Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 21 (2008) 450–457
Neural Networks
www.elsevier.com/locate/neunet

2008 Special Issue

Robust BMPM training based on second-order cone programming and its application in medical diagnosis☆

Xiang Peng*, Irwin King

Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

Received 3 August 2007; received in revised form 1 December 2007; accepted 28 December 2007

Abstract

The Biased Minimax Probability Machine (BMPM) constructs a classifier which deals with the imbalanced learning tasks. It provides a worst-case bound on the probability of misclassification of future data points based on reliable estimates of means and covariance matrices of the classes from the training data samples, and achieves promising performance. In this paper, we develop a novel yet critical extension training algorithm for BMPM that is based on Second-Order Cone Programming (SOCP). Moreover, we apply the biased classification model to medical diagnosis problems to demonstrate its usefulness. By removing some crucial assumptions in the original solution to this model, we make the new method more accurate and robust. We outline the theoretical derivatives of the biased classification model, and reformulate it into an SOCP problem which could be efficiently solved with global optima guarantee. We evaluate our proposed SOCP-based BMPM ($BMPM_{SOCP}$) scheme in comparison with traditional solutions on medical diagnosis tasks where the objectives are to focus on improving the sensitivity (the accuracy of the more important class, say "ill" samples) instead of the overall accuracy of the classification. Empirical results have shown that our method is more effective and robust to handle imbalanced classification problems than traditional classification approaches, and the original Fractional Programming-based BMPM ($BMPM_{FP}$).

Keywords: Biased minimax probability machine; Second-order cone programming; Medical diagnosis

1. Introduction

Classifiers are widely being used in various disciplines with applications such as Information Retrieval (Peng & King, 2006a, 2006b), Bioinformatics (Huang, Yang, King, & Lyu, 2006b; Huang, Yang, King, Lyu, & Chan, 2004c), Text Categorization (Macskassy, Hirsh, Banerjee, & Dayanik, 2001; Nigam, Mccallum, Thrun, & Mitchell, 1999), etc. In particular, biased classifiers, a special kind of classifiers, seek to make the accuracy of the important class, instead of the overall accuracy, as high as possible, while maintaining the accuracy of the less important class at an acceptable level. Recently, a novel biased classification model, Biased Minimax Probability Machine (BMPM), provides a worst-case bound on the probability of misclassification of future data points based on reliable estimates of means and covariance matrices of the classes from the training data points and achieves promising performance (Huang, Yang, King, & Lyu, 2004a, 2006a).

Applying machine learning techniques to medical diagnosis tasks has the advantage of saving time and reducing cost (Kononenko, 2001; West & West, 2000). Many different techniques have been applied to medical diagnosis in the machine learning literature, including Naive Bayesian method (*NB*) (Langley, Iba, & Thompson, 1992), the *k*-Nearest Neighbor method (*kNN*) (Aha, Kibler, & Albert, 1991), the decision tree (Quinlan, 1993) and the logistic regression (Jordan, 1995). The challenging task of medical diagnosis based on machine learning techniques requires an inherent bias, i.e., the diagnosis should favor the positive identification of the "ill" class over the misidentification of the "healthy" class, since a misdiagnosis of an ill patient as a healthy one may delay the therapy and aggravate the illness. Therefore, the objective in the identification task is not to improve the overall accuracy of the classification, but to focus on improving the *sensitivity* (the accuracy of the

☆ An abbreviated version of some portions of this article appeared in Peng and King (2007) as part of the IJCNN 2007 Conference Proceedings, published under IEE copyright.

* Corresponding author. Tel.: +852 26098431; fax: +852 26035024.
E-mail address: xpeng@cse.cuhk.edu.hk (X. Peng).

0893-6080/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2007.12.051

"ill" class) while maintaining an acceptable *specificity* (the accuracy of the "healthy" class) (Grzymala-Busse, Goodwin, & Zhang, 2003). Some current methods adopt roundabout ways to impose a certain bias toward the important class, i.e., they try to utilize some intermediate factors to influence the classification (Cardie & Nowe, 1997; Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Kubat & Matwin, 1997; Maloof, Langley, Binford, Nevatia, & Sage, 2004). However, it remains uncertain whether these methods can improve the classification performance systematically.

In this paper, by employing the Biased Minimax Probability Machine (BMPM), we deal with the issue in a more elegant way and directly achieve the objective of appropriate medical diagnosis. We extend the original BMPM model of Huang et al. (2006b) 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 instead of a Fractional Programming (FP) one (Peng & King, 2007). Under this new proposed framework, the imbalanced classification problem could be modelled and solved efficiently. Moreover, we apply the model to handle the biomedical problems in this work.

The rest of this paper is organized as follows. Section 2 reviews the concept of Biased Minimax Probability Machine (BMPM) and related work. Section 3 presents a robust learning algorithm based on the Second-Order Cone Programming for BMPM. Section 4 gives out the results of our empirical study on the derived learning scheme. Conclusion and future work are given in Section 5.

2. Biased minimax probability machine

In this section, we present the biased minimax framework, designed to achieve the goal of the imbalanced classification. We first introduce and define the linear Biased Minimax Probability Machine (BMPM) model. We then review optimization solutions that solve the linear version of the BMPM model.

