Correspondence

Imbalanced Learning With a Biased Minimax Probability Machine

Kaizhu Huang, Haiqin Yang, Irwin King, and Michael R. Lyu

***Abstract*—Imbalanced learning is a challenged task in machine learning. In this context, the data associated with one class are far fewer than those associated with the other class. Traditional machine learning methods seeking classification accuracy over a full range of instances are not suitable to deal with this problem, since they tend to classify all the data into a majority class, usually the less important class. In this correspondence, the authors describe a new approach named the biased minimax probability machine (BMPM) to deal with the problem of imbalanced learning. This BMPM model is demonstrated to provide an elegant and systematic way for imbalanced learning. More specifically, by controlling the accuracy of the majority class under all possible choices of class-conditional densities with a given mean and covariance matrix, this model can quantitatively and systematically incorporate a bias for the minority class. By establishing an explicit connection between the classification accuracy and the bias, this approach distinguishes itself from the many current imbalanced-learning methods; these methods often impose a certain bias on the minority data by adapting intermediate factors via the trial-and-error procedure. The authors detail the theoretical foundation, prove its solvability, propose an efficient optimization algorithm, and perform a series of experiments to evaluate the novel model. The comparison with other competitive methods demonstrates the effectiveness of this new model.**

***Index Terms*—Fractional programming (FP), imbalanced learning, receiver operating characteristic (ROC) analysis, worst case accuracy.**

I. Introduction

The problem of imbalanced learning, in which nearly all the instances are labeled as one class while much fewer instances are labeled as the other class, usually the more important class, presents a challenge to the community of machine learning. Traditional classifiers seeking classification accuracy over a full range of instances are not suitable to deal with imbalanced-learning tasks, since they tend to classify all the data into the majority class, which is usually the less important class.

In the machine learning literature, there have been several proposals for dealing with the problem of imbalanced learning, which includes: the methods of moving the decision thresholds [29], [33], the methods of adjusting the costs or weights [6], [29], and the methods of sampling [7], [22], [27]. The first school of methods tries to adapt the decision threshold to impose a bias on the minority class. Similarly, the second school of methods improves the prediction performance by adjusting the costs or weights for each class. The third school of methods aims to reduce the data imbalance by downsampling (removing) instances from the majority class or upsampling (duplicating) the training instances from the minority class. A common problem for all the three families of methods is that they lack a rigorous and systematic treatment on imbalanced data. For the methods of adjusting the costs or weights, in order to impose a suitable bias, they have to adapt these factors by trials or, in particular, by cross validations [20]. Hence, it is hard for them to build direct connections between the intermediate factors (e.g., the costs or weights) and the biased-classification accuracy quantitatively. Therefore, these methods cannot rigorously handle imbalanced data. For the sampling method, the problem is that upsampling may introduce excessive weight on the noise data, while downsampling the data probably may lose some critical data points. To solve this problem, Chawla *et al.* proposed synthetic minority over sampling technique (SMOTE) to introduce minority data points and remove redundant majority points "intelligently" [7]. This method is considered as one of the state-of-the-art approach for imbalanced learning [44].

We propose a novel model named the biased minimax probability machine (BMPM) [16], different from the aforementioned approaches, to handle the tasks of learning from imbalanced data. When compared with the sampling methods, the BMPM does not remove or duplicate data. When compared with the methods of changing the thresholds or weights, our model establishes an explicit connection between the classification accuracy and the bias. It thus offers an elegant way to incorporate a certain bias into the classification by directly controlling the classification accuracy. Furthermore, the experiments show that the BMPM method outperforms the first and the second school of methods, and demonstrates the competitive performance against the state-of-the-art method SMOTE.

The rest of this correspondence is organized as follows. In the next section, we introduce the theoretical foundation of this correspondence, namely the BMPM model including the model definition, the solvability, and the techniques to incorporate distributional information. We then apply, in Section III, the BMPM model to deal with the imbalanced-learning tasks. After that, we evaluate the BMPM model based on a series of experiments. In Section V, we discuss some issues and present future work. Finally, we conclude this correspondence in Section VI.

II. Biased Minimax Probability Machine

In this section, we first introduce the model definition of the BMPM. Next, we prove the solvability of the optimization problem associated with BMPM. After that, we propose an efficient algorithm to solve this optimization problem. Finally, we make an additional analysis on the BMPM model when the distributional information for the data is available.

A. Model Definition

We only consider binary classification in this correspondence. 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 $\mathbf{x} \sim (\bar{\mathbf{x}}, \Sigma_{\mathbf{x}})$; and, 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 correspondence, class $\mathbf{x}$ also represents the

Manuscript received August 3, 2004; revised June 23, 2005. This work was supported by grants from the Research Grants Council of Hong Kong SAR, China, Projects CUHK4205/04E and CUHK4235/04E. This paper was recommended by Associate Editor L. Hall.

K. Huang is with the Information Technology Laboratory, Fujitsu Research and Development Center Company Ltd., Beijing 10016, China (e-mail: kzhuang@frdc.fujitsu.com).

H. Yang is with the Titanium Technology Ltd., 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, Chinese University of Hong Kong (e-mail: king@cse.cuhk.edu.hk; lyu@cse.cuhk.edu.hk).

Digital Object Identifier 10.1109/TSMCB.2006.870610

important or minority class, and class $\mathbf{y}$ represents the corresponding less important or majority class.

A decision hyperplane $f(\mathbf{z}) = \mathbf{a}^T\mathbf{z} - b$, where $\mathbf{a} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ and $b \in \mathbb{R}$, is constructed as follows. We try to classify each minority case into a corresponding class $(f(\mathbf{z}) \geq 0)$ with a maximum probability, while to classify the majority case $(f(\mathbf{z}) < 0)$ into a corresponding class with an acceptable accuracy. Since, normally, the distributional information for the data is unavailable, we would like to achieve a decision hyperplane in the worst case scenario. The formulation is described as follows:

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

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

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

Here, α means the lower bound of the probability (accuracy) for the classification of future cases of the class $\mathbf{x}$ with respect to all distributions with the mean and covariance as $\bar{\mathbf{x}}$, $\Sigma_{\mathbf{x}}$, respectively; in other words, α is the worst case accuracy for the class $\mathbf{x}$. Similarly, β is the lower bound of the accuracy of the class $\mathbf{y}$. This optimization aims to maximize the accuracy (the probability α) for the biased class $\mathbf{x}$ while simultaneously maintaining the class $\mathbf{y}$'s accuracy at an acceptable level β_0 by setting a lower bound as (3). This model presents a critical extension of a recently proposed competitive model, the MPM [23], which only considers the balanced data and, therefore, makes α equal to β. Our optimization setting is more useful in incorporating a bias into classifications for imbalanced-learning problems. A typical example can be seen in the epidemic-disease diagnosis problem, which is usually an imbalanced-classification problem as well. The ill cases are usually much fewer than the healthy cases. However, misclassification of the ill class results in more serious consequence than misclassification of the healthy case. Thus, an unequal treatment on different classes, namely making $\alpha \neq \beta$, is obviously necessary.

