Maximizing Sensitivity in Medical Diagnosis Using Biased Minimax Probability Machine

Kaizhu Huang*, Haiqin Yang, Irwin King, *Member, IEEE*, and Michael R. Lyu, *Fellow, IEEE*

***Abstract*—The challenging task of medical diagnosis based on machine learning techniques requires an inherent bias, i.e., the diagnosis should favor the "ill" class over the "healthy" class, since misdiagnosing a patient as a healthy person may delay the therapy and aggravate the illness. Therefore, the objective in this task is not to improve the overall accuracy of the classification, but to focus on improving the sensitivity (the accuracy of the "ill" class) while maintaining an acceptable specificity (the accuracy of the "healthy" class). 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. However, it remains uncertain whether these methods can improve the classification performance systematically. In this paper, by engaging a novel learning tool, 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. More specifically, the BMPM directly controls the worst case accuracies to incorporate a bias toward the "ill" class. Moreover, in a distribution-free way, the BMPM derives the decision rule in such a way as to maximize the worst case sensitivity while maintaining an acceptable worst case specificity. By directly controlling the accuracies, the BMPM provides a more rigorous way to handle medical diagnosis; by deriving a distribution-free decision rule, the BMPM distinguishes itself from a large family of classifiers, namely, the generative classifiers, where an assumption on the data distribution is necessary. We evaluate the performance of the model and compare it with three traditional classifiers: the k-nearest neighbor, the naive Bayesian, and the C4.5. The test results on two medical datasets, the breast-cancer dataset and the heart disease dataset, show that the BMPM outperforms the other three models.**

***Index Terms*—Biased classification, medical diagnosis, minimax probability machine, worst case accuracy.**

I. Introduction

APPLYING machine learning techniques to medical diagnosis tasks has the advantages of saving time and reducing cost. The challenge is, based on data from previous medical cases, to construct a rule which can be used to discriminate between healthy subjects and patients. The decision rule, called the classifier, is trained by using a number of observations with known class labels. This approach is also known as "supervised learning." In the simplest medical context, only two classes, i.e., the "healthy" class and the "ill" class, are considered. Moreover, with these two kinds of data, one is more significant than the other. The "ill" class is obviously more important than the "healthy" class since misdiagnosing a patient as a healthy person may delay the therapy and aggravate the illness. Therefore, the objective in medical diagnosis is inherently biased, i.e., instead of improving the overall accuracy, we should focus on improving the accuracy of the "ill" class, called *sensitivity*, while maintaining the accuracy of the "healthy" class, called *specificity*, at an acceptable level [9].

In the machine learning literature, many different techniques can be applied to medical diagnosis [18], [34], including the naive Bayesian (NB) method [21], the logistic regression [16], the k-nearest neighbor (k-NN) method [1], and the decision tree C4.5 [30]. However, these standard learning tools, originally designed for seeking an accurate performance over a full range of data, need to be modified to favor the more important class, i.e., the "ill" class, over the less important class, i.e., the "healthy" class. Currently, techniques such as the methods of sampling [7], [19], [22], the methods of adapting the thresholds [25], [28], and the methods of adjusting cost matrices [6], [25] can be used to incorporate a certain bias into the learning methods.[1] However, all these methods have shortcomings. For example, the sampling methods favor the more important class by down-sampling (removing) some instances of the less important class or up-sampling (duplicating) some instances of the more important class. Either approach seems to be problematical: down-sampling will lose information while up-sampling may introduce noise. According to [28], one open question is whether simply varying the skewness of the data distribution can improve predictive performance systematically. In the case of the methods of adjusting cost matrices or adapting weights, it is usually hard to build direct quantitative connections between the intermediate factors, i.e., the costs or the weights, and the biased classification performance. These methods, therefore, fail to provide a rigorous approach to the task of medical diagnosis.

In summary, the problems seem to root from the fact that these standard learning tools were not originally designed to achieve the medical diagnosis goal, i.e., the biased configuration, but to maximize the overall accuracy. Either preprocessing the data (in the methods of sampling and the methods of weighting each class) [7], [19], [22] or postprocessing the tools themselves (in the methods of adapting thresholds) [25], [28] and then forcing

Manuscript received April 19, 2005; revised September 25, 2005. This work was supported the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project CUHK4205/04E and Project CUHK4235/04E. *Asterisk indicates corresponding author.*

*K. Huang is with the Information Technology Laboratory, Fujitsu Research and Development Center Co., Ltd., Beijing 100016, China (e-mail: kzhuang@frdc.fujitsu.com).

H. Yang is with Titanium Technology Limited, Shenzhen 518020, China (e-mail: austin.yang@titanium-tech.com).

I. King and M. R. Lyu are with the Department of Computer Science and Engineering, the Chinese University of Hong Kong, Hong Kong, China (e-mail: king@cse.cuhk.edu.hk; lyu@cse.cuhk.edu.hk).

Digital Object Identifier 10.1109/TBME.2006.872819

[1]These methods were originally used to deal with the classification of skewed data, where one class contains far fewer data than the other class. However, in terms of imposing a bias on one class, this task is similar to the medical diagnosis task. Therefore, these techniques are appropriate for the latter task.

these standard tools into the application of medical diagnosis is rather a patchwork approach and, thus, they fail to provide rigorous and systematic treatment of biased classification tasks.

Therefore, in this paper, directly aiming to achieve the goal of solving the medical diagnosis problem, we present a novel learning tool named the biased minimax probability machine (BMPM) [14] to deal with medical diagnosis in a more elegant way. As a significant extension of the minimax probability machine (MPM) [20], the BMPM approach contains several advantages over traditional methods. First, in contrast to the preprocessing and postprocessing methods, the BMPM approaches the medical diagnosis problem directly. As shown later in this paper, the BMPM constructs the decision rule by maximizing the worst case sensitivity under all possible distributions with given mean and covariance matrices, while keeping the specificity acceptable. More importantly, when certain distributions, in particular a Gaussian distribution, are assumed for the data, our model is able to maximize the real sensitivity with respect to future data. The direct control of the accuracies rather than of intermediate factors provides a more rigorous way to handle the biased classification task. Second, by deriving the worst case decision rule under all possible distributions with given mean and covariance matrices, the BMPM becomes distribution-independent. This distinguishes the BMPM from a large family of learning methods, namely, the generative classifiers [11], [12], [15] including the NB classifier and the logistical regression, which have to make specific assumptions about the data distribution and, hence, lack general validity in real tasks. Third, although the BMPM contains the above advantages, it does not sacrifice efficiency for them. The optimization of this model can in practice be transformed to a concave-convex fractional programming (FP) [31] problem or a pseudoconcave problem and, therefore, can be solved efficiently.

The paper is organized as follows. In Section II, we review the MPM briefly. In Section III, we present the linear BMPM (BMPML), showing how to achieve the objective of the medical diagnosis directly. In Section IV, we then kernelize the BMPML to attack nonlinear classification tasks. In Section V, we discuss the evaluation criteria and propose to modify those traditional machine learning approaches for medical diagnosis. In Section VI, we first illustrate our model with a synthetic dataset. We then apply it to real-world medical diagnosis datasets and compare its performance with three traditional classifiers. In Section VII, we make some observations about the BMPM model. Finally, our conclusions are set out in Section VIII.

II. Review of MPM

The notation in this paper will largely follow that of [20]. Suppose two random n-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ represent two classes of data, where $\mathbf{x}$ belongs to the family of distributions with a given mean $\bar{\mathbf{x}}$ and a covariance $\Sigma_{\mathbf{x}}$, denoted as $\mathbf{x} \sim (\bar{\mathbf{x}}, \Sigma_{\mathbf{x}})$; similarly, $\mathbf{y}$ belongs to the family of distributions with a given mean $\bar{\mathbf{y}}$ and a covariance $\Sigma_{\mathbf{y}}$, denoted as $\mathbf{y} \sim (\bar{\mathbf{y}}, \Sigma_{\mathbf{y}})$. Here, $\mathbf{x}, \mathbf{y}, \bar{\mathbf{x}}, \bar{\mathbf{y}} \in \mathbb{R}^n$, and $\Sigma_{\mathbf{x}}, \Sigma_{\mathbf{y}} \in \mathbb{R}^{n\times n}$. In this paper, class $\mathbf{x}$ represents the "ill" class and $\mathbf{y}$ represents the "healthy" class.

The MPM attempts to determine the hyperplane $\mathbf{a}^T\mathbf{z} = b$ ($\mathbf{a} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, $\mathbf{z} \in \mathbb{R}^n$, $b \in \mathbb{R}$, and superscript T denotes the transpose) which can separate two classes of data with the maximal probability. The formulation for the MPM model is written as follows:

$$\max_{\alpha,\mathbf{a}\neq\mathbf{0},b} \quad \alpha \quad \text{s.t.} \quad \inf_{\mathbf{x}\sim(\bar{\mathbf{x}},\Sigma_{\mathbf{x}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x} \geq b\} \geq \alpha \quad \inf_{\mathbf{y}\sim(\bar{\mathbf{y}},\Sigma_{\mathbf{y}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \alpha$$

where α represents the lower bound of the accuracy for future data, namely, the worst case accuracy. Future points $\mathbf{z}$ for which $\mathbf{a}^T\mathbf{z} \geq b$ are then classified as the class $\mathbf{x}$; otherwise they are judged as the class $\mathbf{y}$. This derived decision hyperplane is claimed to minimize the worst case probability of misclassification, or the error rate, of future data. Furthermore, this problem can be transformed to a convex optimization problem, or more specifically, a second-order cone programming problem [23], [27].