2.1. Model definition

We assume that two random vectors $\mathbf{x}$ and $\mathbf{y}$ represent two classes of data with means and covariance matrices as $\{\bar{\mathbf{x}}, \Sigma_{\mathbf{x}}\}$ and $\{\bar{\mathbf{y}}, \Sigma_{\mathbf{y}}\}$, respectively in a two-category classification task, where $\mathbf{x}, \mathbf{y}, \bar{\mathbf{x}}, \bar{\mathbf{y}} \in \mathbb{R}^n$, and $\Sigma_{\mathbf{x}}, \Sigma_{\mathbf{y}} \in \mathbb{R}^{n\times n}$. For convenience, in the following, we use $\mathbf{x}$ and $\mathbf{y}$ to represent the corresponding class of the $\mathbf{x}$ data and the $\mathbf{y}$ data respectively.[1]

Assuming that $\{\bar{\mathbf{x}}, \Sigma_{\mathbf{x}}\}$, $\{\bar{\mathbf{y}}, \Sigma_{\mathbf{y}}\}$ for two classes of data are reliable, Biased Minimax Probability Machine (BMPM) attempts to determine the hyperplane $\mathbf{a}^T\mathbf{z} = b$ ($\mathbf{a} \neq \mathbf{0}$, $\mathbf{z} \in \mathbb{R}^n$, $b \in \mathbb{R}$) 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:

[1] The reader may refer to Huang, Yang, King, Lyu, and Chan (2004d) for a more detailed and complete description of the BMPM model.

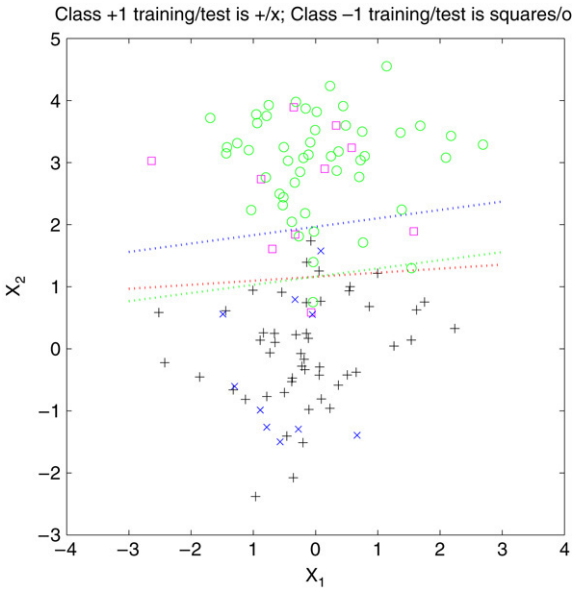

Fig. 1. Decision Lines Comparison: MPM Decision Line (dotted red line), BMPM Decision Line (dotted green line), SVM Decision Line (dotted blue line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$\begin{aligned} \max_{\alpha,\beta,b,\mathbf{a}\neq\mathbf{0}} \quad & \alpha \\ \text{s.t.} \quad & \inf_{\mathbf{x}\sim(\bar{\mathbf{x}},\Sigma_{\mathbf{x}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x} \geq b\} \geq \alpha, \\ & \inf_{\mathbf{y}\sim(\bar{\mathbf{y}},\Sigma_{\mathbf{y}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \beta, \\ & \beta \geq \beta_0, \end{aligned} \tag{1}$$

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

This optimization will maximize the accuracy for the biased class $\mathbf{x}$ (the probability α) while maintaining the class $\mathbf{y}$'s accuracy at an acceptable level by setting a lower bound β_0 as indicated in the third constraint of optimization problem (1). The hyperplane $\mathbf{a}^{*T}\mathbf{z} = b^*$ given by the solution of this optimization will favor the classification of the important class $\mathbf{x}$ over the class $\mathbf{y}$, and will be more suitable in handling biased classification tasks. This is illustrated in Fig. 1.

2.2. Solving the biased minimax probability machine

With the ground work found in Huang et al. (2006b), Huang et al. (2004c) and Lanckriet, Ghaoui, Bhattacharyya, and Jordan (2003), we adopt the transformed optimization problem by using Lemma 1 from Lanckriet et al. (2003) as:

$$\max_{\alpha,\beta,b,\mathbf{a}\neq\mathbf{0}} \quad \alpha \tag{2}$$

$$\text{s.t.} \quad -b + \mathbf{a}^T\bar{\mathbf{x}} \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}, \tag{3}$$

$$b - \mathbf{a}^T\bar{\mathbf{y}} \geq \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}, \tag{4}$$

$$\beta \geq \beta_0, \tag{5}$$

where $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$, $\kappa(\beta) = \sqrt{\frac{\beta}{1-\beta}}$.

From constraints (3) and (4), we eliminate b from this optimization problem. Without considering the influence of magnitude of $\mathbf{a}$ on the optimal solution for the above problem, we set $\mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}) = 1$. In addition, since $\kappa(\alpha)$ increases monotonically with α, maximizing α is equivalent to maximizing $\kappa(\alpha)$. Thus the problem can be finally transformed to the Fractional Programming problem as,

$$\max_{\mathbf{a}\neq\mathbf{0}} \quad \frac{1-\kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}} \tag{6}$$

$$\text{s.t.} \quad \mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}) = 1, \tag{7}$$

$$\kappa(\beta) \geq \kappa(\beta_0), \tag{8}$$

where the objective function is a linear function with respect to $\kappa(\beta)$, and $\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}$ is a positive term.