B. Appealing Features

We summarize the advantages of our biased model in the following. First, this method provides a different treatment on different classes, i.e., the hyperplane $\mathbf{a}^{*T}\mathbf{z} = b^*$ given by the solution of this optimization favors the classification of the important class $\mathbf{x}$ over the less important class $\mathbf{y}$. Second, given reliable mean and covariance matrices, the derived decision hyperplane is directly associated with the two real accuracy indicators, i.e., α and β, for each class. Thus, by varying the lower bound of β, i.e., β_0 and deriving the corresponding classifier, we can quantitatively incorporate a bias into the classification. Third, by considering the worst case accuracy, this model contains a distribution-free feature. With no distribution assumption for the data, the derived hyperplane appears to be more general and valid than a large family of classifiers, namely the generative classifiers [13], [14], [19] including the naive Bayesian (NB) classifier [24]; it has to make specific distribution assumptions. Fourth, as shown shortly in Section III, the best β_0 can either be automatically searched in terms of some standard criteria, or selected by the users based on a tradeoff curve between the accuracies on different classes. Fifth, although the BMPM contains the aforementioned advantages, it does not trade them for efficiency. It is shortly shown that the optimization of BMPM can be cast as a fractional programming (FP) problem [36], [37] and, thus, can be solved efficiently. Finally, although, previously, the decision boundary derived from BMPM is given in a linear configuration, we can apply kernelization techniques to extend it to a nonlinear classification. As shown in [17], [18], and [23], the kernelization trick can be used to map the n-dimensional data points into a high-dimensional feature space, in which a linear classifier corresponds to a nonlinear hyperplane in the original space. Since the kernelization trick is the standard technique, we omit the elaboration of the kernelization and refer the interested readers to [17], [18], and [23]. In short, with these important features, BMPM appears to offer a more direct and rigorous scheme to handle biased-classification tasks, especially for the imbalanced classifications, where the importance or cost for each class is unequal.

C. Model Solvability

In the following, we propose to solve this optimization problem. First, we borrow Lemma 1 from [23].

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

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

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

This lemma can be proven by using the Lagrangian multiplier method and the following theory developed in [32]:

$$\sup_{\mathbf{y}\in\{\bar{\mathbf{y}},\Sigma_{\mathbf{y}}\}} \Pr\{\mathbf{a}^T\mathbf{y} \geq b\} = \frac{1}{1+d^2}$$
$$\text{with} \quad d^2 = \inf_{\mathbf{a}^T\mathbf{y}\geq\mathbf{b}} (\mathbf{y}-\bar{\mathbf{y}})^T\Sigma_{\mathbf{y}}^{-1}(\mathbf{y}-\bar{\mathbf{y}}). \tag{4}$$

Details about the proof can be seen in [23].

By using Lemma 1, we obtain the following transformed optimization problem:

$$\max_{\alpha,\beta,b,\mathbf{a}\neq\mathbf{0}} \quad \alpha \quad \text{s.t.} \tag{5}$$

$$-b + \mathbf{a}^T\bar{\mathbf{x}} \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} \tag{6}$$

$$b - \mathbf{a}^T\bar{\mathbf{y}} \geq \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \tag{7}$$

$$\beta \geq \beta_0 \tag{8}$$

where $\kappa(\alpha) = \sqrt{\alpha/(1-\alpha)}$, $\kappa(\beta) = \sqrt{\beta/(1-\beta)}$. The inequality of (7) is directly obtained from (2) by using Lemma 1. Similarly, by changing $\mathbf{a}^T\mathbf{x} \geq b$ to $\mathbf{a}^T(-\mathbf{x}) \leq -b$, (6) can be obtained from (1). From (6) and (7), we get

$$\mathbf{a}^T\bar{\mathbf{y}} + \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \leq b \leq \mathbf{a}^T\bar{\mathbf{x}} - \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}. \tag{9}$$

If we eliminate b from this inequality, we obtain

$$\mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}}) \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} + \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}. \tag{10}$$

We observe that the magnitude of $\mathbf{a}$ does not influence the solution of (10). 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

$$\max_{\alpha,\beta,\mathbf{a}\neq\mathbf{0}} \quad \kappa(\alpha) \quad \text{s.t.} \tag{11}$$

$$1 \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} + \kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \tag{12}$$

$$\mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}}) = 1 \tag{13}$$

$$\kappa(\beta) \geq \kappa(\beta_0) \tag{14}$$

where (14) is equivalent to (8) due to the monotonic property of the function κ.

Lemma 2: The maximum value of $\kappa(\alpha)$ under the constraints of (12)–(14) is achieved when the right-hand side of (12) is strictly equal to 1.

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

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

$$\max_{\mathbf{a}\neq\mathbf{0}} \frac{1-\kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}} \quad \text{s.t.} \quad \mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}}) = 1. \tag{15}$$

Furthermore, the aforementioned formulation (15) can be written as the so-called FP problem [36]

$$\max_{\mathbf{a}\neq\mathbf{0}} \frac{f(\mathbf{a})}{g(\mathbf{a})} \quad \text{s.t.} \quad \mathbf{a}\in A = \left\{\mathbf{a}|\mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}}) = 1\right\} \tag{16}$$

where $f(\mathbf{a}) = 1-\kappa(\beta_0)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}}$ and $g(\mathbf{a}) = \sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}$. In the following, we propose Lemma 3 to show that this FP problem is solvable.

Lemma 3: The FP problem (16) is strictly a quasi-concave problem and is, thus, solvable.

Proof: It is easy to see that domain A is a convex set on $\mathbb{R}^n$; $f(\mathbf{a})$ and $g(\mathbf{a})$ are differentiable on A. Moreover, since $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{y}}$ can be both considered as positive definite matrices, $f(\mathbf{a})$ 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 quasi-concave on A according to [36]. Therefore, every local maximum is a global maximum [36]. In other words, this FP problem is solvable. ■

D. Practical Solving Method

To solve the FP problem, there are many methods. For example, a conjugate-gradient method can solve this problem in n (the dimension of the data points) steps if the initial point is suitably assigned [3]. In each step, the computational cost to calculate the conjugate gradient is $O(n^2)$. Thus, this method has a worst case $O(n^3)$ time complexity. Adding the time cost to estimate $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$, $\Sigma_{\mathbf{x}}$, and $\Sigma_{\mathbf{y}}$, 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 [23] and the linear support vector machine [39].

In this correspondence, the Rosen gradient projection method [3] is adopted to solve the concave–convex FP problem. This method attains a local maximum with a worse case linear convergence rate [3]. More importantly, the local maximum will be exactly the global maximum in this problem.

In the previous section, we only talk about how to solve $\mathbf{a}$. We now turn to finding the optimal b. From Lemma 2, we can see that the inequalities in (9) will become equalities at the maximum point. The optimal b will thus be obtained by

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

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

E. Assuming Specific Distributions

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

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

$$\begin{aligned}\inf_{\mathbf{x}\sim\mathcal{N}(\overline{\mathbf{x}},\Sigma_{\mathbf{x}})} \Pr\{\mathbf{a}^T\mathbf{x}\geq b\} &= \Pr_{\mathbf{x}\sim\mathcal{N}(\overline{\mathbf{x}},\Sigma_{\mathbf{x}})}\{\mathbf{a}^T\mathbf{x}\geq b\}\\ &= \Pr\left\{\mathcal{N}(0,1)\geq\frac{b-\mathbf{a}^T\overline{\mathbf{x}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}}\right\}\\ &= 1-\Phi\left(\frac{b-\mathbf{a}^T\overline{\mathbf{x}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}}\right)\\ &= \Phi\left(\frac{-b+\mathbf{a}^T\overline{\mathbf{x}}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}}\right)\geq\alpha\end{aligned} \tag{17}$$

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

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

Due to the monotonous nature of $\Phi(z)$, we can further write (17) as

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

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

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

$$-b+\mathbf{a}^T\overline{\mathbf{x}}\geq\Phi^{-1}(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}} \tag{18}$$

$$b-\mathbf{a}^T\overline{\mathbf{y}}\geq\Phi^{-1}(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \tag{19}$$

$$\beta\geq\beta_0. \tag{20}$$

The aforementioned optimization is nearly the same as (5) subjected to the constraints of (6)–(8), 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.