As observed from the above formulation, this model actually assumes that two classes have the same importance. Hence, it makes the worst case accuracies for two classes the same. However, in real applications, especially in medical diagnosis, two classes of data are usually biased, i.e., the disease class is often more important than the healthy class. Therefore, it is more appropriate to take the inherit bias nature into account in this context. In the following, we develop an extension of MPM, i.e., the BMPM, which is more appropriate for medical diagnosis.

III. BMPML for Medical Diagnosis

In this section, we present the linear biased minimax framework, designed to achieve the goal of the medical diagnosis directly. We first introduce the model definition and then propose methods to solve the optimization. Following that, we analyze the case in which a certain distribution is assumed for the data.

A decision hyperplane $f(\mathbf{z}) = \mathbf{a}^T\mathbf{z} - b$, where $\mathbf{a} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ and $b \in \mathbb{R}$, is constructed such that, for an ill case, $f(\mathbf{z}) \geq 0$, and for a healthy case, $f(\mathbf{z}) \leq 0$. Similar to MPM, we aim to achieve a decision hyperplane in the worst case scenario, i.e., 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. This is formulated as follows:

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

$$\inf_{\mathbf{y}\sim(\bar{\mathbf{y}},\Sigma_{\mathbf{y}})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \beta \tag{2}$$

$$\beta \geq \beta_0. \tag{3}$$

Here, α means the lower bound of the probability (accuracy) for the classification of future cases of the class $\mathbf{x}$; in other words,

α is the worst case sensitivity. Similarly, β is the lower bound of the accuracy of the class $\mathbf{y}$, i.e., the worst case specificity.

In our extension, we not only maximize the worst case sensitivity, but also maintain the worst case specificity at a preset acceptable level given by β_0. This is more appropriate in biased classifications. Note that the BMPM only makes assumptions on the mean and covariances and does not assume specific distributions over data. This presents one of the major distinctions between our model and other generative models, whose assumption on data distribution does not always coincide with the real situation. More importantly, as shown shortly in this paper, when a particular distribution (e.g., Gaussianity) is assumed for the data, maximizing the worst case sensitivity results in maximization of the actual sensitivity.

A. Optimization Method

In the following, we discuss the optimization procedure for the BMPM model. We first define two optimization problems as follows.

Definition 1: Optimization problem I is defined as follows:

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

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

$$b-\mathbf{a}^T\bar{\mathbf{y}} \geq \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \tag{6}$$

$$\beta \geq \beta_0 \tag{7}$$

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

Definition 2: Optimization problem II is defined as the following typical linear constrained FP problem[2] :

$$\max_{\mathbf{a}\neq\mathbf{0}} \quad \frac{f(\mathbf{a})}{g(\mathbf{a})}, \quad \text{s.t.} \quad \mathbf{a}\in A=\left\{\mathbf{a}|\mathbf{a}^T(\bar{\mathbf{x}}-\bar{\mathbf{y}})=1\right\} \tag{8}$$

where $f(\mathbf{a})=1-\kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}$, $g(\mathbf{a})=\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}$.

Lemma 1: The following statements are true: a) Optimization problem I is equivalent to the optimization of BMPM. b) Optimization problem I is equivalent to Optimization problem II.

See the Appendix for the detailed proof.

In the following, we further show that the FP problem of (8) is solvable.

Lemma 2: The above FP problem (8) is a strictly quasiconcave problem and is, thus, solvable.

Proof: It is easy to see that the domain A is a convex set on $\mathbb{R}^n$, and $f(\mathbf{a})$ and $g(\mathbf{a})$ are differentiable on A. Moreover, although in theory $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{y}}$ can only be guaranteed as positive semi-definite matrices, in practice they can be made positive definite matrices by adding a small positive turbulence in their diagonal elements [20]. Therefore, $f(\mathbf{a})$ is a concave function on A and $g(\mathbf{a})$ is a convex function on A. Then $f(\mathbf{a})/g(\mathbf{a})$ is a concave-convex FP or a pseudoconcave problem. Hence, it is strictly quasiconcave on A according to [31].[3]

Therefore, every local maximum is a global maximum [31]. In other words, this FP problem is solvable. □

[2]A linear constrained FP problem can be informally defined as a family of optimization problems, where the optimization objective is in a fractional form and the constraints are in linear forms.

[3]A function $g(x)$ is quasiconcave if we have $g(\lambda\mathbf{x}'+(1-\lambda)\mathbf{x}'') \geq \min\{g(\mathbf{x}'),g(\mathbf{x}'')\}$, where $0\leq\lambda\leq 1$.

Many methods can be used to solve this problem. For example, a conjugate gradient method can solve this problem in at most n (the data dimension) steps if the initial point is suitably assigned [3]. In each step, the computational cost to calculate the conjugate gradient is $O(n^2)$. Thus, this method will have a worst case time complexity of $O(n^3)$. Adding the time cost to estimate $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$, $\Sigma_{\mathbf{x}}$, $\Sigma_{\mathbf{y}}$, the total cost is $O(n^3+Nn^2)$, where N is the number of the data points. This computational cost is in the same order as the MPM [20] and the quadratic program in solving the linear support vector machine [32].

In this paper, we use the Rosen gradient projection method [3] to find the solution of this concave-convex FP problem, which is proved to converge to a local maximum with a linear convergence rate in the worst case. Moreover, the local maximum will be exactly the global maximum in this problem.

With reference to the proof of Lemma 1 in the Appendix, the optimal b, denoted by b^*, is obtained by

$$\begin{aligned} b^* &= \mathbf{a}^{*T}\bar{\mathbf{y}}+\kappa(\beta^*)\sqrt{\mathbf{a}^{*T}\Sigma_{\mathbf{y}}\mathbf{a}^*} \\ &= \mathbf{a}^{*T}\bar{\mathbf{x}}-\kappa(\alpha^*)\sqrt{\mathbf{a}^{*T}\Sigma_{\mathbf{x}}\mathbf{a}^*} \end{aligned}$$

where $\mathbf{a}^*$ is the optimal solution of the FP problem in (8).[4]

B. Assuming Specific Distributions

Although the BMPM model assumes no specific distribution for the data, it is interesting to explore the properties of BMPM when some specific distribution is assumed. In the following, we show that when certain distributions, in particular a Gaussian distribution, are assumed for the data, maximizing the worst case sensitivity strictly leads to maximizing the real sensitivity with respect to future data.

Assuming $\mathbf{x}$ and $\mathbf{y}$ are two sets of data with Gaussian distributions $\mathcal{N}(\bar{\mathbf{x}},\mathbf{\Sigma_x})$ and $\mathcal{N}(\bar{\mathbf{y}},\mathbf{\Sigma_y})$, respectively, (1) becomes

$$\begin{aligned} \inf_{\mathbf{x}\sim\mathcal{N}(\bar{\mathbf{x}},\mathbf{\Sigma_x})} \mathbf{Pr}\{\mathbf{a}^T\mathbf{x}\geq b\} &= \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\bar{\mathbf{x}},\mathbf{\Sigma_x})}\{\mathbf{a}^T\mathbf{x}\geq b\} \\ &= \mathbf{Pr}\left\{\mathcal{N}(0,1)\geq\frac{b-\mathbf{a}^T\bar{\mathbf{x}}}{\sqrt{\mathbf{a}^T\mathbf{\Sigma_x}\mathbf{a}}}\right\} \\ &= 1-\Phi\left(\frac{b-\mathbf{a}^T\bar{\mathbf{x}}}{\sqrt{\mathbf{a}^T\mathbf{\Sigma_x}\mathbf{a}}}\right) \\ &= \Phi\left(\frac{-b+\mathbf{a}^T\bar{\mathbf{x}}}{\sqrt{\mathbf{a}^T\mathbf{\Sigma_x}\mathbf{a}}}\right)\geq\alpha \end{aligned} \tag{9}$$

where $\Phi(z)$ is the cumulative distribution function for the standard Gaussian distribution

$$\Phi(z)=\mathbf{Pr}\{\mathcal{N}(0,1)\leq z\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp\left(-\frac{s^2}{2}\right)ds.$$

Due to the monotonic nature of $\Phi(z)$, we can further write (9) as

$$-b+\mathbf{a}^T\bar{\mathbf{x}}\geq\Phi^{-1}(\alpha)\sqrt{\mathbf{a}^T\mathbf{\Sigma_x}\mathbf{a}}.$$

[4]The inequalities of (17) in the Appendix will become equalities at the maximum point.

