A Class of Single-Class Minimax Probability Machines for Novelty Detection

James T. Kwok, Ivor Wai-Hung Tsang, and Jacek M. Zurada, *Fellow, IEEE*

***Abstract*—Single-class minimax probability machines (MPMs) offer robust novelty detection with distribution-free worst case bounds on the probability that a pattern will fall inside the normal region. However, in practice, they are too cautious in labeling patterns as outlying and so have a high false negative rate (FNR). In this paper, we propose a more aggressive version of the single-class MPM that bounds the best case probability that a pattern will fall inside the normal region. These two MPMs can then be used together to delimit the solution space. By using the hyperplane lying in the middle of this pair of MPMs, a better compromise between false positives (FPs) and false negatives (FNs), and between recall and precision can be obtained. Experiments on the real-world data sets show encouraging results.**

***Index Terms*—Kernel methods, minimax probability machines (MPMs), novelty detection.**

I. Introduction

In recent years, there has been a lot of interest on using kernels in various aspects of machine learning, such as classification, regression, clustering, ranking, and principal component analysis [1]. A well-known example in supervised learning are the support vector machines (SVMs). The basic idea of kernel methods is to map the data from an input space to a feature space $\mathcal{F}$ via some map φ, and then apply a linear procedure there. It is now well known that the computations do not involve φ explicitly, but depend only on the inner product defined in $\mathcal{F}$, which in turn can be obtained efficiently from a suitable *kernel* function (the "kernel trick").

In this paper, we will focus on the use of kernels in novelty detection. Here, the goal is to differentiate the known objects (*normal* patterns) from the unknown objects (*outliers*) [2], [3]. Novelty detection has found many real-world applications, such as the detection of unusual vibration signatures in jet engines [4]. Traditionally, novel patterns are detected by either estimating the density function of the normal patterns or by finding a small set $\mathcal{Q}$ such that $P(\mathbf{x} \in \mathcal{Q}) = \alpha$ for some fixed $\alpha \in (0, 1]$ (quantile estimation). However, they both depend critically on the parametric form of the density function and can fail miserably when this assumption is incorrect.

Manuscript received June 13, 2005; revised March 20, 2006 and October 27, 2006; accepted November 7, 2006. This work was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region under the Grant 615005. The work of J. M. Zurada was supported in part by the Institute of Systems Science, the Polish Academy of Sciences, Warsaw, Poland.

J. T. Kwok and I. W.-H. Tsang are with the Department of Computer Science, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong (e-mail: ivor@cs.ust.hk; jamesk@cs.ust.hk).

J. M. Zurada is with the Department of Electrical and Computer Engineering, the University of Louisville, Louisville, KY 40292 USA (e-mail: jacek.zurada@louisville.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2007.891191

Instead of estimating the density or quantile, a simpler task is to model the support of the data distribution directly. Tax and Duin proposed the support vector data description (SVDD) [5], which uses a small ball to enclose most of the data. Computationally, this leads to a quadratic programming (QP) problem, which has the important advantage that the solution obtained is always globally optimal. Moreover, as with other kernel methods, SVDD works well with high-dimensional data and can be easily kernelized by replacing the dot products between patterns with the corresponding kernel evaluations.

Besides balls, hyperplanes have also been used. Schölkopf *et al.* proposed the one-class SVM [6] that uses a hyperplane to separate normal patterns from the outliers with maximum margin. Again, this leads to a QP problem. Moreover, when Gaussian kernels are used, the one-class SVM solution is equivalent to that of the SVDD. Instead of using QP formulations as in both SVDD and one-class SVM, a linear programming (LP) formulation has also been proposed [7].