We further provide an analysis on BMPM when other general distributions are assumed. Similarly, assuming $\mathbf{x}\sim\mathcal{S}(\overline{\mathbf{x}},\Sigma_{\mathbf{x}})$, $\mathbf{y}\sim\mathcal{S}(\overline{\mathbf{y}},\Sigma_{\mathbf{y}})$, where $\mathcal{S}$ means a specific distribution, we have

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

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

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

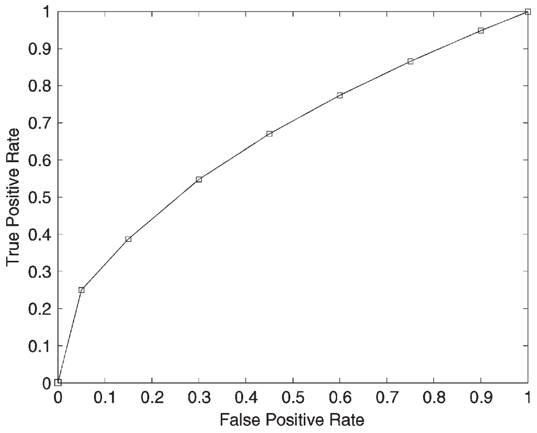


Fig. 1. Artificially generated ROC Curve.

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. This topic deserves further exploration.

III. Learning From Imbalanced Data by Using BMPM

In this section, we propose to apply the novel BMPM model into the tasks of learning from imbalanced data. We first review four standard imbalanced-learning criteria. We then, based on two of them, apply BMPM into imbalanced-learning tasks.

A. Four Criteria to Evaluate Learning From Imbalanced Data

In general, four criteria are exploited to evaluate the imbalanced learning. They are: 1) the criterion of minimum cost (MC); 2) the criterion of the maximum geometry mean (MGM) of the accuracies on the majority class and the minority class; 3) the criterion of the maximum sum (MS) of the accuracies on the majority class and the minority class; and 4) the criterion of the receiver operating characteristic (ROC) analysis. We review these criteria as follows.

Aiming to solve problems caused by maximizing the accuracy over a full range of data, instead, Grzymala-Busse *et al.* [11] maximized the sum of the accuracies on the minority class and the majority class (or maximized the difference between the true-positive and false-positive accuracies). This criterion is also widely used in other fields, e.g., graph detection, especially line detection and arc detection, where it is called vector recovery index [8], [28]. Similarly, Kubat *et al.* [21] proposed to use the geometric mean instead of the sum of the accuracies. However, compared to maximizing the sum, this criterion has a nonlinear form, which is not easy to be automatically optimized. On the other hand, when the cost of misclassification is known, an MC measure defined in (21) should be used [5]

$$\text{Cost} = F_{\text{p}} \cdot C_{F_{\text{p}}} + F_{\text{n}} \cdot C_{F_{\text{n}}} \tag{21}$$

where F_{p} is the number of the false positive, $C_{F_{\text{p}}}$ is the cost of the false positive, F_{n} is the number of the false negative, and $C_{F_{\text{n}}}$ is the cost of the false negative. However, because the cost of misclassification is generally unknown in real cases, the usage of this measure is somewhat restricted. Considering this point, some researchers introduced the ROC analysis [29], [41]. This criterion plots a so-called ROC curve to visualize the tradeoff between the false-positive rate and the true-positive rate and leaves the task of the selection of a specific tradeoff to the practitioners. Fig. 1 illustrates an artificially generated ROC curve. It has been suggested that the area beneath an ROC curve can be used as a measure of accuracy in many applications [34], [40]. Thus, a good classifier for imbalanced learning should have a larger area.

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

B. BMPM for Maximizing the Sum of the Accuracies

In the following, we first modify the formulation of BMPM to maximize the sum of the accuracy for the two classes. We then propose to solve the optimization associated with the modification version.

1) Model Modification: When applying the BMPM for the criterion of MS, we can modify the formulation of BMPM as follows:

$$\max_{\alpha,\beta,b,\mathbf{a}\neq\mathbf{0}} \quad \alpha+\beta \quad \text{s.t.} \tag{22}$$

$$\inf_{\mathbf{x}\sim\{\overline{\mathbf{x}},\Sigma_{\mathbf{x}}\}} \Pr\{\mathbf{a}^T\mathbf{x} \geq b\} \geq \alpha \tag{23}$$

$$\inf_{\mathbf{y}\sim\{\overline{\mathbf{y}},\Sigma_{\mathbf{y}}\}} \Pr\{\mathbf{a}^T\mathbf{y} \leq b\} \geq \beta. \tag{24}$$

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

2) Solving Method: Similar to the standard BMPM, we can transform the aforementioned optimization problem as follows:

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

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

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

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

Similarly, we can eliminate b and obtain the following inequality:

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

Furthermore, we transform the optimization into the following formulation by setting $\mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}})=1$:

$$\max_{\alpha,\beta,\mathbf{a}\neq\mathbf{0}} \quad \alpha+\beta \quad \text{s.t.} \tag{25}$$

$$1 \geq \kappa(\alpha)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{x}}\mathbf{a}}+\kappa(\beta)\sqrt{\mathbf{a}^T\Sigma_{\mathbf{y}}\mathbf{a}} \tag{26}$$

$$\mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}})=1. \tag{27}$$

It is easily verified that the maximum value of $\alpha+\beta$ under the constraints of (26) and (27) is achieved when the right-hand side of (26) is strictly equal to 1. Therefore, the optimization problem can be further transformed as follows:

$$\max_{\beta,\mathbf{a}\neq\mathbf{0}} \quad \frac{\kappa^2(\alpha)}{\kappa^2(\alpha)+1}+\beta \quad \text{s.t.} \quad \mathbf{a}^T(\overline{\mathbf{x}}-\overline{\mathbf{y}})=1 \tag{28}$$

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

The optimization of (28) corresponds to finding an optimal β^*, making $f(\beta^*)=\kappa^2(\alpha)/(\kappa^2(\alpha)+1)+\beta^*$ maximal. Therefore, if we

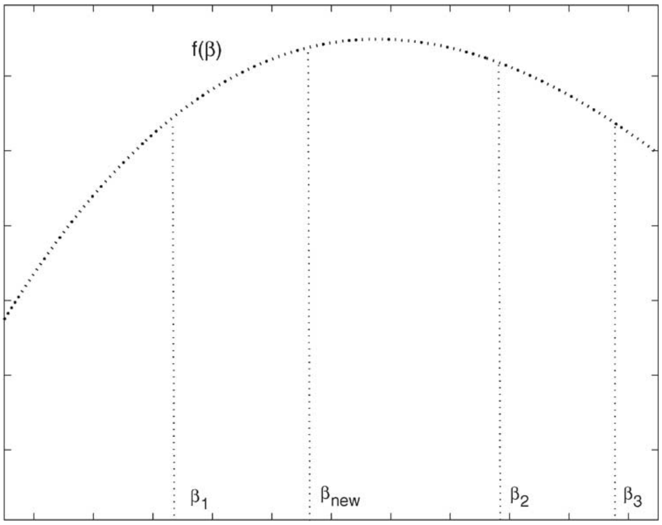


Fig. 2. Three-point pattern and quadratic line-search method. A β_{new} is obtained, and a new three-point pattern is constructed by β_{new} and two of β_1, β_2, and β_3.

fix β to a specific value within [0, 1), the optimization is equivalent to maximizing $\kappa^2(\alpha)/(\kappa^2(\alpha)+1)$, and is further equivalent to maximizing $\kappa(\alpha)$, which is exactly the BMPM problem. We then change β and repeat the BMPM optimization procedure until an optimal β^* is found, such that $f(\beta^*)$ is maximized. The aforementioned procedure is also the so-called line-search problem [3]. Many methods can be used to solve the line-search problem. In this correspondence, we use the quadratic interpolation (QI) method [3]. As illustrated in Fig. 2, QI finds the maximum point by updating a three-point pattern $(\beta_1, \beta_2, \beta_3)$ repeatedly. The new β, denoted by β_{new}, is given by the QI from the three-point pattern. Then, a new three-point pattern is constructed by β_{new} and two of β_1, β_2, and β_3. This method is shown to converge superlinearly [3].