(2) can be reformulated in a similar fashion. The optimization of the BMPM model is then changed to

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

$$-b+\mathbf{a}^T\overline{\mathbf{x}} \geq \Phi^{-1}(\alpha)\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{a}} \tag{10}$$

$$b-\mathbf{a}^T\overline{\mathbf{y}} \geq \Phi^{-1}(\beta)\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{a}} \tag{11}$$

$$\beta \geq \beta_0. \tag{12}$$

The above optimization is nearly the same as (4) subjected to the constraints of (5)–(7), except that $\kappa(\alpha)$ is equal to $\Phi^{-1}(\alpha)$, instead of $\sqrt{\alpha/(1-\alpha)}$. Thus, it can be similarly solved based on the proposed FP method. From the proof of Lemma 1 (in which (17) will change to an equality), we can know (10) and (11) will eventually become equalities. Traced back to (9), the equalities imply that α and β will have achieved their upper bounds. This means that the worst case accuracy (sensitivity) eventually changes to the real accuracy.

It is interesting to make an analysis of BMPM when other general distributions are assumed. By analogy with the Gaussian case, assuming $\mathbf{x}\sim\mathcal{S}(\overline{\mathbf{x}},\boldsymbol{\Sigma}_{\mathbf{x}})$, $\mathbf{y}\sim\mathcal{S}(\overline{\mathbf{y}},\boldsymbol{\Sigma}_{\mathbf{y}})$, where $\mathcal{S}$ means a specific distribution, we have

$$\inf_{\mathbf{x}\sim\mathcal{S}(\overline{\mathbf{x}},\boldsymbol{\Sigma}_{\mathbf{x}})}\mathbf{Pr}\{\mathbf{a}^T\mathbf{x}\geq b\}=\mathbf{Pr}_{\mathbf{x}\sim\mathcal{S}(\overline{\mathbf{x}},\boldsymbol{\Sigma}_{\mathbf{x}})}\{\mathbf{a}^T\mathbf{x}\geq b\}.$$

We note that the random variable $\mathbf{a}^T\mathbf{x}$ contains the mean $\mathbf{a}^T\overline{\mathbf{x}}$ and the variance $\mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{a}$. Thus, the normalized random variable $(\mathbf{a}^T\mathbf{x}-\mathbf{a}^T\overline{\mathbf{x}})/\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{a}}$ will have the mean 0 and the variance 1. If the distribution of the normalized random variable $(\mathbf{a}^T\mathbf{x}-\mathbf{a}^T\overline{\mathbf{x}})/\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{a}}$, denoted as $\mathcal{NS}$, is independent of $\mathbf{a}$, as the case in Gaussian distribution, a formulation similar to that in the Gaussian case can be easily derived, except that $\Phi(z)$ is changed to $\mathbf{Pr}\{\mathcal{NS}(0,1)\leq z\}$. Otherwise, it may not be easy to incorporate the distributional information into the optimization of BMPM. Further exploration on this topic is deserved. In summary, we incorporate the above analysis into Lemma 3.

Lemma 3: Assuming the data are under a specific distribution $\mathcal{S}$, if the distribution of the normalized random variable $(\mathbf{a}^T\mathbf{x}-\mathbf{a}^T\overline{\mathbf{x}})/\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{a}}$, denoted as $\mathcal{NS}$, is independent of $\mathbf{a}$, the BMPM model optimization maximizes the real sensitivity with respect to future data.

Another interesting finding is that, given an n-dimensional random variable $\mathbf{x}$, a linear combination of its component variable $\mathbf{x}_i$, $1\leq i\leq n$, namely $\mathbf{a}^T\mathbf{x}$ tends toward a Gaussian distribution, as n grows. This shows that, when the dimension n grows and the data distribution is unknown, it may be suitable to use $\Phi^{-1}(\alpha)$, the inverse function of the normal cumulative distribution, instead of $\sqrt{\alpha/(1-\alpha)}$, to perform the optimization of BMPM.

IV. Kernelized BMPM

The classifier derived above from the BMPM is given in a linear configuration. To handle more general cases, namely nonlinear classification tasks, we need to develop methods to extend the BMPML. Therefore, in this section, we first seek to use the kernelization trick 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. Then, we propose a feasible algorithm to solve the kernelized optimization problem.

A. Kernelizing the BMPM

Let $\{\mathbf{x}_i\}_{i=1}^{N_\mathbf{x}}$ and $\{\mathbf{y}_j\}_{j=1}^{N_\mathbf{y}}$ represent the training data for the class $\mathbf{x}$ and the class $\mathbf{y}$, respectively, and be mapped as $\mathbf{x}\rightarrow\varphi(\mathbf{x})\sim(\overline{\varphi(\mathbf{x})},\boldsymbol{\Sigma}_{\varphi(\mathbf{x})})$, and $\mathbf{y}\rightarrow\varphi(\mathbf{y})\sim(\overline{\varphi(\mathbf{y})},\boldsymbol{\Sigma}_{\varphi(\mathbf{y})})$, where $\varphi:\mathbb{R}^n\rightarrow\mathbb{R}^f$ is a mapping function. The corresponding linear classifier in $\mathbb{R}^f$ is $f(\mathbf{z})=\mathbf{a}^T\varphi(\mathbf{z})-b$, where $\mathbf{a}\in\mathbb{R}^f\setminus\{\mathbf{0}\}$, $\varphi(\mathbf{z})\in\mathbb{R}^f$, and $b\in\mathbb{R}$. Likewise, the transformed FP optimization of BMPM can be written as

$$\max_{\mathbf{a}\neq\mathbf{0}}\frac{1-\kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_{\varphi(\mathbf{y})}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_{\varphi(\mathbf{x})}\mathbf{a}}}\ \text{s.t.}\ \mathbf{a}^T\left(\overline{\varphi(\mathbf{x})}-\overline{\varphi(\mathbf{y})}\right)=1. \tag{13}$$

To make the kernelization trick work, we need to represent the optimization and the final decision hyperplane into a kernel form, $K(\mathbf{z}_1,\mathbf{z}_2)=\varphi(\mathbf{z}_1)^T\varphi(\mathbf{z}_2)$, i.e., an inner product form of the mapping data points.

We reformulate the optimization and the decision hyperplane in the kernelized form as follows.

Let $\mathbf{a}=\mathbf{a}_\mathbf{p}+\mathbf{a}_\mathbf{v}$, where $\mathbf{a}_\mathbf{p}$ is the projection of $\mathbf{a}$ in the space spanned by all the training data, i.e., $\{\varphi(\mathbf{x}_i)\}_{i=1}^{N_\mathbf{x}}$ and $\{\varphi(\mathbf{y}_j)\}_{j=1}^{N_\mathbf{y}}$ and $\mathbf{a}_\mathbf{v}$ is the orthogonal component of $\mathbf{a}$ in this span space. The component $\mathbf{a}_\mathbf{v}$ vanishes in the optimization (13) by using $\mathbf{a}_\mathbf{v}^T\varphi(\mathbf{x}_i)=0$ and $\mathbf{a}_\mathbf{v}^T\varphi(\mathbf{y}_j)=0$. This implies that the optimal $\mathbf{a}$ is in the space spanned by all the training data and, thus, can be written as a linear combination of the training data. We write this linear combination as follows:

$$\mathbf{a}=\sum_{i=1}^{N_\mathbf{x}}\mu_i\varphi(\mathbf{x}_i)+\sum_{j=1}^{N_\mathbf{y}}\upsilon_j\varphi(\mathbf{y}_j) \tag{14}$$

where $\mu_i,\upsilon_j\in\mathbb{R}$, $i=1,\ldots,N_\mathbf{x}$, and $j=1,\ldots,N_\mathbf{y}$ are coefficients. Moreover, four plug-in estimated parameters for the mean and covariance matrices can be written as:

$$\overline{\varphi(\mathbf{x})}=\frac{1}{N_\mathbf{x}}\sum_{i=1}^{N_\mathbf{x}}\varphi(\mathbf{x}_i),\quad \overline{\varphi(\mathbf{y})}=\frac{1}{N_\mathbf{y}}\sum_{j=1}^{N_\mathbf{y}}\varphi(\mathbf{y}_j)$$

$$\Sigma_{\varphi(\mathbf{x})}=\frac{1}{N_\mathbf{x}}\sum_{i=1}^{N_\mathbf{x}}\left(\varphi(\mathbf{x}_i)-\overline{\varphi(\mathbf{x})}\right)\left(\varphi(\mathbf{x}_i)-\overline{\varphi(\mathbf{x})}\right)^T$$

$$\Sigma_{\varphi(\mathbf{y})}=\frac{1}{N_\mathbf{y}}\sum_{j=1}^{N_\mathbf{y}}\left(\varphi(\mathbf{y}_j)-\overline{\varphi(\mathbf{y})}\right)\left(\varphi(\mathbf{y}_j)-\overline{\varphi(\mathbf{y})}\right)^T.$$

Substituting (14) and the above four plug-in estimated parameters into the optimization problem (13), we can obtain a kernelized version

$$\max_{\mathbf{w}\neq\mathbf{0}}\frac{1-\kappa(\beta_0)\sqrt{\frac{1}{N_\mathbf{y}}\mathbf{w}^T\tilde{\mathbf{K}}_\mathbf{y}^T\tilde{\mathbf{K}}_\mathbf{y}\mathbf{w}}}{\sqrt{\frac{1}{N_\mathbf{x}}\mathbf{w}^T\tilde{\mathbf{K}}_\mathbf{x}^T\tilde{\mathbf{K}}_\mathbf{x}\mathbf{w}}}$$
$$\text{s.t.}\ \mathbf{w}^T(\tilde{\mathbf{k}}_\mathbf{x}-\tilde{\mathbf{k}}_\mathbf{y})=1. \tag{15}$$