In the earlier work of this model, Rosen Gradient projection method (Bertsekas, 2004) is employed to find the solution of this concave-convex FP problem. Furthermore it is observed that the inequalities in (3) and (4) will become equalities at the optimal point. The optimal b will thus be obtained by

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

3. Robust BMPM training

3.1. Motivation

Biased Minimax Probability Machine (BMPM) has been extensively studied as a state-of-the-art learning techniques in various areas, such as bioinformatics (Huang et al., 2006b, 2004c), information retrieval (Peng & King, 2006a, 2006b) and statistical learning (Huang, Yang, King, & Lyu, 2004b). Most of the recent studies on BMPM are generally 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 assumptions which may lead to failure of the model solution. Another issue is the Fractional Programming-based $BMPM_{FP}$ would be very sensitive to data dimension and time consuming when applied to some domain-specific applications. As we can see in its original solution, it directly sets $\mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}) = 1$. However this is not necessarily the case in every learning problem.

Motivated from these shortcomings of FP-based BMPM solutions, we formulate the model into a Second-Order Cone Programming (SOCP) problem without any loss of model information. With this, the BMPM could be effectively trained and applied to biased machine learning problems with more accurate results.

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 $1 - \alpha^* = 1$. Otherwise, an optimal hyperplane $H(\mathbf{a}^*, b^*)$ exists, and can be determined by solving the convex optimization problem:*

$$\begin{aligned} \min_{t,\mathbf{a}\neq\mathbf{0}} \quad & t - \mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}) \\ s.t. \quad & \|\Sigma_{\mathbf{x}}^{\frac{1}{2}}\mathbf{a}\| \leq 1, \\ & \|\Sigma_{\mathbf{y}}^{\frac{1}{2}}\mathbf{a}\| \leq \sqrt{\frac{1-\beta_0}{\beta_0}}t, \end{aligned} \tag{9}$$

and setting b to the value

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

where $\mathbf{a}^*$ is the optima of (9), and $t \in \mathbb{R}$ is a new optimization variable. The optimal worst-case misclassification probability for class $\mathbf{x}$ and $\mathbf{y}$ is

$$\mathbf{Pr}(Misclassification_{\mathbf{x}}) = 1 - \alpha^*, \tag{10}$$

$$\mathbf{Pr}(Misclassification_{\mathbf{y}}) = 1 - \beta_0, \tag{11}$$

respectively. Furthermore, if either $\Sigma_{\mathbf{x}}$ or $\Sigma_{\mathbf{y}}$ is positive definite, the optimal hyperplane is unique.

Proof. It is observed that the optimization problem (1) could be transformed to the following format:

$$\begin{aligned} \max_{\alpha,b,\mathbf{a}\neq\mathbf{0}} \quad & \alpha \\ \text{s.t.} \quad & \inf_{\mathbf{x}\sim(\overline{\mathbf{x}},\Sigma_{\mathbf{x}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x} \geq b\} \geq \alpha, \\ & \inf_{\mathbf{y}\sim(\overline{\mathbf{y}},\Sigma_{\mathbf{y}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \beta_0. \end{aligned} \tag{12}$$

By using Lemma 1 in Lanckriet et al. (2003), the above optimization becomes:

$$\begin{aligned} \max_{\alpha,\mathbf{a}\neq\mathbf{0}} \quad & \alpha \\ \text{s.t.} \quad & \sqrt{\frac{\alpha}{1-\alpha}}\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} \\ & + \sqrt{\frac{\beta_0}{1-\beta_0}}\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \leq \mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}). \end{aligned} \tag{13}$$

Since $\sqrt{\frac{\alpha}{1-\alpha}}$ is a monotonic increasing function of α, we can change variables and rewrite our problem as

$$\begin{aligned} \max_{\alpha,\mathbf{a}\neq\mathbf{0}} \quad & \sqrt{\frac{\alpha}{1-\alpha}} \\ \text{s.t.} \quad & \sqrt{\frac{\alpha}{1-\alpha}}\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} \\ & + \sqrt{\frac{\beta_0}{1-\beta_0}}\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \leq \mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}). \end{aligned} \tag{14}$$

Considering $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{y}}$ can be viewed as positive definite matrices, we formulate the optimization as following:

$$\begin{aligned} \max_{\alpha,\mathbf{a}\neq\mathbf{0}} \quad & \sqrt{\frac{\alpha}{1-\alpha}} \\ \text{s.t.} \quad & \sqrt{\frac{\alpha}{1-\alpha}} \leq \frac{\mathbf{a}^T(\overline{\mathbf{x}} - \overline{\mathbf{y}}) - \sqrt{\frac{\beta_0}{1-\beta_0}}\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}}, \end{aligned} \tag{15}$$

which allow us to eliminate $\sqrt{\frac{\alpha}{1-\alpha}}$,

$$\max_{\mathbf{a}\neq\mathbf{0}} \frac{\mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}}) - \sqrt{\frac{\beta_0}{1-\beta_0}}\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}}. \tag{16}$$

It is observed that the optimization problem (16) is equivalent to bound the denominator to 1, and then maximize its numerator. Otherwise if the denominator has no bound, we would have no way to obtain the optimal solution.[2] Furthermore, the maximization of an item is equivalent to minimize its opponent. Hence, we obtain the transformed problem as,

$$\begin{aligned} \min_{\mathbf{a}\neq\mathbf{0}} \quad & -\mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}}) + \sqrt{\frac{\beta_0}{1-\beta_0}}\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \\ \text{s.t.} \quad & \sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} \leq 1. \end{aligned} \tag{17}$$