C. BMPM for ROC Analysis

It is straightforward to apply the BMPM model to plot the ROC curve, since the lower bounds α and β directly and quantitatively control the accuracies for two classes. We only need to adapt the acceptable level for β, namely β_0, from 0 to 1, to obtain a sequence of tradeoffs between the accuracy of the important class and the negative class. We address again, since β_0 represents the lower bound of the accuracy of the less important class, and varying β_0 provides a direct and quantitative way to move the decision plane with different tradeoffs. Directly associating accuracies, with the moving of the hyperplane while assuming no distribution, is one of the advantages of BMPM over the other methods by adapting the weights or thresholds.

IV. Experimental Results

In this section, we first illustrate the BMPM model with a toy example. We then evaluate the performance of BMPM on five real-world imbalanced datasets in comparison with the SMOTE method, the NB classifier, the k-nearest neighbor (k-NN) method [1], and the decision-tree classifier C4.5 [35]. Note that the NB, the k-NN, and the C4.5 are all modified so that they can be applied to imbalanced learning.

A. Toy Example

We present a toy example to illustrate the BMPM model in this section. Suppose, 15 data points of the class $\mathbf{x}$ are generated from a two-dimensional Gaussian distribution with the mean and covariance matrix as $\overline{\mathbf{x}} = [0\ 1.5]^T$ and $\Sigma_{\mathbf{x}} = [0.5\ 0; 0\ 0.5]$, and 65 data points of the class $\mathbf{y}$ from another two-dimensional Gaussian distribution with $\overline{\mathbf{y}} = [0\ 0]^T$ and $\Sigma_{\mathbf{y}} = [0.5\ 0; 0\ 0.5]$.

By adapting the lower bound accuracy β_0 for the class $\mathbf{y}$ and then optimizing the corresponding BMPM, we obtain a series of decision boundaries for the toy example when using the Gaussian kernel $\exp[-\|\mathbf{x}-\mathbf{y}\|^2/\sigma]$ with the parameter σ as 5. These boundaries are illustrated in Fig. 3. Shaded regions are classified as the class $\mathbf{x}$ represented by crosses, whereas those outside shaded regions are judged as the class $\mathbf{y}$ plotted as squares. It is clear to observe that the lower bound β_0 directly controls the accuracy of the class $\mathbf{y}$. More specifically, when β_0 is set to small values such as 10.00%, 60.00%, and 95.00%, the boundary is biased toward the class $\mathbf{x}$. When β_0 is set to larger values such as 99.00%, the classification is biased toward the class $\mathbf{y}$. Moreover, we demonstrate in Table I that the lower bounds β_0 and α can serve as the accuracy indicators. It is observed that these lower bounds work well, i.e., the corresponding accuracy is slightly higher than the lower bounder except in the case when $\beta_0 = 95.00\%$. The exception, i.e., the value of α, which is 99.16%, is greater than the real accuracy 93.33%, is understandable due to the relatively smaller number of training samples. One single misclassification will influence the classification results significantly. This toy example demonstrates that, by changing β_0, BMPM provides an elegant and direct way to incorporate the bias into the classification.

B. Evaluations on Real-World Imbalanced Datasets

1) Modification on Learning Techniques: We first investigate and modify three learning techniques. The NB classifier [15], [24] is proposed based on a very simple assumption, i.e., each attribute is conditionally independent of each other when given the class variable. The decision in a two-category prediction task is made according to the calculation of the posterior probability $p(C|\mathbf{z})$, where C is the class variable and $\mathbf{z}$ represents the observation. When $p(C_1|\mathbf{z}) \geq 0.5$ or another equivalent yet more convenient rule is satisfied, i.e., $p(C_1)p(\mathbf{z}|C_1) \geq p(C_2)p(\mathbf{z}|C_2)$, $\mathbf{z}$ is classified into C_1; otherwise, it is judged as C_2. Even with the strong conditionally independent assumption, the NB classifier demonstrates a surprisingly good performance when compared with state-of-the-art classifiers [10], [25], such as support vector machines [42] and C4.5 in many domains. By simply introducing a parameter τ into the decision rule $p(C_1)p(\mathbf{z}|C_1) \geq \tau p(C_2)p(\mathbf{z}|C_2)$, the NB classifier can be adapted into the imbalanced learning. For example, specifying $\tau < 1$ imposes a bias toward the C_1 class, whereas specifying $\tau > 1$ imposes a bias toward the C_2 class.

In the k-NN classification [1], based on certain distance measures (e.g, the Euclidean distance measure), k data points closest to the query point are selected. The query point is then labeled as the most frequent class among the chosen k points. Although this method is very simple and may suffer from difficulties in high dimensions, it achieves satisfactory performance in many real domains. Following [29], we alter the distance measure δ_j for class C_j to handle imbalanced-learning tasks according to

$$\delta_j = d_E(\mathbf{z}, \mathbf{z}_j) - \tau_j d_E(\mathbf{z}, \mathbf{z}_j) \tag{29}$$

where $\mathbf{z}_j$ is the closest point from class C_j to the query point, and $d_{\text{E}}(\mathbf{z}, \mathbf{z}_j)$ represents the Euclidean distance measure. Similar to the NB classifier, by modifying τ_j, the NN method can build biased classifiers.

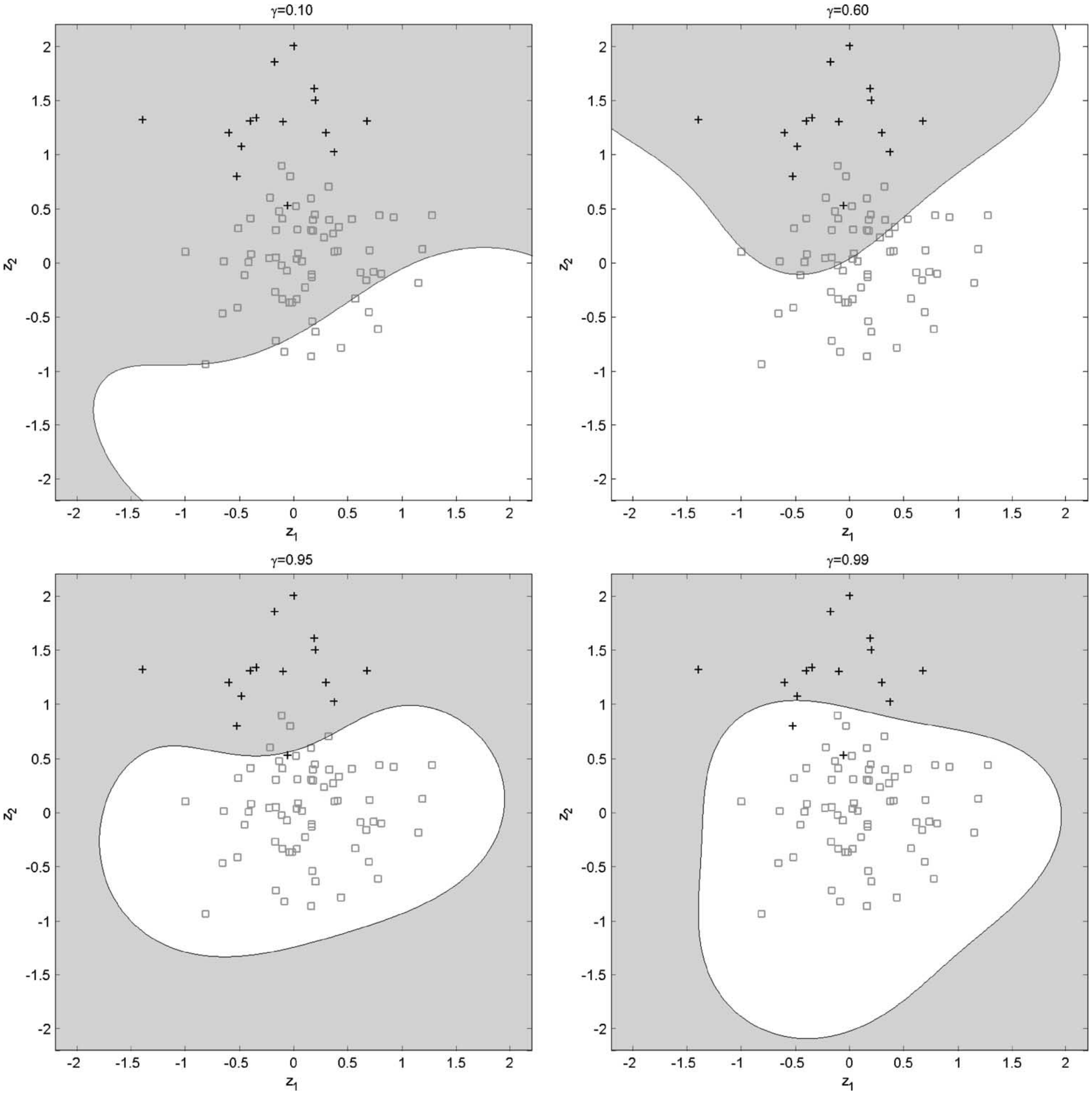


Fig. 3. Toy example to illustrate BMPM. The data of the class $\mathbf{x}$ are plotted as crosses, and the data of class $\mathbf{y}$ as squares. The shaded area represents the classification region of the class $\mathbf{x}$, while the area outside the shaded region is classified as the class $\mathbf{y}$.