In the above, $\mathbf{w} = [\mu_1, \ldots, \mu_{N_\mathbf{x}}, \upsilon_1, \ldots, \upsilon_{N_\mathbf{y}}]^T$ and $\tilde{\mathbf{k}}_\mathbf{x}, \tilde{\mathbf{k}}_\mathbf{y} \in \mathbb{R}^{N_\mathbf{x}+N_\mathbf{y}}$ with

$$[\tilde{\mathbf{k}}_\mathbf{x}]_i = \frac{1}{N_\mathbf{x}} \sum_{j=1}^{N_\mathbf{x}} \mathbf{K}(\mathbf{x}_j, \mathbf{z}_i), \quad [\tilde{\mathbf{k}}_\mathbf{y}]_i = \frac{1}{N_\mathbf{y}} \sum_{j=1}^{N_\mathbf{y}} \mathbf{K}(\mathbf{y}_j, \mathbf{z}_i)$$

where $\mathbf{z}_i = \mathbf{x}_i$ for $i = 1, 2, \ldots, N_\mathbf{x}$ and $\mathbf{z}_i = \mathbf{y}_{i-N_\mathbf{x}}$ for $i = N_\mathbf{x}+1, N_\mathbf{x}+2, \ldots, N_\mathbf{x}+N_\mathbf{y}$. $\tilde{\mathbf{K}}$ is given by

$$\tilde{\mathbf{K}} = \begin{pmatrix} \tilde{\mathbf{K}}_\mathbf{x} \\ \tilde{\mathbf{K}}_\mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{K}_\mathbf{x} - \mathbf{1}_{N_\mathbf{x}} \tilde{\mathbf{k}}_\mathbf{x}^T \\ \mathbf{K}_\mathbf{y} - \mathbf{1}_{N_\mathbf{y}} \tilde{\mathbf{k}}_\mathbf{y}^T \end{pmatrix},$$

where $\mathbf{1}_{N_\mathbf{x}}$ ($\mathbf{1}_{N_\mathbf{y}}$) is an $N_\mathbf{x}$-dimension ($N_\mathbf{y}$-dimension) column vector with the values of all elements equal to one. $\mathbf{K}_\mathbf{x}$ and $\mathbf{K}_\mathbf{y}$ are the matrices formed by the first $N_\mathbf{x}$ rows and the last $N_\mathbf{y}$ rows of the Gram matrix $\mathbf{K}$, respectively, which is defined as $\mathbf{K}_{ij} = \varphi(\mathbf{z}_i)^T \varphi(\mathbf{z}_j)$.

Similarly, the optimal b in the kernelized version, represented by b^*, can be obtained as

$$\begin{aligned} b^* &= \mathbf{w}^{*T} \tilde{\mathbf{k}}_\mathbf{y} + \kappa(\beta^*) \sqrt{\frac{1}{N_\mathbf{y}} \mathbf{w}^{*T} \tilde{\mathbf{K}}_\mathbf{y}^T \tilde{\mathbf{K}}_\mathbf{y} \mathbf{w}^*} \\ &= \mathbf{w}^{*T} \tilde{\mathbf{k}}_\mathbf{x} - \kappa(\alpha^*) \sqrt{\frac{1}{N_\mathbf{x}} \mathbf{w}^{*T} \tilde{\mathbf{K}}_\mathbf{x}^T \tilde{\mathbf{K}}_\mathbf{x} \mathbf{w}^*}, \end{aligned}$$

where $\mathbf{w}^*$, α^*, and β^* are the optimum values given by the above optimization procedure. The kernelized decision hyperplane can be written as:

$$f(\mathbf{z}) = \sum_{i=1}^{N_\mathbf{x}} \mathbf{w}_i^* K(\mathbf{z}, \mathbf{x}_i) + \sum_{i=1}^{N_\mathbf{y}} \mathbf{w}_{N_\mathbf{x}+i}^* K(\mathbf{z}, \mathbf{y}_i) - b^*.$$

B. Solving the Kernelized BMPM

In this section, we present a parametric method to solve the FP problem [31] involved in the kernelized BMPM. When compared with Gradient methods, this approach is relatively slow, but it need not calculate the gradient in each step and, hence, may avoid accumulated errors. Moreover, for brevity, we still use the unkernelized version to present the algorithm since (16) has a form similar to the unkernelized version of (8).

According to the parametric method, the fractional function, $f(\mathbf{a})/g(\mathbf{a})$ can be iteratively optimized in two steps:

Step 1) Find $\mathbf{a}$ by maximizing $f(\mathbf{a}) - \lambda g(\mathbf{a})$ in the domain A, where $\lambda \in \mathbb{R}$ is the newly introduced parameter.

Step 2) Update λ by setting it to $f(\mathbf{a})/g(\mathbf{a})$.

According to [31], the maximum of λ, namely, the maximum solution of the FP problem, is guaranteed to converge when these steps are iterated.

In the following, we propose to solve the maximization problem in Step 1. Replacing $f(\mathbf{a})$ and $g(\mathbf{a})$, we expand the optimization problem as

$$\begin{aligned} \max_{\mathbf{a} \neq \mathbf{0}} \quad & 1 - \kappa(\beta_0) \sqrt{\mathbf{a}^T \Sigma_\mathbf{y} \mathbf{a}} - \lambda \sqrt{\mathbf{a}^T \Sigma_\mathbf{x} \mathbf{a}} \\ \text{s.t.} \quad & \mathbf{a}^T (\bar{\mathbf{x}} - \bar{\mathbf{y}}) = 1. \end{aligned} \tag{16}$$

Equation (16) is equivalent to $\min_{\mathbf{a} \neq \mathbf{0}} \kappa(\beta_0) \sqrt{\mathbf{a}^T \Sigma_\mathbf{y} \mathbf{a}} + \lambda \sqrt{\mathbf{a}^T \Sigma_\mathbf{x} \mathbf{a}}$ under the same constraint. By writing $\mathbf{a} = \mathbf{a}_0 + \mathbf{F}\mathbf{u}$, where $\mathbf{a}_0 = (\bar{\mathbf{x}} - \bar{\mathbf{x}})/\|\bar{\mathbf{x}} - \bar{\mathbf{x}}\|_2^2$ and $\mathbf{F} \in \mathbb{R}^{n \times (n-1)}$ is an orthogonal matrix whose columns span the subspace of vectors orthogonal to $\bar{\mathbf{x}} - \bar{\mathbf{x}}$, an equivalent form without the constraint can be obtained

$$\min_\mathbf{u} \quad \lambda \left\| \Sigma_\mathbf{x}^{\frac{1}{2}} (\mathbf{a}_0 + \mathbf{F}\mathbf{u}) \right\|_2 + \kappa(\beta_0) \left\| \Sigma_\mathbf{y}^{\frac{1}{2}} (\mathbf{a}_0 + \mathbf{F}\mathbf{u}) \right\|_2.$$

The above optimization can further be transformed as follows:

$$\min_{\mathbf{u}, \eta > 0, \xi > 0} \left\{ \eta + \frac{\lambda^2}{\eta} \left\| \Sigma_\mathbf{x}^{\frac{1}{2}} (\mathbf{a}_0 + \mathbf{F}\mathbf{u}) \right\|_2^2 + \xi + \frac{\kappa(\beta_0)^2}{\xi} \left\| \Sigma_\mathbf{y}^{\frac{1}{2}} (\mathbf{a}_0 + \mathbf{F}\mathbf{u}) \right\|_2^2 \right\}.$$

This optimization form is very similar to the one in the MPM [20] and can also be solved by using an iterative least-squares approach [3], [20].

V. Modifying Learning Algorithms for Medical Diagnosis

In this section, we first discuss two practical performance metrics in order to evaluate the performance of the BMPM model against other traditional learning algorithms. These traditional algorithms are the k-NN method, the NB classifier, and the C4.5 method in this paper. Next, we propose to tailor the BMPM and those traditional learning models to these two metrics.

A. Practical Performance Metrics

In real applications, the objective of medical diagnosis is to maximize the sensitivity while maintaining the specificity at a prespecified level set by experts. However, when there are no experts at hand, we have to plot a series of sensitivities against the corresponding specificities in order to evaluate the biased learning. This performance metric is well-known as the receiver operating characteristic (ROC) analysis [29], [33]. More precisely, the ROC curve plots a series of sensitivities (true positive rates) against the corresponding one minus specificities (the false positive rates). Fig. 1 illustrates an artificially generated ROC curve. As discussed in [25], if the ROC curves are generated with good shapes evenly distributed along their length, they can be used to evaluate biased learning models 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 model's performance.