Recently, Lanckriet *et al.* proposed the single-class minimax probability machine (MPM) [8] that is also based on the use of hyperplanes. However, it is distinctive in that a distribution-free probability bound, based on the use of generalized Chebychev inequalities [9], can be provided. Specifically, given only the mean $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{\Sigma}$ of a distribution and *without* making any other distributional assumption, it seeks the smallest half-space $\mathcal{Q}(\mathbf{a}, b) = \{\mathbf{z} \mid \mathbf{a}'\mathbf{z} \geq b\}$ for the normal patterns, not containing the origin, that bounds the worst case probability of a data pattern falling inside of $\mathcal{Q}$ (Fig. 1). Mathematically, given $\alpha \in (0, 1)$, the single-class MPM ensures that

$$\inf_{\mathbf{x} \sim (\bar{\mathbf{x}}, \mathbf{\Sigma})} P(\mathbf{a}'\mathbf{x} \geq b) \geq \alpha \tag{1}$$

where $\mathbf{x} \sim (\bar{\mathbf{x}}, \mathbf{\Sigma})$ denotes the class of distributions with the given values of mean $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{\Sigma}$.

Despite its interesting theoretical properties, the single-class MPM has a high false negative rate (FNR) in practice (in this paper, outliers are treated as positives while normal patterns as negatives) [10]. This can be explained by noting that (1) is equivalent to

$$\sup_{\mathbf{x} \sim (\bar{\mathbf{x}}, \mathbf{\Sigma})} P(\mathbf{a}'\mathbf{x} < b) \leq 1 - \alpha. \tag{2}$$

In other words, the probability of having a data pattern falling outside $\mathcal{Q}$ (i.e., an outlier), for every distribution with given values of $\bar{\mathbf{x}}$ and $\mathbf{\Sigma}$, is upper bounded by $1 - \alpha$. Thus, single-class MPMs are too cautious in labeling a pattern as outlying. To alleviate this problem, Lanckriet *et al.* [10] suggested the provision of some uncertainty information on the covariance matrix. Another possibility is to also use higher order moments (where

1045-9227/$25.00 © 2007 IEEE

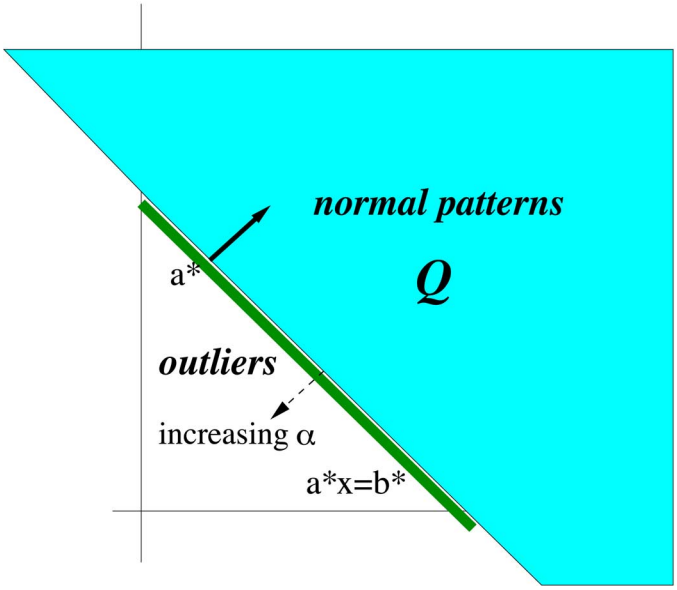

Fig. 1. Single-class MPM. Here, we follow the convention that the outliers are treated as positives while normal patterns are treated as negatives.

order $k > 2$) to better characterize the tail probability behavior of the data distribution. However, when $k \geq 4$, it can be shown that obtaining tight bounds for the corresponding generalized Chebychev inequalities is NP-hard [9]. Moreover, since typically we do not know the moments in advance, they have to be replaced by plug-in estimates based on the empirical data. Very often, only the first- and second-order moments can be reliably estimated, making the use of higher order moments difficult in practice.

In this paper, we address this problem by considering the other side of the coin. We first propose a variant of the single-class MPM that bounds instead the distribution-free, *best* case probability of a pattern falling inside the normal region. It will be seen that this MPM will be more aggressive in labeling patterns as outliers. By using this aggressive version in tandem with the traditional, more conservative single-class MPM, improved novelty detection performance can be obtained. The rest of this paper is organized as follows. First, Section II reviews the traditional single-class MPM. Then, Section III describes the two proposed variants. Experimental results on real-world data sets are presented in Section IV, and Section V gives some concluding remarks.