And it could be further transformed to

$$\begin{aligned} \min_{t,\mathbf{a}\neq\mathbf{0}} \quad & t - \mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}}) \\ \text{s.t.} \quad & \sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} \leq 1, \\ & \sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \leq \sqrt{\frac{1-\beta_0}{\beta_0}}t. \end{aligned} \tag{18}$$

Eq. (18) is exactly a Second-Order Cone Programming problem in the form of:

$$\begin{aligned} \min_{t,\mathbf{a}\neq\mathbf{0}} \quad & t - \mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}}) \\ \text{s.t.} \quad & \|\Sigma_{\mathbf{x}}^{\frac{1}{2}}\mathbf{a}\| \leq 1, \\ & \|\Sigma_{\mathbf{y}}^{\frac{1}{2}}\mathbf{a}\| \leq \sqrt{\frac{1-\beta_0}{\beta_0}}t. \end{aligned} \tag{19}$$

The above problem is convex, feasible, and its objective is linear; therefore, there exists an optimal point, $\mathbf{a}^*$. The linearity of the objective function which is strict convex implies that the optimal point is unique. This ends our proof of the theorem. □

Lemma 1. *The Second-Order Cone Programming problem with linear objective function and norm constraints is a convex optimization problem and thus is solvable.*

Proof. This can be directly observed from the properties of convex optimization. □

A number of software packages can be used to solve this problem. For example, SeDuMi can solve the transformed BMPM model efficiently with the global optima guarantee (Sturm, 1999).

3.3. Kernelized biased minimax probability machine and its solution

We use the kernelization technique to map the n-dimensional data points into a high-dimensional feature space $\mathbb{R}^f$, in which a linear classifier corresponds to a nonlinear hyperplane in the original space.

Assuming that the training data points are represented by $\{\mathbf{x}_i\}_{i=1}^{N_{\mathbf{x}}}$ and $\{\mathbf{y}_j\}_{j=1}^{N_{\mathbf{y}}}$ for class $\mathbf{x}$ and class $\mathbf{y}$, respectively, we can formulate the kernel mapping as:

$$\begin{aligned} &\mathbf{x} \to \varphi(\mathbf{x}) \sim (\overline{\varphi(\mathbf{x})}, \Sigma_{\varphi(\mathbf{x})}), \\ &\mathbf{y} \to \varphi(\mathbf{y}) \sim (\overline{\varphi(\mathbf{y})}, \Sigma_{\varphi(\mathbf{y})}), \end{aligned}$$

where $\varphi : \mathbb{R}^n \to \mathbb{R}^f$ is a mapping function. The corresponding linear classifier in $\mathbb{R}^f$ is $\mathbf{a}^T\varphi(\mathbf{z}) = b$, where $\mathbf{a}$, $\varphi(\mathbf{z}) \in \mathbb{R}^f$ and $b \in \mathbb{R}$. Similarly, the transformed SOCP optimization in BMPM can be written as:

$$\begin{aligned} \min_{t,\mathbf{a}\neq\mathbf{0}} \quad & t - \mathbf{a}^T(\overline{\varphi(\mathbf{x})} - \overline{\varphi(\mathbf{y})}) \\ \text{s.t.} \quad & \|\Sigma_{\varphi(\mathbf{x})}^{\frac{1}{2}}\mathbf{a}\| \leq 1, \\ & \|\Sigma_{\varphi(\mathbf{y})}^{\frac{1}{2}}\mathbf{a}\| \leq \sqrt{\frac{1-\beta_0}{\beta_0}}t. \end{aligned} \tag{20}$$

To make the kernel work, we represent the final decision hyperplane and the optimization into a kernel form, $K(\mathbf{z}_1, \mathbf{z}_2) = \varphi(\mathbf{z}_1)^T\varphi(\mathbf{z}_2)$, namely an inner product form of the mapping data points. Due to the limited scope of this paper, we will not present a detailed kernelization procedure here. Readers interested in the details can refer to Huang et al. (2004d).

We now outline the kernelized optimization function for Biased Minimax Probability Machine as follows:

$$\begin{aligned} \min_{t,\mathbf{a}\neq\mathbf{0}} \quad & t - \mathbf{w}^T(\tilde{\mathbf{k}}_{\mathbf{x}} - \tilde{\mathbf{k}}_{\mathbf{y}}) \\ \text{s.t.} \quad & \sqrt{\frac{1}{N_{\mathbf{x}}}\mathbf{w}^T\tilde{\mathbf{K}}_{\mathbf{x}}^T\tilde{\mathbf{K}}_{\mathbf{x}}\mathbf{w}} \leq 1, \\ & \sqrt{\frac{1}{N_{\mathbf{y}}}\mathbf{w}^T\tilde{\mathbf{K}}_{\mathbf{y}}^T\tilde{\mathbf{K}}_{\mathbf{y}}\mathbf{w}} \leq \sqrt{\frac{1-\beta_0}{\beta_0}}t, \end{aligned} \tag{21}$$

which is a Second-Order Cone Program (SOCP) that has a similar form as the SOCP in (9) and can thus be solved in a similar way.