We can also 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) [2], [9]. As seen in Fig. 1, the filled red point in the ROC curve represents the one that achieves the largest sum of the sensitivity and the specificity. 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

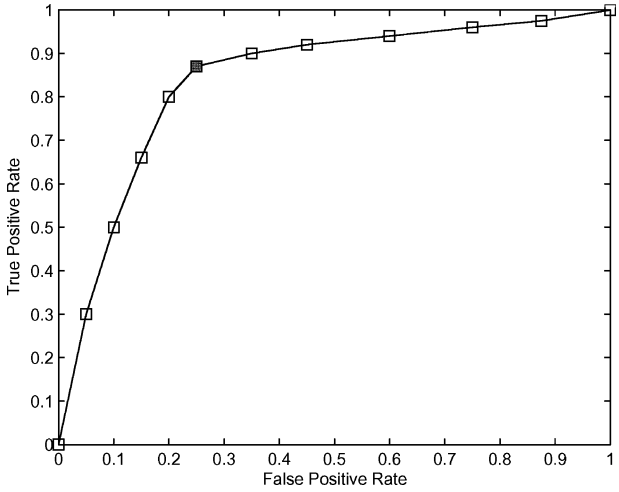


Fig. 1. Artificially generated ROC curve. Either the area under this curve or the critical point (the filled square point) achieving the highest difference between the true positive rate and the false positive rate can be used as the performance metric to evaluate the biased performance.

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 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).

B. Modifications

In this section, in order to generate the test models' true and false positive rates for both metrics, we need to modify the models. For the BMPM model, it is very convenient to obtain the true positive rates and false positive rates pairs by shifting β_0 from 0.0 to 1.0. For other traditional models, i.e., the NB classifier, the k-NN, and C4.5, we follow the methods proposed in [25], i.e., we adopt selective sampling, adjust the cost matrices, or adapt weights to modify them to generate the ROC curves. We introduce the modification procedures in the following.

The k-NN is one of the most common classifiers. For an input vector, the k-NN calculates the k closest vectors in the test set by using a distance measure, and labels the input vector as the most frequent class among the kNNs. Similar to [25], we use an altered distance as follows

$$\delta_{\mathbf{c}} = (1 - \tau_{\mathbf{c}}) d_E(\mathbf{z}, \mathbf{z_c})$$

where $\mathbf{z_c}$ is the closest point from the class $\mathbf{c}$ ($\mathbf{c} = \mathbf{x}$ or $\mathbf{y}$), and $d_E(\mathbf{z}, \mathbf{z_c})$ is the Euclidean distance from $\mathbf{z}$ to $\mathbf{z_c}$. By changing $\tau_{\mathbf{c}}$ from 0.0 to 1.0, a series of true positive rates and false positive rates can be generated.

In the NB classifier, a new point $\mathbf{z}$ is classified into the class $\mathbf{x}$ when the posterior probability $p(\mathbf{x}|\mathbf{z}) \geq 0.5$; otherwise, it is judged as the class $\mathbf{y}$. Here, we introduce a new parameter p_0, where $p_0 \in [0.0, 1.0]$ and change the decision criterion to $p(\mathbf{x}|\mathbf{z}) \geq p_0$. By scanning p_0 from 0.0 to 1.0, we can obtain a set of true and false positive rates.

C4.5 is a kind of decision tree, introduced by Quinlan [30]. We train the C4.5 by changing the prior probability to favor the corresponding class. The method is similar to [25].

It is again observed that, by trying to utilize intermediate factors such as the prior distributions, the thresholds, and the weights rather than controlling the accuracy directly to impose a bias, these traditional methods lack a rigorous treatment of biased classifications and, thus, it remains uncertain whether they can deal with the medical diagnosis problem systematically.

Remark: According to [5] and [24], a connection has been established among the distribution of the training data, the prior probability of each class, the costs of misclassification of each class, and the setup of the decision threshold. Changing one of these factors is equivalent to changing other factors. Thus, in the above, it is sufficient that we tailor the above traditional machine learning methods to unbalanced classifications by only using one of the approaches.

VI. Experiments

In this section, we first illustrate our model with a synthetic toy dataset. Then we apply the BMPM to two real-world medical diagnosis datasets, the breast-cancer dataset and the heart disease dataset, and compare the performance with other traditional learning models, the k-Nearest Neighbor method, the NB classifier, and the decision tree, C4.5.

A. Model Illustration With a Synthetic Toy Dataset

A synthetic toy dataset is generated by the two-dimensional Gamma distribution. Two classes of data are generated under the same Gamma distribution with the shape and scale parameter $\Gamma(5, 4)$ for the first dimension and $\Gamma(6, 3)$ for the second dimension. To illustrate the algorithm clearly, we transform the data by displacement and rotation to distinguish the two classes as illustrated in Fig. 2. We assume that the class $\mathbf{x}$, which is the more important class, is represented by filled $\square$'s (training points) and o's (test points). The other class $\mathbf{y}$, which is the less important class, is represented by $+$'s (training points) and $\times$'s (test points). The acceptance level is set to 90%. As a performance baseline, we also implement the MPM on this dataset. Several observations from Fig. 2 are noteworthy. First, the solid line/curve (the decision hyperplane of BMPM for linear/Gaussian kernel) is further from the important class $\mathbf{x}$ than the corresponding dashed line/curve (the decision hyperplane of MPM for linear/Gaussian kernel). This demonstrates that a bias is imposed in favor of the class $\mathbf{x}$. More specifically, as shown in Table I, the accuracies of the class $\mathbf{x}$, i.e., both the worst case (α) and real sensitivities (the test-set accuracies $\mathrm{TSA}_{\mathbf{x}}$), are significantly increased in BMPM when compared to those of MPM. Second, the test-set accuracies for the less important class $\mathbf{y}$, $\mathrm{TSA}_{\mathbf{y}}$, i.e., the specificities, remain at an acceptable level, i.e., 91.1% and 93.3%, for linear and Gaussian kernel, respectively, with the lower bound set to 90.0%. Third, the worst case accuracies given by $\alpha_{\mathbf{x}}$, $\alpha_{\mathbf{y}}$, and α are all smaller than the real test-set accuracies. This clearly demonstrates how the worst case probability can quantitatively control the classification accuracy with respect to future data and rigorously incorporate a bias in the medical diagnosis. Fourth, as seen in Table I, the overall test-set

TABLE I
LOWER BOUND α AND TEST-SET ACCURACY WITH BMPM AND MPM ON THE SYNTHETIC DATASET

Kernel	BMPM					MPM			
	α		Accuracy			α	Accuracy		
	$\alpha_{\mathbf{x}}$	$\alpha_{\mathbf{y}}$	$\mathrm{TSA}_{\mathbf{x}}$	$\mathrm{TSA}_{\mathbf{y}}$	TSA	α	$\mathrm{TSA}_{\mathbf{x}}$	$\mathrm{TSA}_{\mathbf{y}}$	TSA
Linear(%)	94.9 ↑	90.0	97.8 ↑	91.1	94.4	92.7	93.3	95.6	94.4
Gaussian(%)	96.9 ↑	90.0	97.8 ↑	93.3	95.6	93.1	93.3	95.6	94.4

TABLE II
COMPARISON OF MODEL PERFORMANCE BASED ON THE MS CRITERION ON THE BREAST-CANCER DATASET

Model	Sensitivity	Specificity	(Sensitivity+Specificity)/2
BMPML	0.982 ± 0.002	0.972 ± 0.004	**0.977 ± 0.003**
BMPMG	0.9916 ± 0.003	0.962 ± 0.001	**0.977 ± 0.002**
k-NN(5)	0.980 ± 0.005	0.967 ± 0.004	**0.973** ± 0.004
k-NN(7)	0.978 ± 0.007	0.964 ± 0.007	0.971 ± 0.006
k-NN(3)	0.976 ± 0.008	0.961 ± 0.006	0.968 ± 0.006
NB	0.987 ± 0.005	0.969 ± 0.007	**0.978** ± 0.005
C4.5	0.968 ± 0.008	0.936 ± 0.006	0.952 ± 0.007

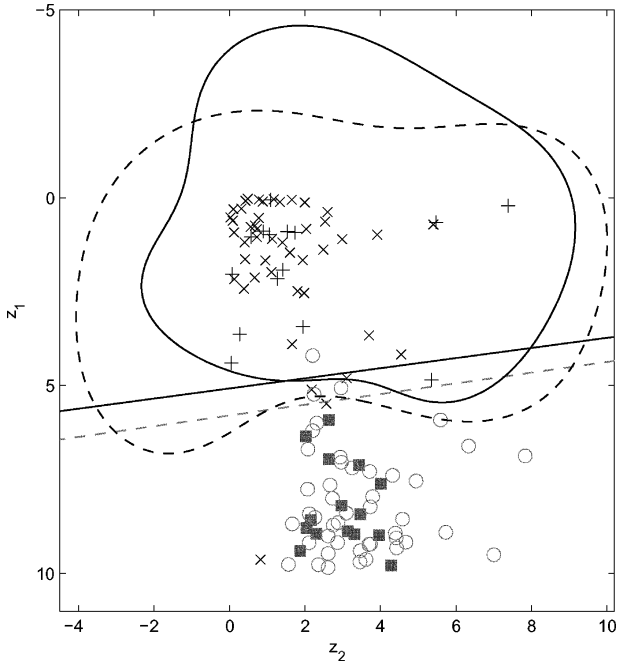


Fig. 2. Example to illustrate the BMPM. The solid red line is the decision hyperplane for the BMOML while the dashed line is the decision hyperplane for the linear MPM. The solid black curve is the decision hyperplane for the Gaussian kernel BMPM, while the dashed curve is the decision hyperplane for the Gaussian kernel MPM. Training points are indicated with filled □'s for the class $\mathbf{x}$ and +'s for the class $\mathbf{y}$. Test points are indicated with o's for the class $\mathbf{x}$ and ×'s for the class $\mathbf{y}$. The parameter σ for the Gaussian kernel is found by cross validation. The solid line and the solid curve are pushed away from the favored class $\mathbf{x}$. The results show a qualitative accuracy indicator $\alpha_{\mathbf{x}} = 94.9\%$ and $\alpha_{\mathbf{x}} = 96.9\%$ for the BMPM.

accuracies, i.e., TSA, of BMPM are not necessarily lower than those of MPM. An interesting interpretation can be seen in [13].

