMPML	70.4	57.5	**63.9**	BMPML	**68.4**	BMPML	80.2 ± 0.6	82.3 ± 0.6	$\mathbf{81.2 \pm 0.6}$	BMPML	$\mathbf{87.9 \pm 0.6}$
BMPMG	72.0	57.7	**64.9**	BMPMG	**68.0**	BMPMG	80.0 ± 0.9	84.1 ± 1.0	$\mathbf{82.0 \pm 0.9}$	BMPMG	$\mathbf{88.2 \pm 0.9}$



(a) (b) (c) (d)

Figure 1: ROC curves for the Recidivism and the Rooftop dataset. Sub£gures (a) and (c) show a full range of the ROC curves for Recidivism and Rooftop respectively, while sub£gures (b) and (d) show a critical portion of the ROC curves for Recidivism and Rooftop respectively, which is more important in real applications. All £gures demonstrate the superiority of the BMPM model.

References

[1] D. P. Bertsekas. *Nonlinear Programming*. Athena Scienti£c, Belmont, Massashusetts, 2nd edition, 1999.

[2] A. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithm. *Pattern Recognition*, 30(7):1145–1159, 1997.

[3] L. Breiman. Arcing classi£ers. Technical Report 460, Statistics Department, University of California, 1997.

[4] J. W. Grzymala-Busse, L. K. Goodwin, and X. Zhang. Increasing sensitivity of preterm birth by changing rule strengths. *Pattern Recognition Letters*, (24):903–910, 2003.

[5] Kaizhu Huang, Haiqin Yang, Irwin King, Michael R. Lyu, and Laiwan Chan. Biased minimax probability machine for medical diagonsis. In *the Eighth International Symposium on Artif icial Intelligence and Mathematics*, 2003.

[6] M. Kubat, R. Holte, and S. Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, (30):195–215, 1998.

[7] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the Fourteen International Conference on Machine Learning (ICML-1997)*, pages 179–186. San Francisco, CA: Morgan Kaufmann, 1997.

[8] G. R. G. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classi£cation. *Journal of Machine Learning Research*, 3:555–582, 2002.

[9] M. A. Maloof, P. Langley, T. O. Binford, R. Nevatia, and S. Sage. Improved rooftop detection in aerial images with machine learning. *Machine Learning*, 53:157–191, 2003.

[10] F. Provost. Learning from imbanlanced data sets. In *Proceedings of The Seventeenth National Conference on Arti£cial Intelligence (AAAI 2000)*, 2000.

[11] S. Schaible. *Fractional programming*. Nonconvex Optimization and its Applications. Kluwer Academic Publishers, Dordrecht-Boston-London, 1995.

[12] J. Swets. Measureing the accuracy of diagnostic systems. *Science*, (240):1285–1293, 1988.