Remark. We omit the introduction of some notations here due to the space limitations. Interested readers could refer to Huang et al. (2006b).

4. Experimental results

In this section we discuss the experimental evaluation of our proposed biased learning algorithm in comparison to some state-of-the-art approaches. For a consistent evaluation, we conduct our empirical comparisons on two standard datasets for medical diagnosis: the breast-cancer dataset, and the heart disease dataset. The traditional algorithms are the *NB* classifier, the *kNN* method, and the Minimax Probability Machine (*MPM*) in this paper, along with the two BMPMs, $BMPM_{SOCP}$ and $BMPM_{FP}$.

[2] This is a common technique to tackle optimization problems.

Table 1
An overview of the breast-cancer dataset

Dataset	#Instances	#Features	#Classes
Breast-cancer dataset	699	9	2

Table 2
An overview of the heart disease dataset

Dataset	#Instances	#Features	#Classes
Heart disease dataset	270	13	2

4.1. Experimental testbeds

Two medical datasets, the breast-cancer dataset and the heart disease dataset, obtained from the UCI machine learning repository (Asuncion & Newman, 2007), are used in this experiment (Table 1).

4.1.1. Breast-cancer dataset

It has been widely employed as a benchmark dataset for evaluating biomedical diagnosis problems. The breast-cancer dataset consists of 458 instances of the benign class and 241 instances of the malignant class. Each instance is described by 9 attributes.

4.1.2. Heart disease dataset

The heart disease dataset includes 120 instances with heart disease and 150 instances without heart disease. Each instance is described by 13 attributes (Table 2).

We pre-process the datasets by removing any instances with missing attribute values since handling of these missing attribute values is out of the scope of this paper. For these two datasets, the preferred class **x** is the malignant class and the heart disease class, respectively. Therefore, the sensitivity, or the true positive rate, corresponds to the accuracy of the class **x**, and the specificity is the accuracy of the class **y**.

4.2. Experimental settings

Applying BMPM-based technique in medical diagnosis is a very straightforward task, where we separate the two classes of cases by maximizing the worst-case (minimal) probability that an "ill" case is correctly classified into the "ill" class with respect to all distributions with these means and covariance matrices, while maintaining acceptable the worst-case (minimal) probability that a healthy case is also correctly diagnosed. These probabilities can also be considered as the corresponding accuracies, namely, the sensitivity and the specificity. Therefore we just need to assume the patients to be the more important class (**x**) in the biased classification framework while assuming the healthy ones to be the less important class (**y**).

We use three performance measurements to evaluate the BMPM model. They are: (1) the Receiver Operating Characteristic (ROC), (2) Maximum Sum (MS), and (3) Test-Set Accuracy (TSA). The ROC curve plots a series of sensitivities against the corresponding one minus specificities, or the true positive rates versus the false positive rates for short. Moreover, if the ROC curves are generated with good shapes evenly distributed along their length, they can be used to evaluate biased learning algorithms by using the area under the curve. The larger the area under the curves, the higher the sensitivity for a given specificity, and hence the better the method's performance (Huang et al., 2006b).

We use another metric to perform evaluations, namely the criterion of maximum sum (MS). Instead of using the area as the metric in the ROC curve analysis, this criterion uses a typical point that achieves the largest sum of the sensitivity and the specificity (or the maximum difference between the true positive rate and the false positive rate) (Grzymala-Busse et al., 2003; Huang et al., 2006b). This criterion is originally designed to evaluate the performance for imbalanced data. In this context, the data associated with one class are far fewer than those associated with the other class. If using the traditional metric, i.e., the metric of maximizing the overall accuracy of data, the learning algorithms tend to classify all the data into the majority yet less important class; such cases can be avoided by using the MS criterion. Note that, in medical diagnosis tasks there also exist cases in which the number of the disease data is far smaller than the number of the healthy data (e.g., for certain peculiar diseases that occur rarely) (Huang et al., 2006b).

The other measurement which is used to demonstrate the effectiveness of our proposed model and strategy is Test-Set Accuracy (TSA). It consists of three measurements, i.e., Test-Set Accuracy on Class **x** ($TSA_{\mathbf{x}}$), Test-Set Accuracy on Class **y** ($TSA_{\mathbf{y}}$) and the overall Test-Set Accuracy on both classes (TSA).

To examine the effectiveness and efficiency of the learning model and proposed solving strategy, we use three reference models in our experiments. The first reference model is the Naive Bayesian classifier (NB)[3] which is an efficient classification model based on Bayes Theorem. The second reference model is based on kNN[4] which is a traditional classification tool. We also include Minimax Probability Machine (MPM)[5] for performance comparison intention. Finally, BMPM has been conducted based on both FP and SOCP frameworks. By comparing with these three models, we are able to demonstrate that the BMPM model is more reliable to handle the imbalanced medical diagnosis classification problem, and the advantages of our proposed training strategy.

To implement the SOCP-based BMPM algorithm, we adopt the standard optimization package, i.e., SeDuMe (Sturm, 1999) and YALMIP (Lofberg, 2004), to solve the Second-Order Cone Programming problem in our algorithm. The FP-based BMPM framework is based on the Rosen Gradient Projection method described in Huang et al. (2006b).

[3] http://bnt.sourceforge.net/.
[4] http://people.revoledu.com/kardi/tutorial/KNN/resources.html.
[5] http://cosmal.ucsd.edu/~gert/publications.html.