B. Evaluation on Real Medical Datasets

After demonstrating the BMPM algorithm with linear and Gaussian kernels on synthetic data, we apply it to two real medical datasets. Two medical datasets, the breast-cancer dataset and the heart disease dataset, obtained from the UCI machine learning repository [4], are used in this experiment. 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. The heart disease dataset includes 120 instances with heart disease and 150 instances without heart disease. Each instance is described by 13 attributes. Since handling the missing attribute values is out of the scope of this paper, we simply remove any instances with missing attribute values in the datasets. For these two datasets, the preferred class $\mathbf{x}$ 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 $\mathbf{x}$, and the specificity is the accuracy of the class $\mathbf{y}$.

We use tenfold cross validation [17] to evaluate the performance of different learning algorithms. We compare our BMPM including the BMPML and the Gaussian kernel BMPM (BMPMG), with three other approaches: NB, C4.5, and k-NN. In the BMPMG, the parameter σ in Gaussian kernel, $e^{-\|\mathbf{z}_i - \mathbf{z}_j\|^2/\sigma}$, is obtained by cross validation. In the k-NN, k is set to an odd number from 1 to 19; only the best three results are shown for brevity.

1) 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 II. It can be seen that the BMPML, BMPMG, NB, and 5-NN achieves the best performance. Although NB is slightly higher than the BMPM, a significance analysis according to the traditional analysis of variance (ANOVA) shows that difference of the means of BMPML, BMPMG, NB, and k-NN, for $k = 5$ are insignificant $(p < 0.05)$. In addition, a further ANOVA analysis shows that the means of BMPML, BMPMG, C4.5, 3-NN, and 7-NN are significantly different $(p < 0.05)$.

The results of the heart disease dataset are shown in Table III. In this dataset, the BMPM model demonstrates a superiority to the other three models. The BMPML and BMPMG achieves the best results of 0.851 and 0.838. They are both greater than 0.826, the best result of other learning algorithms given by NB. Furthermore, the ANOVA test shows that the difference of the

TABLE III
COMPARISON OF MODEL PERFORMANCE BASED ON THE MS CRITERION ON THE HEART DISEASE DATASET

Model	Sensitivity	Specificity	(Sensitivity+Specificity)/2
BMPML	0.815 ± 0.002	0.888 ± 0.004	**0.851 ± 0.003**
BMPMG	0.840 ± 0.003	0.835 ± 0.006	**0.838 ± 0.003**
k-NN(7)	0.822 ± 0.008	0.801 ± 0.009	0.815 ± 0.008
k-NN(5)	0.830 ± 0.006	0.788 ± 0.007	0.809 ± 0.007
k-NN(11)	0.850 ± 0.005	0.753 ± 0.008	0.801 ± 0.007
NB	0.813 ± 0.004	0.840 ± 0.006	0.826 ± 0.005
C4.5	0.908 ± 0.014	0.699 ± 0.005	0.804 ± 0.010

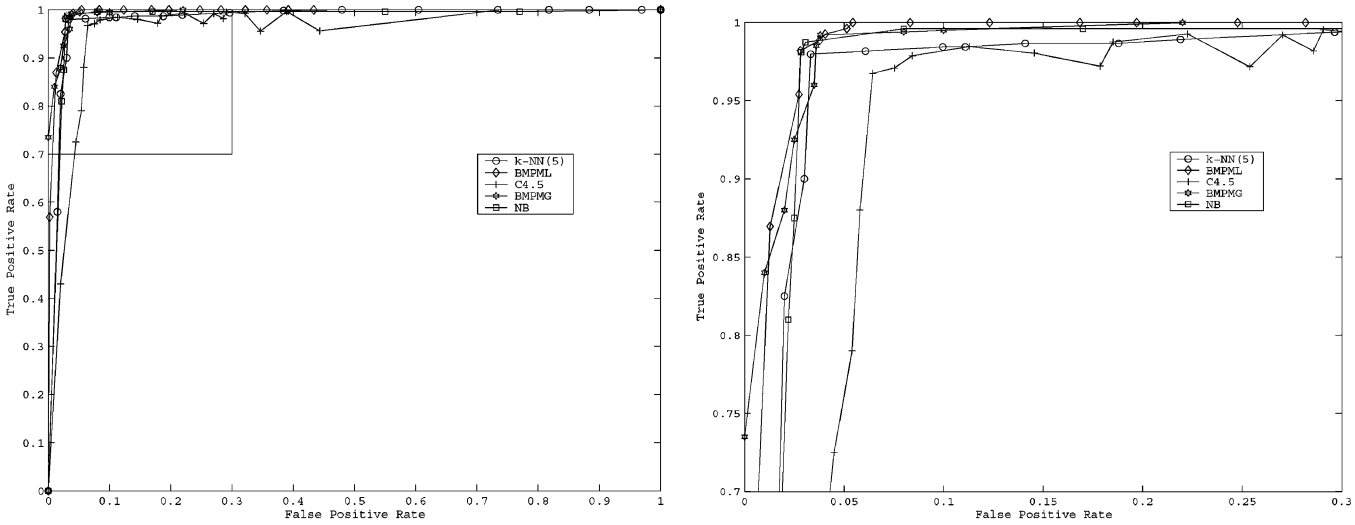


Fig. 3. ROC curves on the breast-cancer dataset. The ROC curves of the BMPML and the BMPMG are higher than those of other models and the BMPML yields the largest area under the ROC curve.

BMPML, BMPMG, and the other algorithms is significant ($p < 0.05$).

In summary, in terms of the MS criterion, our BMPM model demonstrates better performance when compared with other algorithms in both the breast cancer and heart disease datasets.

2) ROC Curve Analysis: We now compare our BMPM model with the NB, k-NN, and C4.5 in terms of the ROC curve analysis. We generate the ROC curves as illustrated in Fig. 3(a) and Fig. 4(a). It is observed that the BMPML and BMPMG perform better than other classifiers for both datasets, since at most points, the BMPM curves are above those of other methods. More specifically, we calculate the areas under the ROC curves as illustrated in Table IV. For the breast-cancer dataset, it produces a curve with an area of 0.994 in the linear setting and a curve with an area of 0.992 in the Gaussian kernel, whereas the NB forms a curve with a smaller area equal to 0.983, the best result from the other models. For the heart disease dataset, the BMPM shows a curve with an area of 0.893 in the linear setting and a curve with an area of 0.906 in the Gaussian kernel setting. These two areas are both greater than those of the other methods.

In addition, usually not all the portions of the ROC curve are of great interest [26]. Generally, those with a small false positive rate and a high true positive rate are most important [35]. In Fig. 3(b) and Fig. 4(b) we show the critical portion of Figs. 3(a) and 4(a), respectively, when the false positive rate is in the range of 0.0 to 0.3 and the true positive rate is in the range of 0.7 to 1.0. In this critical region, most parts of the ROC curves of the BMPM are above the corresponding curves of other models in both datasets, which again demonstrates the superiority of the BMPM model.

To judge whether these results are significant, we follow [25] and conduct an analysis using LabMRMC [8]. This uses the Jackknife method [10] to account for case-sample variance and then applies traditional ANOVA to determine significance. This analysis shows that the means of BMPM, C4.5, NB, k-NN, for $k = 11$, are significantly different ($p < 0.05$).

Remark: Note that we do not compare BMPM and MPM in the above. Due to the balanced nature of MPM, it cannot easily generate an ROC curve. Moreover, the sensitivity and specificity output by the MPM model has been incorporated in the ROC curve: Its result corresponds to a certain point in the ROC curve, where the worst case sensitivity and the worst case specificity is equal. Therefore, the BMPM will usually be better than MPM.