TABLE I
LOWER BOUNDS OF ACCURACIES, α, β_0, AND THE REAL ACCURACIES

β_0(%)	True Negative Rate(%)	α(%)	True Positive Rate(%)
10.00	13.85	100.00	100.00
60.00	63.08	100.00	100.00
95.00	95.38	99.16	93.33
99.00	100.00	81.94	86.67

C4.5 is a kind of algorithm, introduced by Quinlan, for inducing classification models, also called decision trees, from the data [35]. By selecting the attributes according to the gain ratio criterion, an information measure of homogeneity, C4.5 builds up a decision tree, where each path from the root to a leaf represents a specific classification rule. We adapt C4.5 to learn from imbalanced dataset based on the similar method in [29], i.e., by changing the prior probability to bias the classification.

2) Evaluations on Five Real-World Datasets: We evaluate the BMPM model on five real-world datasets including audiology, glass, hepatitis, recidivism, and rooftop datasets. The first three datasets are obtained from the University of California at Irvine machine learning repository [4]. Originally, they are multiclass datasets. In order to generate imbalanced two-class data, we intentionally consider the 19th, 7th, and 1st classes as the minority or the positive class, respectively, for these datasets, while all other classes are regarded as the majority class or the negative class. The recidivism dataset was obtained from a cohort of releasees of the North Carolina prison system during the time

TABLE II
DESCRIPTION OF THE DATASETS USED IN THIS CORRESPONDENCE

Dataset	Attributes	Positive Instances#	Negative Instances#
Audiology19	69	20	206
Glass7	9	29	185
Hepatitis1	19	32	123
Recidivism	9	570	970
Rooftop	9	781	17,048

period from July 1, 1977 to June 30, 1978 [38]. The rooftop dataset consists of 17 829 overhead images of Fort Hood, TX, collected as part of the RADIUS project [9], which are of a military base. Regarding whether they are buildings (with a detected rooftop) or not, 781 images in this dataset are labeled as positive examples while 17 048 images are labeled as negative examples [26]. For audiology, glass, hepatitis, and rooftop datasets, we randomly split them into a training set with 60% data and a test set with 40% data. We then construct classifiers from the imbalanced data based on the training dataset, and perform evaluations on the test data. We repeat this procedure ten times and use the average of the results as the performance metric. For recidivism, training data containing 72.5% data and test data containing 27.5% are provided when the data are first released. For easy comparison with the literature [38], we, therefore, do not perform a hold-out evaluation. The detailed information about the datasets is described in Table II.

We compare the performance of our proposed BMPM model, in both the linear BMPM (BMPML) and the Gaussian kernel setting (BMPMG), with the SMOTE method, the modified NB classifier, the modified C4.5, and the modified k-NN method. The latter three methods are adjusted to the imbalanced learning according to the methods introduced in the previous section. For SMOTE, the best amount of SMOTE is searched from 1 to 4, and only the best result is presented for brevity. Similarly, we run k-NN methods for $k = 1, 3, 5, \ldots, 21$ and only present the best result. The width parameter for the Gaussian kernel is tuned via cross validation [20]. For C4.5, we use default parameter setting.

We first report the experimental results based on the MS criterion in Table III. To be more comparable, we show the average of the true-positive (Tp) rate and the true-negative (Tn) rate (instead of their sum) when each classifier attains the point of MS. When compared with the state-of-the-art imbalance learning method SMOTE, the BMPML and BMPMG demonstrate a competitive performance. In more details, the BMPM approach performs the best in audiology, glass, and recidivism, while SMOTE performs the best in hepatitis and rooftop. When compared with the other three methods, i.e., the modified NB, the modified k-NN, and the modified C4.5, the BMPML and BMPMG consistently demonstrate a better performance. Moreover, the t-test analysis shows that the accuracy of BMPML and BMPMG is significantly different from that of the modified NB, the modified k-NN, and the modified C4.5 at $p \leq 0.1$ in audiology, glass, hepatitis, and rooftop.

In Table III, we also report the results of MPM in the linear case (MPML) and the Gaussian kernel case (MPMG). BMPM demonstrates a better performance than MPM in both the linear case and the MPMG. Moreover, the t test shows that, except in the linear case of rooftop, the differences between BMPM and MPM (i.e., between BMPML and MPML, and between BMPMG and MPMG) are significant at $p \leq 0.1$ in audiology, glass, hepatitis, and rooftop.

Note that, in the previous section, our BMPM model directly achieves the objective of MS by maximizing the sum of the worst case accuracies for two classes. In contrast, MPM forces the equal worst case accuracies and, therefore, does not necessarily maximize the sum of accuracies. Hence, it is natural that BMPM outperforms MPM in terms of the MS criterion.

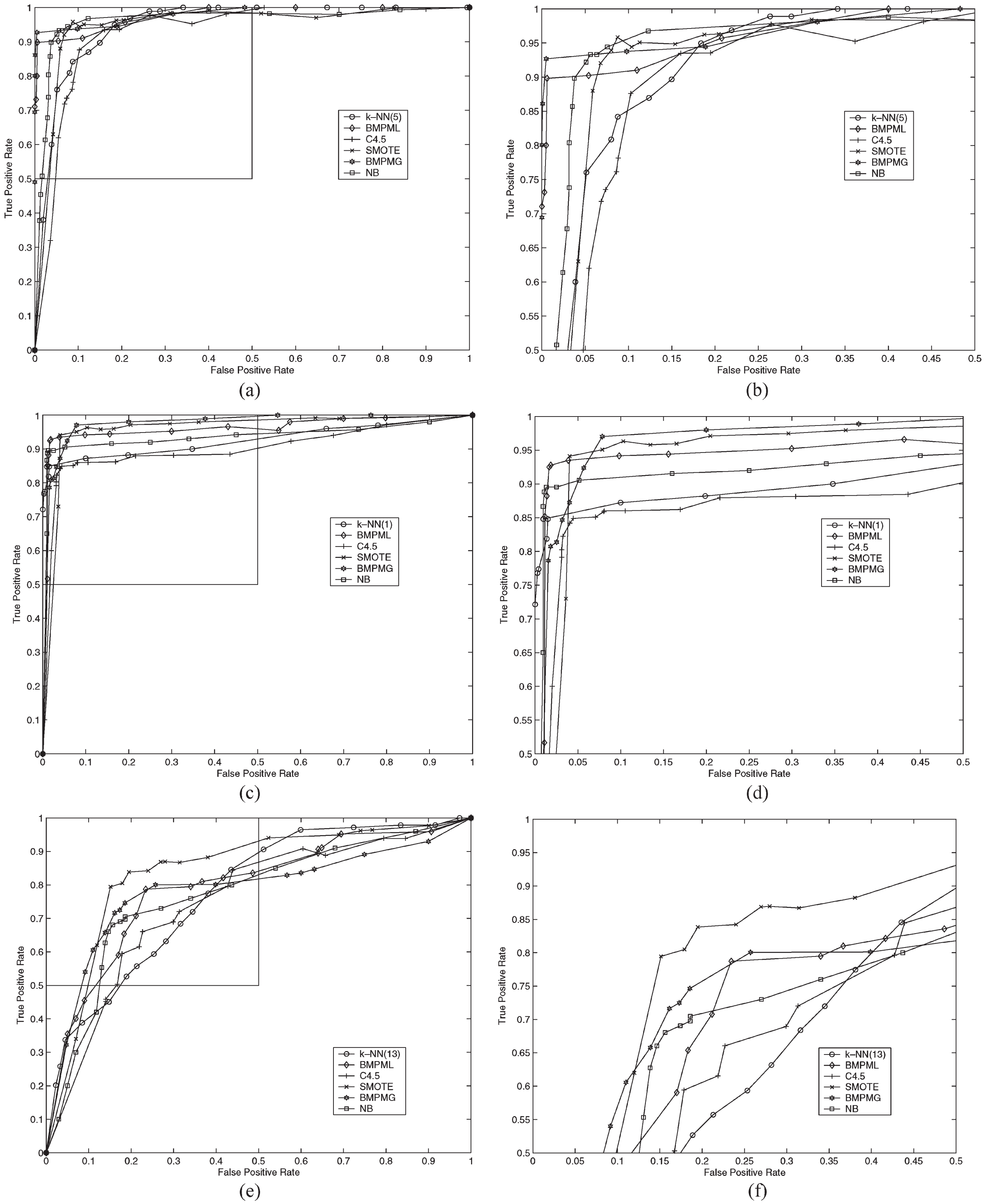
TABLE III
PERFORMANCE BASED ON THE MS CRITERION