4.3. Performance evaluation

4.3.1. Test-set accuracy comparison

Table 3 shows the experimental results of Test-Set Accuracy (TSA) performance over the breast-cancer dataset.

First, as listed in the first and the second columns of Table 3, we observe that the two classifiers, $BMPM_{SOCP}$ and $BMPM_{FP}$, outperform the other three models. Take the parameter α for example, $BMPM_{SOCP}$ and $BMPM_{FP}$ achieves noticeably better performance than *MPM*, which makes the worst-case (maximum) misclassification probability much lower with the value $1-\alpha$ reduced. Second, we compare the performance of the two BMPM classifiers with the traditional classifiers, i.e., *NB* and *kNN*. The results are listed in the fourth and fifth columns of Table 3. We find that the average TSA performance, which is indicated as *TSA* in the table, of these two learning methods become closer than the BMPM models. But for the TSA of the more important class indicated as $TSA_{\mathbf{x}}$ is much lower than BMPM models. For example, the $TSA_{\mathbf{x}}$ of $BMPM_{SOCP}$ is much better than *NB* though it shows the shortcoming in the *TSA* measurement. Finally, we compare the performance of the proposed Second-Order Cone-Programming-based algorithm, i.e., $BMPM_{SOCP}$, to the Fractional-Programming-based methodology $BMPM_{FP}$. It is evident that the proposed learning algorithm outperforms its original approach.

In order to evaluate the performance substantially, the classification results of the heart disease dataset is listed in Table 4. From the experimental results, 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.

4.3.2. MS analysis

We first evaluate the BMPM approach against other algorithms based on the MS criterion. The results of breast-cancer dataset are shown in Table 5. It can be seen that the BMPMs, i.e., $BMPM_{SOCP}$ and $BMPM_{FP}$, achieve the best performance. Although NB is very close with the $BMPM_{FP}$, a significance analysis according to the traditional analysis of variance (ANOVA) shows that difference of the means of $BMPM_{SOCP}$, *NB* and *kNN* are significantly different ($p < 0.05$).

The results of the heart disease dataset are shown in Table 6. In this dataset, the BMPM models demonstrate a superiority to the other learning models. The $BMPM_{SOCP}$ and $BMPM_{FP}$ achieve the best results of 0.872 and 0.838, respectively. They are both greater than 0.827, the best result from *NB* and *kNN*. Furthermore, the ANOVA test shows that the difference of the $BMPM_{SOCP}$, $BMPM_{FP}$, and the other algorithms are significant ($p < 0.05$).

In summary, in terms of the MS criterion, our *BMPM* models demonstrate better performance when compared with other algorithms in both the breast-cancer and heart disease datasets, while the $BMPM_{SOCP}$ outperforms the original $BMPM_{FP}$.

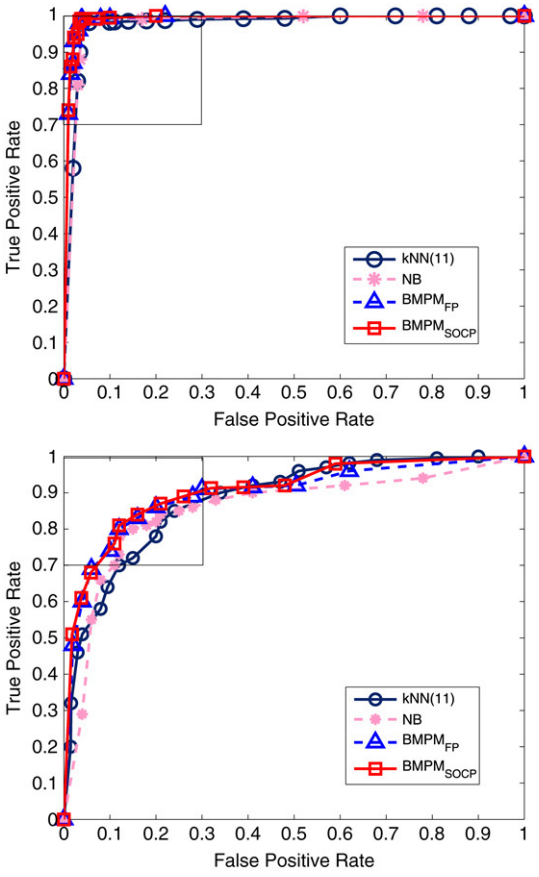

Fig. 2. Full range of the ROC curves on two datasets.

4.3.3. ROC curve analysis

It is difficult for *MPM* to generate the ROC curves due to its balanced setting. Without loss of generality, since the worst-case sensitivity and the worst-case specificity output by the *MPM* model is equal, *BMPM* will be equal or better than *MPM* in most cases, we have omitted MPM in this set of experiments.

We generate the ROC curves as illustrated in Fig. 2. It is observed that the $BMPM_{SOCP}$ and $BMPM_{FP}$ perform better than the *NB* and *kNN* classifiers for the two data collections, since the BMPM curves are above of the ones for *NB* and *kNN* methods at most cases. In addition, usually not all the portions of the ROC curve are of great interest. In general, those with a small false positive rate and a high true positive rate are most important. In light of this, we show the critical portions of Fig. 2 in Fig. 3 with more detail when the false positive rate is in the range of [0, 0.3] and the true positive rate is in the range of [0.7, 1.0], respectively. In this critical region, most parts of the ROC curve of BMPMs are above the corresponding curves of *NB* and *kNN* models in both datasets along with the $BMPM_{SOCP}$ curve is above the one of $BMPM_{FP}$, which again demonstrates the superiority of the BMPM models and our proposed $BMPM_{SOCP}$ algorithm.