VII. OPEN PROBLEMS AND FUTURE WORK

Two main problems related to the BMPM are worth discussion. First, although we propose efficient algorithms to solve the BMPM optimization problems, one interesting question for both MPM and BMPM is whether any techniques can be used to speed up the training process, especially for the kernelized models. More specifically, can some decomposable techniques

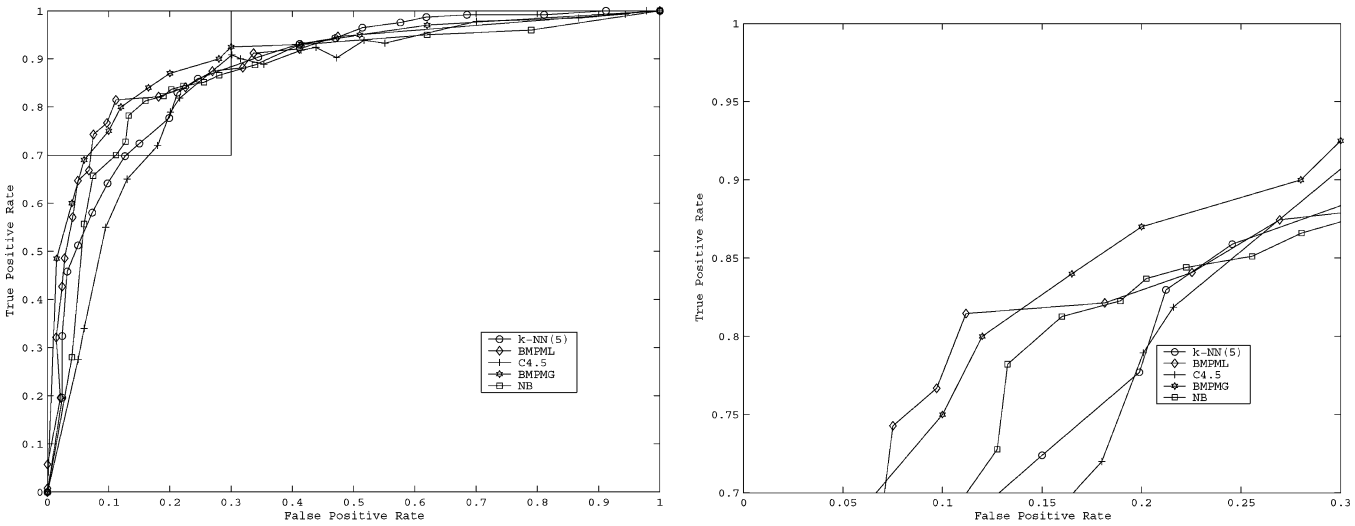


Fig. 4. ROC curves on the heart disease dataset. The ROC curves of the BMPML and the BMPMG are higher than those of other models and the BMPMG yields the largest area under the ROC curve.

TABLE IV
COMPARISON OF MODEL PERFORMANCE BASED ON THE ROC ANALYSIS

breast-cancer		heart	
Model	ROC Area	Model	ROC Area
BMPML	$\mathbf{0.994 \pm 0.002}$	BMPML	$\mathbf{0.893 \pm 0.005}$
BMPMG	$\mathbf{0.992 \pm 0.002}$	BMPMG	$\mathbf{0.906 \pm 0.006}$
k-NN(5)	0.963 ± 0.008	k-NN(5)	0.881 ± 0.004
k-NN(7)	0.959 ± 0.007	k-NN(7)	0.876 ± 0.006
k-NN(3)	0.956 ± 0.012	k-NN(9)	0.868 ± 0.005
NB	0.983 ± 0.006	NB	0.867 ± 0.005
C4.5	0.957 ± 0.016	C4.5	0.849 ± 0.007

be applied in the kernel matrix and, thus, speed up the least-squares training? Another problem in training BMPM with kernels, e.g., the Gaussian kernel, is that the parameter, σ, has to be determined via time-consuming cross validation. Speeding up these processes remains one of the open problems for both the MPM and the BMPM models.

Second, to assure a tight lower bound of the accuracy, both MPM and BMPM require that the mean and covariance matrices estimated from the dataset can reliably represent the true mean and covariance matrices. It has been empirically verified that direct plug-in estimation achieves satisfactory performance in many real classification tasks [13]. However, there exist cases, where the estimation will be inaccurate and cause problems, i.e., the worst case accuracy does not represent the lower limit of the real test-set accuracy [13]. To attack this problem, some robust estimation techniques need to be applied. For example, a specific uncertainty model is proposed in [20] to correct the plug-in estimations. However, seeking more robust estimation based on general uncertainty models remains an open problem and is, therefore, one of our planned research topics for the future.

VIII. CONCLUSION

In this paper, we address the problem of biased classification needed with the medical data and present a novel learning tool, the BMPM, for the medical diagnosis. In contrast to the traditional methods, the BMPM does not adopt an indirect approach, but directly controls the worst case classification accuracy in order to impose a certain bias in favor of the important class. This provides a more elegant way to handle biased classifications. Specifically, the BMPM is able to maximize the worst case sensitivity while maintaining the specificity within a lower bound. Importantly, when certain distributions, e.g., a Gaussian distribution, are assumed for the data, the BMPM maximizes the real sensitivity, which is the goal of medical diagnosis. We evaluate the performance of the BMPM based on the ROC analysis and the MS criterion and compare it with three traditional classifiers: the k-Nearest Neighbor, the NB, and the C4.5. The results on two medical datasets, the breast-cancer dataset and the heart disease dataset, both show that the BMPM outperforms the other three models.

APPENDIX

We append the proof of Lemma 1 in this section. First, we present the following corollary obtained from [20].

Corollary 1: Given $\mathbf{a} \neq \mathbf{0}$, b such that $\mathbf{a}^T\mathbf{y} \leq b$ and $\beta \in [0, 1)$, the condition

$$\inf_{\mathbf{y} \sim (\bar{\mathbf{y}}, \Sigma_{\mathbf{y}})} P_r\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \beta$$

holds if and only if $b - \mathbf{a}^T\bar{\mathbf{y}} \geq \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}$ with $\kappa(\beta) = \sqrt{\beta/(1-\beta)}$.

Proof of Lemma 1: Lemma 1(1) can easily be proved by using Corollary 1 directly. We now prove Lemma 1(2). From (5) and (6), we get

$$\mathbf{a}^T\bar{\mathbf{y}} + \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \leq b \leq \mathbf{a}^T\bar{\mathbf{x}} - \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}. \quad (17)$$

If we eliminate b from this inequality, we obtain

$$\mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}}) \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} + \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}. \quad (18)$$

We observe that the magnitude of $\mathbf{a}$ will not influence the solution of (18). Without loss of generality, we can set $\mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}}) = 1$. In addition, since $\kappa(\alpha)$ increases monotonically with α, maximizing α is equivalent to maximizing $\kappa(\alpha)$. Thus, the problem can further be modified to

$$\begin{aligned} \max_{\alpha,\beta,\mathbf{a}\neq\mathbf{0}} \quad & \kappa(\alpha) \\ \text{s.t.} \quad & 1 \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} + \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \quad (19) \\ & \mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}}) = 1 \quad (20) \\ & \kappa(\beta) \geq \kappa(\beta_0) \quad (21) \end{aligned}$$

where (21) is equivalent to (7) due to the monotonic property of the κ function.

In the above, the maximum value of $\kappa(\alpha)$ under the constraints of (19)–(21) is achieved when the right-hand side of (19) is strictly equal to 1; otherwise, assuming the maximum is achieved when $1 > \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} + \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}$, 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.

Moreover, $\mathbf{\Sigma_x}$ and $\mathbf{\Sigma_y}$ can be assumed as positive definite matrices; otherwise, we can always add a small positive amount to the diagonal elements of these two matrices and make them positive definite. Therefore, we obtain $\kappa(\alpha) = (1 - \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}})/\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}$. Obviously, this optimization function is a linear function with respect to $\kappa(\beta)$ and $\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}$ is a positive term; therefore, this optimization function is maximized when $\kappa(\beta)$ is set to its lower bound $\kappa(\beta_0)$. Finally, the optimization problem is easily verified to be the FP problem. □

ACKNOWLEDGMENT

The authors thank G. R. G. Lanckriet for providing the Matlab source code of the MPM on the web. They also extend their thanks to M. A. Maloof for his useful suggestions on the generation of the ROC curves.

REFERENCES

[1] D. Aha, D. Kibler, and M. Albert, Instance-Based Learning Algorithms vol. 6, pp. 37–66, 1991.

[2] R. Bairagi and C. M. Suchindram, "An estimator of the cutoff point maximizing sum of sensitivity and specificity," *Sankhya, Ser. B, Indian J. Statist.*, vol. 51, pp. 263–269, 1989.

[3] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999.

[4] C. L. Blake and C. J. Merz, UCI Repository of Machine Learning Databases Univ. California, Irvine, Dept. Inf. Comput. Sci., 1998 [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[5] L. Breiman, Arcing Classifiers Statist. Dept., Univ. California, Tech. Rep. 460, 1997.

[6] C. Cardie and N. Howe, "Improving minority class prediction using case specific feature weights," Proc. 14th Int. Conf. Machine Learning (ICML-1997) pp. 57–65. San Francisco, CA, 1997.

[7] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "Smote: sythetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[8] D. Dorfman, K. Berbaum, and C. Metz, "Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method," *Invest. Radiol.*, vol. 27, pp. 723–731, 1992.

[9] J. W. Grzymala-Busse, L. K. Goodwin, and X. Zhang, "Increasing sensitivity of preterm birth by changing rule strengths," *Pattern Recognit. Lett.*, vol. 24, pp. 903–910, 2003.

[10] D. Hinkley, S. Kotz, N. Johnson, and C. Read, Eds., "Jackknife methods," *Encyclopedia of Statistical Sciences*, vol. 4, pp. 280–287, 1983.