Method	Audiology	Glass	Hepatitis	Recidivism	Rooftop
NB	93.85 ± 0.84	93.75 ± 0.70	76.21 ± 1.24	62.77	80.73 ± 0.66
KNN	91.38 ± 0.62	91.65 ± 1.04	70.68 ± 0.16	58.90	78.22 ± 0.72
C4.5	88.73 ± 1.66	90.21 ± 1.55	71.68 ± 2.43	61.53	80.59 ± 0.51
SMOTE	93.54 ± 1.37	95.04 ± 2.13	82.18 ± 2.66	60.95	85.97 ± 0.71
MPML	85.51 ± 1.56	93.43 ± 1.52	73.52 ± 1.17	63.26	80.97 ± 0.51
MPMG	90.03 ± 1.63	94.19 ± 1.45	77.17 ± 1.83	64.01	81.27 ± 0.74
BMPML	94.76 ± 1.20	95.48 ± 1.43	77.64 ± 1.85	63.91	81.23 ± 0.60
BMPMG	96.34 ± 1.41	95.73 ± 1.51	78.04 ± 1.84	64.90	82.01 ± 0.91

We now present the experimental results based on the ROC analysis. By setting the thresholds or costs by trials for NB, k-NN, and C4.5, the ROC curves are generated with good shapes as evenly distributed along their length as possible. The ROC curve of SMOTE is generated according to [7], i.e., it is created by first "smoting" the minority class and then undersampling the majority class gradually. For the BMPM model, since the lower bound β_0 serves as the accuracy indicators, we simply vary it from 0 to 1 to generate the corresponding ROC curve. The ROC curves for audiology, glass, and hepatitis are plotted in Fig. 4, and the ROC curves for recidivism and rooftop are drawn in Fig. 5. As seen in these two figures, the performance of BMPML and BMPMG is comparable with the SMOTE method: The ROC curves of BMPML and BMPMG covers that of SMOTE in audiology, glass, and recidivism in most parts, while the ROC curve of SMOTE covers those of BMPML and BMPMG in hepatitis and rooftop in most of the parts. When compared with the other three methods, BMPML and BMPMG once again demonstrate the better performance, since their ROC curves dominate those of other models in most parts. To quantitatively demonstrate the difference, we also show the areas beneath the ROC curves approximated by using the trapezoid rule in Table IV. The BMPML and BMPMG outperform SMOTE in audiology, glass, and recidivism, while their performances are not as good as that of SMOTE in hepatitis and rooftop. Furthermore, the BMPML and BMPMG show a consistent superiority to the other three models. The t test on the areas shows that the values of BMPML and BMPMG is significantly different from that of the modified NB, the modified k-NN, and the modified C4.5 at $p \leq 0.1$ in audiology, glass, hepatitis, and rooftop.

In addition, in real applications, not all portions of the ROC curve are of great interest [31]. Usually, those with a small false-positive rate and a high true-positive rate should be of more interest and importance [45]. We, thus, show the portion of the ROC curve in the range when the false-positive rate $\text{Fp} \in [0, 0.5]$ and the true-positive rate $\text{Tp} \in [0.5, 1]$. As seen in Fig. 4(b), (d), and (f) and Fig. 5(b) and (d) in this range, the superiority of the BMPML and BMPMG to the modified NB, the modified k-NN, and the modified C4.5 is more obvious than the whole ROC-curve analysis. In comparison with the SMOTE, the previous conclusion can also be drawn, i.e., the BMPML and BMPMG outperforms SMOTE in audiology, glass, and recidivism, but their performance is not as good as SMOTE's in hepatitis and rooftop.

Remark: Note that, in the previous section, we do not compare BMPM with MPM in terms of the ROC criterion. Due to the balanced nature of MPM, it cannot generate an ROC curve. Moreover, the true-positive and the true-negative outputs by the MPM model have been incorporated in the ROC curve: its result corresponds to a certain point in the ROC curve, where the worst case true positive and the worst case true negative are equal.



(a) (b) (c) (d) (e) (f)