More specifically, we calculate the areas under the ROC curves as illustrated in Table 7. For the breast-cancer dataset, it produces a curve with an area of 0.996 in $BMPM_{SOCP}$ and a curve with an area of 0.992 in $BMPM_{FP}$, which are both greater

Table 3
Lower bound α and test-set accuracy on the breast-cancer dataset (%)

Parameter	$BMPM_{SOCP}$	$BMPM_{FP}$	MPM	$kNN(11)$	NB
α	**98.4±0.2**↑	97.3 ± 0.1 ↑	90.1 ± 0.3	–	–
β	**50.0±0.0**	50.0 ± 0.0	90.1 ± 0.3	–	–
TSA_x	**100.0±0.1**↑	99.7 ± 0.2 ↑	96.2 ± 0.2	94.3 ± 0.1	96.1 ± 0.2
TSA_y	**92.1±0.1**	88.4 ± 0.1	97.2 ± 0.3	95.1 ± 0.2	98.6 ± 0.1
TSA	**97.2±0.2**	94.8 ± 0.1	96.7 ± 0.2	94.6 ± 0.1	97.7 ± 0.2

Table 4
Lower bound α and test-set accuracy on the heart disease dataset (%)

Parameter	$BMPM_{SOCP}$	$BMPM_{FP}$	MPM	$kNN(11)$	NB
α	**65.2±0.2**↑	61.4 ± 0.1 ↑	58.2 ± 0.1	–	–
β	**50.0±0.0**	50.0 ± 0.0	58.2 ± 0.1	–	–
TSA_x	**87.1±0.1**↑	83.5 ± 0.2 ↑	81.3 ± 0.2	81.7 ± 0.3	82.3 ± 0.2
TSA_y	**85.6±0.2**	85.2 ± 0.1	86.6 ± 0.3	82.1 ± 0.2	80.7 ± 0.1
TSA	**86.2±0.1**	84.3 ± 0.1	85.2 ± 0.1	81.4 ± 0.2	82.4 ± 0.2

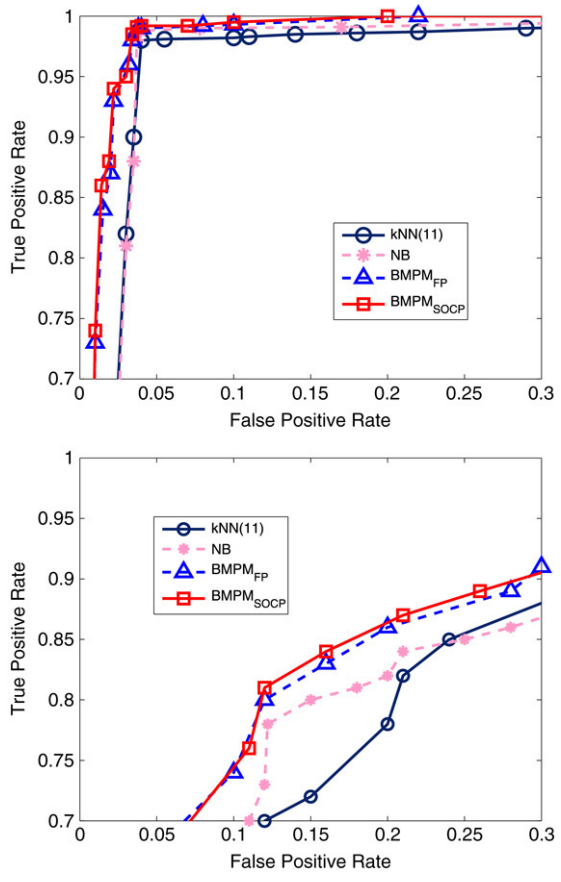

Fig. 3. Crucial part of the ROC curves on two datasets.

than those of the other methods. We also obtain similar results for the heart disease dataset.

5. Conclusion and future work

The computational complexity of our method for Biased Minimax Probability Machine (BMPM) is comparable to the quadratic program that one has to solve for the support vector machine (SVM) and Minimax Probability Machine (MPM).

Table 5
Comparison of model performance based on the MS criterion on the breast-cancer dataset

Model	Sensitivity	Specificity	(Sensitivity + Specificity)/2
$BMPM_{SOCP}$	0.995 ± 0.001	0.975 ± 0.003	**0.985 ± 0.003**
$BMPM_{FP}$	0.991 ± 0.002	0.962 ± 0.002	**0.976 ± 0.002**
$kNN(11)$	0.978 ± 0.006	0.966 ± 0.004	0.972 ± 0.004
NB	0.983 ± 0.004	0.967 ± 0.007	0.975 ± 0.006

Table 6
Comparison of model performance based on the MS criterion on the heart disease dataset

Model	Sensitivity	Specificity	(Sensitivity + Specificity)/2
$BMPM_{SOCP}$	0.892 ± 0.005	0.852 ± 0.003	**0.872 ± 0.004**
$BMPM_{FP}$	0.840 ± 0.002	0.835 ± 0.006	**0.838 ± 0.003**
$kNN(11)$	0.850 ± 0.005	0.753 ± 0.008	0.802 ± 0.006
NB	0.813 ± 0.004	0.842 ± 0.006	0.827 ± 0.005