[11] K. Huang, I. King, and M. R. Lyu, "Discriminative training of bayesian chow-liu tree multinet classifiers," in *Proc. Int. Joint Conf. Neural Network (IJCNN-2003)*, Portland, OR, 2003, vol. 1, pp. 484–488.

[12] K. Huang, I. King, M. R. Lyu, and H. Yang, "Improving Chow-Liu tree performance based on association rules," in *Neural Information Processing: Research and Development*. Berlin, Germany: Springer-Verlag, 2004, pp. 94–112.

[13] K. Huang, H. Yang, I. King, M. R. Lyu, and L. Chan, "The minimum error minimax probability machine," *J. Mach. Learn. Res.*, vol. 5, pp. 1253–1286, 2004.

[14] K. Huang, H. Yang, I. King, M. R. Lyud, and L. W. Chan, "Biased minimax probability machine for medical diagnosis," in *Proc. 8th Int. Symp. Artificial Intelligence and Mathematics (AMAI-2004)*, 2004.

[15] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Advances in Neural Information Processing Systems (NIPS 11)*. Cambridge, MA: MIT Press, 1998.

[16] M. I. Jordan, Why the Logistic Function? A Tutorial Discussion on Probabilities and Neural Networks Tech. Rep. 9503, 1995, MIT Computational Cognitive Science Report.

[17] R. Kohavi, "A study of cross validation and bootstrap for accuracy estimation and model selection," Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-1995) pp. 338–345. San Francisco, CA, Morgan Kaufmann, 1995.

[18] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, pp. 89–109, 2001.

[19] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," Proc. 14th Int. Conf. Machine Learning (ICML-1997) pp. 179–186. San Francisco, CA, 1997.

[20] G. R. G. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan, "A robust minimax approach to classification," *J. Mach. Learn. Res.*, vol. 3, pp. 555–582, 2002.

[21] P. Langley, W. Iba, and K. Thompson, "An analysis of Bayesian classifiers," in *Proc. National Conf. Artificial Intelligence (AAAI-1992)*, 1992, pp. 223–228.

[22] C. Ling and C. Li, "Data mining for direct marketing: problems and solutions," in *Proc. 4th Int. Conf. Knowledge Discovery and Data Mining (KDD-1998)*, 1998, pp. 73–79.

[23] M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra Applicat.*, vol. 284, pp. 193–228, 1998.

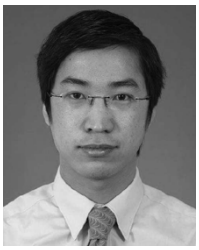
[24] M. A. Maloof, "On machine learning, ROC analysis, and statistical tests of significance," in *Proc. 16th Int. Conf. Pattern Recognition (ICPR-2002)*, 2002, pp. 204–207.

[25] M. A. Maloof, P. Langley, T. O. Binford, R. Nevatia, and S. Sage, "Improved rooftop detection in aerial images with machine learning," *Mach. Learn.*, vol. 53, pp. 157–191, 2003.

[26] D. Mcclish, "Analyzing a portion of the roc curve," *Med. Decision Making*, vol. 9, pp. 190–195, 1989.

[27] Y. Nesterov and A. Nemirovsky, *Interior Point Polynomial Methods in Convex Programming: Theory and Applications*. Philadelphia, PA: SIAM, 1994, Studies in Applied Mathematics.

[28] F. Provost, "Machine learning from imbalanced data sets 101," in *Proc. 17th Nat. Conf. Artificial Intelligence (AAAI)' Workshop on Imbalanced Data Sets*, 2000.
[29] F. Provost and T. Fawcett, "Analysis and visulization of classifier performance: comparison under imprecise class and cost distributions," in *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-1997)*, 1997, pp. 43–48.
[30] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.
[31] S. Schaible, *Fractional Programming. Nonconvex Optimization and Its Applications*. Dordrecht, The Netherlands: Kluwer Academic, 1995.
[32] B. Scholkopf and A. Smola, *Learning With Kernels*. Cambridge, MA: MIT Press, 2002.
[33] J. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, pp. 1285–1293, 1988.
[34] D. West and V. West, "Model selection for a medical diagnostic decision support system: a breast cancer detection case," *Artif. Intell. Med.*, vol. 20, no. 3, pp. 183–204, 2000.
[35] K. Woods, W. P. Kegelmeyer, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Tans. Pattern Anal. Mach. Intell.*, vol. 19, no. 4, pp. 405–410, Apr. 1997.

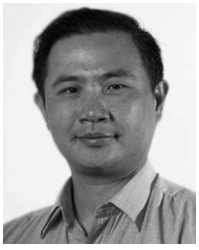


Kaizhu Huang received the B.E. degree in automation from Xi'an Jiaotong University, Xi'an, China, in 1997, the M.E. degree in pattern recognition and intelligent systems from Institute of Automation, the Chinese Academy of Sciences, Beijing, China, in 2000, and the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2004.

He is currently a Researcher in the Information Technology Laboratory, Fujitsu Research and Development Center Co., Ltd. His research interests include machine learning, pattern recognition, image processing, and information retrieval.

Haiqin Yang received the B.S. degree in computer science and technology from Nanjing University, Nanjing, China, in 2001 and the M.Phil. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2003.

He currently works in face recognition and related techniques with Titanium Technology Limited, Shenzhen, China. His research interests include machine learning, pattern recognition, and financial time series analysis.



Irwin King (S'91–M'93) received the B.Sc. degree in engineering and applied science from California Institute of Technology, Pasadena, in 1984, and the M.Sc. and Ph.D. degrees in computer science from the University of Southern California, Los Angeles, in 1988 and 1993 respectively.

He joined the Chinese University of Hong Kong, Hong Kong, in 1993. His research interests include content-based retrieval methods for multimedia databases, distributed multimedia information retrieval in peer-to-peer systems, and statistical learning theory.

He is a member of ACM, IEEE Computer Society, International Neural Network Society (INNS), and Asian Pacific Neural Network Assembly (APNNA). He is also a governing board member of the Asian Pacific Neural Network Assembly (APNNA). He is a founding member of the Intelligence Data Engineering and Learning Laboratory (IDEAL) and the Multimedia Information Processing Laboratory (MIP Lab). He is a member of the Editorial Board of the *Neural Information Processing-Letters and Reviews* journal (NIP-LR). He has served as program and/or organizing member in international conferences and workshops, e.g., WWW, ICASSP, IJCANN, ICONIP, etc. Currently, he is the Program Co-Chair of ICONIP2006. He has also served as reviewer for international conferences as well as journals, e.g., Information Fusion, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, SIGMOD, IEEE TRANSACTIONS ON NEURAL NETWORKS, IEEE PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON SYSTEM, MAN, AND CYBERNETICS, etc. He also served in the Neural Network Technical Committee (NNTC) under the IEEE Computational Intelligence Society (formerly the IEEE Neural Network Society).

Michael R. Lyu (S'84–M'88–SM'97–F'04) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1981, the M.S. degree in computer engineering from University of California, Santa Barbara, in 1985, and the Ph.D. degree in computer science from University of California, Los Angeles, in 1988.

He is currently a Professor in the Department of Computer Science and Engineering, the Chinese University of Hong Kong, Hong Kong. He is also Director of the Video over InternEt and Wireless (VIEW) Technologies Laboratory. He was with the Jet Propulsion Laboratory as a Technical Staff Member from 1988 to 1990. From 1990 to 1992, he was with the Department of Electrical and Computer Engineering, The University of Iowa, Iowa City, as an Assistant Professor. From 1992 to 1995, he was a Member of the Technical Staff in the applied research area of Bell Communications Research (Bellcore), Morristown, NJ. From 1995 to 1997, he was a Research Member of the Technical Staff at Bell Laboratories, Murray Hill, NJ. His research interests include software reliability engineering, distributed systems, fault-tolerant computing, mobile networks, Web technologies, multimedia information processing, and E-commerce systems. He has published over 200 refereed journal and conference papers in these areas. He has participated in more than 30 industrial projects, and helped to develop many commercial systems and software tools. He was the editor of two book volumes: *Software Fault Tolerance* (Wiley, 1995) and *The Handbook of Software Reliability Engineering* (IEEE and McGraw-Hill, 1996).

Dr. Lyu received Best Paper Awards in ISSRE'98 and ISSRE'2003. He initiated the First International Symposium on Software Reliability Engineering (ISSRE) in 1990. He was the Program Chair for ISSRE'96 and General Chair for ISSRE'2001. He was also PRDC'99 Program Co-Chair, WWW10 Program Co-Chair, SRDS'2005 Program Co-Chair, and PRDC'2005 General Co-Chair, and served in program committees for many other conferences including HASE, ICECCS, ISIT, FTCS, DSN, ICDSN, EUROMICRO, APSEC, PRDC, PSAM, ICCCN, ISESE, and WI. He has been frequently invited as a keynote or tutorial speaker to conferences and workshops in U.S., Europe, and Asia. He served on the Editorial Board of IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, and has been an Associate Editor of IEEE TRANSACTIONS ON RELIABILITY and the *Journal of Information Science and Engineering*.