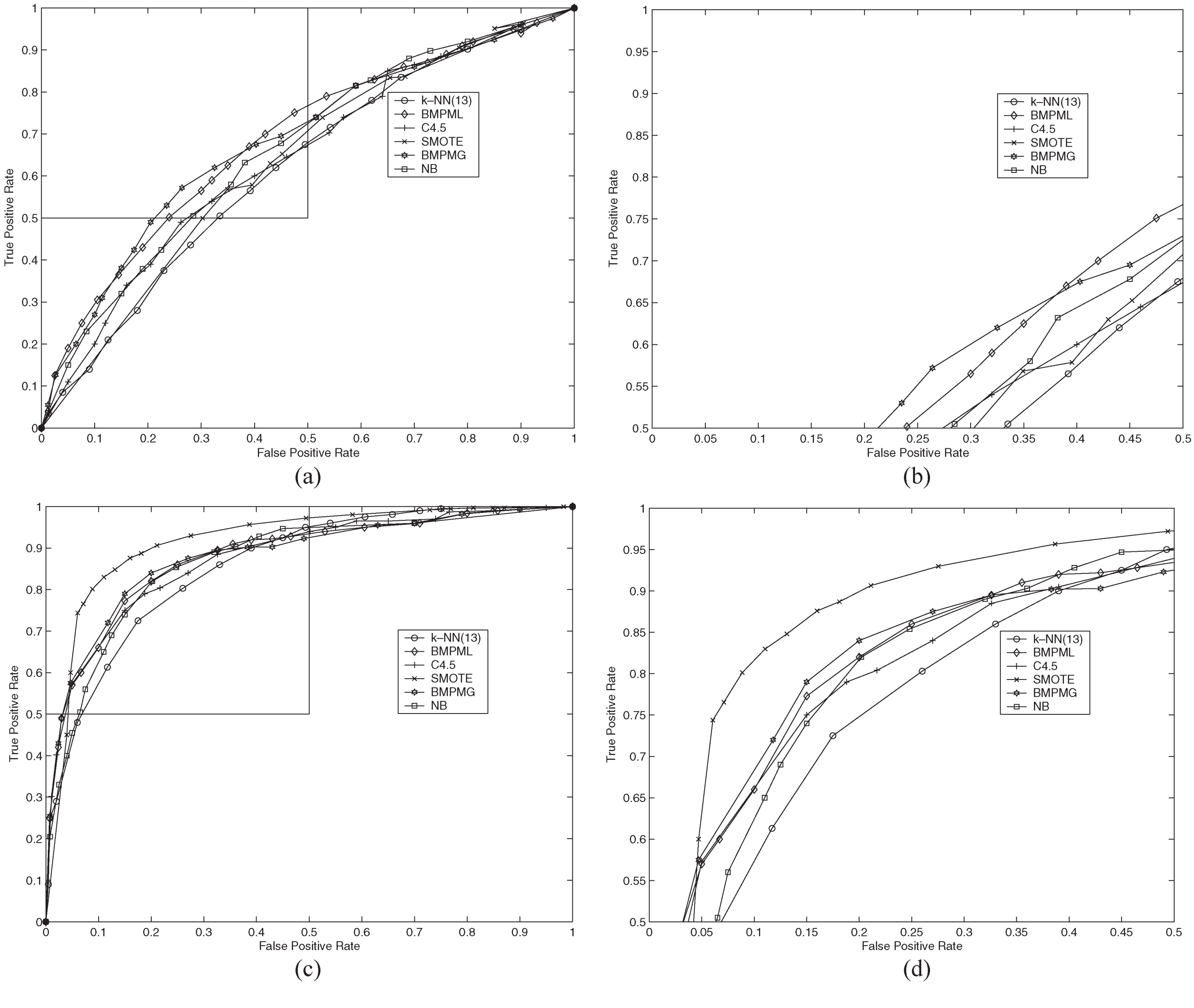
Fig. 4. ROC curves for the audiology, glass, and hepatitis datasets. (a), (c), and (e) show a full range of the ROC curves, while (b), (d), and (f) show a critical proportion of the ROC curves, which is of more interest in real applications. (a) Audiology: $0 \leq \mathrm{Tp}, \mathrm{Tn} \leq 1$. (b) Audiology: $0.5 \leq \mathrm{Tp}, \mathrm{Tn} \leq 1$. (c) Glass: $0 \leq \mathrm{Tp}, \mathrm{Tn} \leq 1$. (d) Glass: $0.5 \leq \mathrm{Tp}, \mathrm{Tn} \leq 1$. (e) Hepatitis: $0 \leq \mathrm{Tp}, \mathrm{Tn} \leq 1$. (f) Hepatitis: $0.5 \leq \mathrm{Tp}, \mathrm{Tn} \leq 1$.

V. Discussion

In this section, we first show that the BMPM model can easily be adapted when the cost for each class is known. Next, we discuss the limitations of the BMPM model and present future work.

A. *When the Cost for Each Class Is Known*

There exists cases in which the cost for each class can be given by experts. In the following, we show that the BMPM model can naturally be adapted into this type of tasks.



(a)

(b)

(c)

(d)

Fig. 5. ROC curves for recidivism and rooftop datasets. ROC curves for the audiology, glass, and hepatitis datasets. (a) and (c) show a full range of the ROC curves, while (b) and (d) show a critical proportion of the ROC curves, which is of more interest in real applications. (a) Recidivism: $0 \leq$ Tp, Tn ≤ 1. (b) Recidivism: $0.5 \leq$ Tp, Tn ≤ 1. (c) Rooftop: $0 \leq$ Tp, Tn ≤ 1. (d) Rooftop: $0.5 \leq$ Tp, Tn ≤ 1.

TABLE IV
PERFORMANCE BASED ON THE AREA OF THE ROC CURVE

Method	Audiology	Glass	Hepatitis	Recidivism	Rooftop
NB(%)	96.27 ± 0.61	93.79 ± 0.47	77.10 ± 1.13	66.46	86.78 ± 0.60
k-NN(%)	92.72 ± 0.46	92.13 ± 0.78	77.71 ± 0.87	61.89	86.01 ± 0.91
C4.5(%)	93.01 ± 1.02	90.19 ± 0.82	75.07 ± 1.34	63.83	$87.44. \pm 0.62$
SMOTE (%)	94.72 ± 0.98	95.95 ± 1.52	84.26 ± 1.46	63.29	91.75 ± 0.77
BMPML(%)	97.66 ± 0.95	96.09 ± 1.10	79.21 ± 0.83	68.42	87.91 ± 0.61
BMPMG(%)	98.14 ± 1.21	97.64 ± 1.14	78.00 ± 0.98	67.98	88.19 ± 0.87

Assuming $\mathbf{x}$ and $\mathbf{y}$ are the minority class and the majority class, respectively, it is easily verified that minimizing the optimization function given by (21) is equivalent to maximizing the following formulation:

$$\max \quad r_{\mathbf{x}} K_{\mathbf{x}} + r_{\mathbf{y}} K_{\mathbf{y}}$$

where $r_{\mathbf{x}}$ is the true-positive rate or the accuracy of the class $\mathbf{x}$, $r_{\mathbf{y}}$ is the true-negative rate or the accuracy of the class $\mathbf{y}$, and $K_{\mathbf{x}}$ and $K_{\mathbf{y}}$ are two constants, which are equal to $C_{F_{\mathrm{p}}} N_{\mathbf{y}}$ and $C_{F_{\mathrm{n}}} N_{\mathbf{x}}$, respectively ($N_{\mathbf{x}}$ and $N_{\mathbf{y}}$ are the number of data points labeled as class $\mathbf{x}$ and $\mathbf{y}$, respectively). Similar to the optimization procedure of MS, we can naturally modify the BMPM model into the following formulation:

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

The aforementioned optimization derives the classification boundary by maximizing the weighted lower bound of the real accuracies or the weighted worst case real accuracies so as to minimize the overall classification risk. Moreover, similar to the MS case, it is easily validated that this optimization problem can be cast as a sequential BMPM problem. Hence, it can similarly be solved based on the method presented in Section III.

B. Open Problems and Future Work

We discuss the limitations of the BMPM model and present future work. First, currently, the applications of the MPM model and the

BMPM model are restricted in the two-category classification domain. Although there are systematic methods, e.g., one-versus-all or one-versus-one [2], [12], to extend the two-category classifications into multiway classifications, for BMPM, it needs to be careful. To derive a multiway boundary efficiently while maintaining a tight lower bound is not straightforward. Further explorations and considerations on this topic are deserved.

Second, although we propose efficient algorithms to solve the BMPM optimization problems, one interesting question for both MPM and BMPM is that whether any techniques can be used to speed up the training process, especially the kernelized training process. Another problem in training the MPM or the BMPM model with kernels, e.g., the Gaussian kernel, is that the parameter σ has to be determined via the time-consuming cross-validation procedure. How to speed up these processes is one of the open problems for both MPM and BMPM.

Third, to assure a tight lower bound of the accuracy, both the MPM and the BMPM models require that the mean and the covariance matrices estimated from the dataset can reliably represent the true mean and covariance matrices. It is empirically verified that the direct plug-in estimation achieves a satisfactory performance on many real classification tasks [18], [23]. However, there exist cases where the estimation will be inaccurate and cause problems, i.e., the worst case accuracy cannot bind the real test-set accuracy. To tackle this problem, some robust-estimation techniques could be applied. For example, under the computational consideration, a specific uncertainty model in [23] is proposed to correct the plug-in estimations. However, seeking more robust estimation based on general uncertainty models remains to be an open problem and, therefore, is one of our research topics in the future.

Finally, we have mainly compared our proposed BMPM model with three competitive machine learning techniques, NB, C4.5, and k-NN. Although these methods are widely used in machine learning and even refereed to as state-of-the-art classifiers in some literature [25], [43], there are still many other competitive approaches that can be adopted and modified for imbalanced learning such as those found in [30]. Evaluating our BMPM against other competitive approaches is interesting. We leave this topic as a future work.