Table 7
Comparison of model performance based on the ROC analysis

Breast-cancer dataset		Heart disease dataset	
Model	ROC area	Model	ROC area
$BMPM_{SOCP}$	**0.996 ± 0.008**	$BMPM_{SOCP}$	**0.921 ± 0.007**
$BMPM_{FP}$	**0.992 ± 0.002**	$BMPM_{FP}$	**0.902 ± 0.004**
$kNN(11)$	0.957 ± 0.003	$kNN(11)$	0.867 ± 0.005
NB	0.983 ± 0.006	NB	0.876 ± 0.006

While we have viewed this model from the viewpoint of a convex optimization problem, we believe that there is much to gain from exploiting analogies to the SVM and developing specialized optimization procedures for our model. Another direction that we are currently investigating is the extension of our model to multiway classification.

Acknowledgments

The authors thank G.R.G. Lanckriet for providing the Matlab source code of the MPM on the web, and Kaizhu

Huang and Haiqin Yang for the FP-based BMPM code. 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 VIEW Technologies Lab and the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

References

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, *6*(1), 37–66.

Asuncion, A., & Newman, D. J. (2007). *UCI machine learning repository*.

Bertsekas, D. P. (2004). *Nonlinear programming* (2nd ed.). Athena Scientific.

Cardie, C., & Nowe, N. (1997). Improving minority class prediction using case-specific feature weights. In *Proceedings of the 14th international conference on machine learnin* (pp. 57–65).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence and Research*, *16*, 321–357.

Grzymala-Busse, J. W., Goodwin, L. K., & Zhang, X. (2003). Increasing sensitivity of preterm birth by changing rule strengths. *Pattern Recognition Letters*, *24*(6), 903–910.

Huang, K., Yang, H., King, I., & Lyu, M. (2004a). Learning classifiers from imbalanced data based on biased minimax probability machine. In *Proceedings of IEEE conference on computer vision and pattern recognition*: *Vol. 2* (pp. 558–563).

Huang, K., Yang, H., King, I., & Lyu, M. (2004b). Learning large margin classifiers locally and globally. In *Proceedings of international conference on machine learning* (pp. 51–59).

Huang, K., Yang, H., King, I., & Lyu, M. (2006a). Imbalanced learning with a biased minimax probability machine. *IEEE Transactions on Systems, Man and Cybernetics (Part B)*, *36*(4), 913–923.

Huang, K., Yang, H., King, I., & Lyu, M. (2006b). Maximizing sensitivity in medical diagnosis using biased minimax probability machine. *IEEE Transactions on Biomedical Engineering*, *53*(5), 821–831.

Huang, K., Yang, H., King, I., Lyu, M., & Chan, L. (2004c). Biased minimax probability machine for medical diagnosis. In *Proceedings of annals of mathematics and artificial intelligence* (pp. 1103–1110).

Huang, K., Yang, H., King, I., Lyu, M., & Chan, L. (2004d). The minimum error minimax probability machine. *Journal of Machine Learning Research*, *5*, 1253–1286.

Jordan, M. (1995). Why the logistic function? A tutorial discussion on probabilities and neural networks (*Technical report*). MIT Computational Cognitive Science.

Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine*, *23*(1), 89–109.

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of 14th international conference on machine learning* (pp. 179–186).

Lanckriet, G., Ghaoui, L. E., Bhattacharyya, C., & Jordan, M. (2003). A robust minimax approach to classification. *Journal of Machine Learning Research*, *3*, 555–582.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of bayesian classifiers. In *Proceedings of national conference on artificial intelligence* (pp. 223–228).

Lofberg, J. (2004). YALMIP: A toolbox for modeling and optimization in MATLAB. In *Proceedings of IEEE international symposium on computer aided control system design* (pp. 284–289).

Macskassy, S. A., Hirsh, H., Banerjee, A., & Dayanik, A. A. (2001). Using text classifiers for numerical classification. In *Proceedings of the seventeenth international joint conference on artificial intelligence*.

Maloof, M., Langley, P., Binford, T., Nevatia, R., & Sage, S. (2004). Improved rooftop detection in aerial images with machine learning. *Machine Learning*, *53*, 157–191.

Nigam, K., Mccallum, A., Thrun, S., & Mitchell, T. (1999). Text classification from labeled and unlabeled documents using em. *Machine Learning*, *39*(2), 103–134.

Peng, X., & King, I. (2006a). Biased minimax probability machine active learning for relevance feedback in content-based image retrieval. In *Proceedings of international conference on intelligent data engineering and automated learning* (pp. 953–960).

Peng, X., & King, I. (2006b). Imbalanced learning in relevance feedback with biased minimax probability machine for image retrieval tasks. In *Proceedings of international conference on neural information processing* (pp. 342–351).

Peng, X., & King, I. (2007). Large scale imbalanced classification with biased minimax probability machine. In *Proceedings of 20th international joint conference on neural networks* (pp. 1685–1690).

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Sturm, J. F. (1999). Using sedumi 1.02: A matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software*, *11*, 625–653.

West, D., & West, V. (2000). Model selection for a medical diagnostic decision support system: A breast cancer detection case. *Artificial Intelligence in Medicine*, *20*(3), 183–204.