VI. Conclusion

A novel model named the BMPM has been proposed to deal with imbalanced-learning problems. This new model can incorporate a certain bias into classification by directly and quantitatively controlling the lower bound of the real accuracy. Therefore, it provides a systematic and rigorous treatment on imbalanced data. We have proved the solvability, provided the technique to incorporate distributional information, and proposed an efficient algorithm to solve the optimization problem of BMPM. Moreover, we have illustrated our approach on a synthetic toy dataset. We have also evaluated our novel model on five real-world datasets in terms of two criteria. In both criteria, the performance is shown to be competitive with the state-of-the-art imbalanced-learning approach, SMOTE. When compared with the other three competitive methods, such as the modified NB classifier, the modified k-NN method, and the modified decision-tree classifier, C4.5, our method, significantly outperforms them.

Acknowledgment

The authors would like to thank M. I. Jordan (U. C. Berkeley) for his encouragement and comments, especially for pointing out the imbalanced direction. Gratitude is also extended to M. A. Maloof (Georgetown University), for his insightful suggestions on the generations of the ROC curves and his distribution of the rooftop dataset, as well as to the editor and the anonymous reviewers for their valuable comments.

References

[1] D. Aha, D. Kibler, and M. Albert, “Instance-based learning algorithms,” *Mach. Learn.*, vol. 6, no. 1, pp. 37–66, Jan. 1991.

[2] E. L. Allwein, R. E. Schapire, and Y. Singer, “Reducing multiclass to binary: A unifying approach for margin classifiers,” *J. Mach. Learn. Res.*, vol. 1, pp. 113–141, Dec. 2000.

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

[4] C. L. Blake, C. J. Merz, (1998). “UCI repository of machine learning databases,” Dept. Inf. Comput. Sci., Univ. California, Irvine. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[5] A. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithm,” *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.

[6] C. Cardie and N. Howe, “Improving minority class prediction using case specific feature weights,” in *Proc. 14th Int. Conf. Machine Learning (ICML)*, Nashville, TN, 1997, pp. 57–65.

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

[8] D. Dori and W. Liu, “Sparse pixel vectorization: An algorithm and its performance evaluation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 3, pp. 202–215, Mar. 1999.

[9] O. Firschein and T. Strat, Eds. *RADIUS: Image Understanding for Imagery Intelligence.* San Francisco, CA: Morgan Kaufmann 1996.

[10] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian network classifiers,” *Mach. Learn.*, vol. 29, no. 2/3, pp. 131–161, Nov. 1997.

[11] J. W. Grzymala-Busse, L. K. Goodwin, and X. Zhang, “Increasing sensitivity of preterm birth by changing rule strengths,” *Pattern Recognit. Lett.*, vol. 24, no. 6, pp. 903–910, Mar. 2003.

[12] C. Hsu and C. Lin, “A comparison of methods for multi-class support vector machines,” *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[13] H. Huang and C. Hsu, “Bayesian classification for data from the same unknown class,” *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 32, no. 2, pp. 137–145, Apr. 2002.

[14] K. Huang, I. King, and M. R. Lyu, “Discriminative training of Bayesian Chow-Liu tree multinet classifiers,” in *Proc. IJCNN*, Portland, OR, 2003, vol. 1, pp. 484–488.

[15] ——, “Finite mixture model of bound semi-naive Bayesian network classifier,” in *Proc. ICANN.* Berlin, Germany: Springer-Verlag, vol. 2714, Lecture Notes in Computer Science, 2003, pp. 115–122.

[16] K. Huang, H. Yang, I. King, and M. R. Lyu, “Learning classifiers from imbalanced data based on biased minimax probability machine,” in *Proc. IEEE Computer Society CVPR*, Washington, DC, 2004, vol. 2, pp. 558–563.

[17] ——, “Learning large margin classifiers locally and globally,” in *Proc. 21st ICML,* R. Greiner and D. Schuurmans, Eds., Banff, AB, Canada, 2004, pp. 401–408.

[18] 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, Oct. 2004.

[19] T. S. Jaakkola and D. Haussler, “Exploiting generative models in discriminative classifiers,” in *Proc. Advances NIPS*, Denver, CO, 1998, vol. 11, pp. 487–493.

[20] R. Kohavi, “A study of cross validation and bootstrap for accuracy estimation and model selection,” in *Proc. 14th IJCAI*, Montreal, QC, Canada, 1995, pp. 338–345.

[21] M. Kubat, R. Holte, and S. Matwin, “Machine learning for the detection of oil spills in satellite radar images,” *Mach. Learn.*, vol. 30, no. 2/3, pp. 195–215, Feb. 1998.

[22] M. Kubat and S. Matwin, “Addressing the curse of imbalanced training sets: One-sided selection,” in *Proc. 14th ICML*, Nashville, TN, 1997, pp. 179–186.

[23] 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.

[24] P. Langley, W. Iba, and K. Thompson, “An analysis of Bayesian classifiers,” in *Proc. Nat. Conf. AAAI*, San Jose, CA, 1992, pp. 223–228.

[25] B. Lerner and N. D. Lawrence, "A comparison of state-of-the-art classification techniques with application to cytogenetics," *Neural Comput. Appl.*, vol. 10, no. 1, pp. 39–47, 2001.

[26] C. Lin and R. Nevatia, "Building detection and description from a single intensity image," *Comput. Vis. Image Understanding*, vol. 72, no. 2, pp. 101–121, Nov. 1998.

[27] C. Ling and C. Li, "Data mining for direct marketing: Problems and solutions," in *Proc. 4th KDD*, New York, 1998, pp. 73–79.

[28] W. Liu and D. Dori, "A protocol for performance evaluation of line detection algorithms," *Mach. Vis. Appl.*, vol. 9, no. 5/6, pp. 240–250, 1997.

[29] 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, no. 1/2, pp. 157–191, Oct./Nov. 2003.

[30] L. Mangasarian, "Linear and nonlinear separation of patterns by linear programming," *Oper. Res.*, vol. 13, no. 3, pp. 444–452, 1965.

[31] D. Mcclish, "Analyzing a portion of the ROC curve," *Med. Decis. Making*, vol. 9, no. 3, pp. 190–195, Jul.–Sep. 1989.

[32] I. Popescu and D. Bertsimas, "Optimal inequalities in probability theory: A convex optimization approach," INSEAD, Dept. Math. O.R., Cambridge, MA, Tech. Rep. TM62, 2001.

[33] F. Provost, "Machine learning from imbalanced data sets 101," in *Proc 17th Nat. Conf. AAAI, Workshop on Imbalanced Data Sets*, Austin, TX, 2000.

[34] F. Provost and T. Fawcett, "Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions," in *Proc. 3rd KDD*, Newport Beach, CA, 1997, pp. 43–48.

[35] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[36] S. Schaible, "Fractional programming," in *Nonconvex Optimization and Its Applications*. Boston, MA: Kluwer, 1995.

[37] S. Schaible and W. T. Ziemba, *Generalized Concavity in Optimization and Economics*. New York: Academic, 1981.

[38] P. Schmidt and A. Witte, *Predicting Recidivism Using Survival Models*. New York: Springer-Verlag, 1988.

[39] B. Scholkopf and A. Smola, *Learning With Kernels*. Cambridge, MA: MIT Press, 2002.

[40] J. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, no. 4857, pp. 1285–1293, Jun. 1988.

[41] J. Swets and R. Pickett, *Evaluation of Diagnostic Systems: Methods From Signal Detection Theory*. New York: Springer-Verlag, 1982.

[42] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York: Springer-Verlag, 1999.

[43] S. Viaene, R. A. Derrig, B. Baesens, and G. Dedene, "A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection," *J. Risk Insur.*, vol. 69, no. 3, pp. 373–421, Sep. 2002.

[44] B. X. Wang and N. Japkowicz, "Imbalanced data set learning with synthetic examples," in *Proc. IRIS Machine Learning Workshop*, Ottawa, ON, Canada, 2004.

[45] K. Woods, W. P. Kegelmeyer, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 4, pp. 405–410, Apr. 1997.