II. Single-Class MPM

Similar to the one-class SVM [6], the single-class MPM seeks to separate the normal patterns from the origin with the maximum margin. Given the mean $\bar{\mathbf{x}}$ and the covariance matrix $\mathbf{\Sigma}$ of a distribution (and without making any other distributional assumption), the single-class MPM finds the smallest half-space (Fig. 1)

$$\mathcal{Q}(\mathbf{a}, b) = \{\mathbf{z} | \mathbf{a}'\mathbf{z} \geq b\} \tag{3}$$

for the normal patterns, not containing the origin, that minimizes the worst case probability of a data pattern falling inside $\mathcal{Q}$. The size of $\mathcal{Q}$ in (3) can be minimized by maximizing the distance of $\mathbf{a}'\mathbf{z} = b$ from the origin, i.e., $b/\sqrt{\mathbf{a}'\mathbf{\Sigma}\mathbf{a}}$. Hence, for a given $\alpha \in (0, 1)$, this leads to the following constrained optimization problem:

$$\max_{\mathbf{a} \neq 0, b} \frac{b}{\sqrt{\mathbf{a}'\mathbf{\Sigma}\mathbf{a}}} : \inf_{\mathbf{x} \sim (\bar{\mathbf{x}}, \mathbf{\Sigma})} P(\mathbf{a}'\mathbf{x} \geq b) \geq \alpha. \tag{4}$$

For novelty detection, α is typically chosen close to 1 so that most of the patterns are contained in the half-space $\mathcal{Q}$ while those outside $\mathcal{Q}$ are the outliers.

An important tool in deriving the MPM is the generalized Chebychev inequality [9]. Over all distributions of $\mathbf{x}$ having the given values of mean $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{\Sigma}$, the generalized Chebychev inequality bounds the probability of a random pattern $\mathbf{x}$ falling in a convex set $\mathcal{S}$. Mathematically

$$\sup_{\mathbf{x} \in (\bar{\mathbf{x}}, \mathbf{\Sigma})} P(\mathbf{x} \in \mathcal{S}) = \frac{1}{1 + d^2}$$

where $d^2 = \inf_{\mathbf{x} \in S} (\mathbf{x} - \bar{\mathbf{x}})' \mathbf{\Sigma}^{-1} (\mathbf{x} - \bar{\mathbf{x}})$. Lanckriet *et al.* [11] showed that the constraint in (4) is the same as

$$\mathbf{a}'\bar{\mathbf{x}} - 1 \geq \kappa(\alpha) \sqrt{\mathbf{a}'\mathbf{\Sigma}\mathbf{a}} \tag{5}$$

where

$$\kappa(\alpha) = \sqrt{\frac{\alpha}{1 - \alpha}}. \tag{6}$$

Denote

$$\zeta = \sqrt{\bar{\mathbf{x}}' \mathbf{\Sigma}^{-1} \bar{\mathbf{x}}}. \tag{7}$$

They also showed that if

$$\zeta > \kappa(\alpha) \tag{8}$$

the optimal values of $\mathbf{a}$ and b (denoted $\mathbf{a}_1^*$ and b_1^*, respectively) in (4) are obtained as

$$\mathbf{a}_1^* = \frac{\mathbf{\Sigma}^{-1} \bar{\mathbf{x}}}{\zeta^2 - \kappa(\alpha)\zeta} \quad b_1^* = 1 \tag{9}$$

otherwise, the problem is infeasible. Note that one can multiply $\mathbf{a}_1^*$ and b_1^* by the same constant without changing the hyperplane. On testing, a pattern $\mathbf{z}$ will be predicted as an outlier if it lies on the side of the hyperplane containing the origin, i.e., $\mathbf{a}_1^{*\prime}\mathbf{z} \leq b_1^*$.

Similar to the traditional techniques of the principal component analysis and Fisher discriminant analysis [12], MPM also relies heavily on the use of the covariance matrix. In practice, it is not known *a priori* and has to be estimated from the data. To improve robustness, Lanckriet *et al.* [8], [11] suggested the incorporation of uncertainties on $\bar{\mathbf{x}}$ and $\mathbf{\Sigma}$ into the optimization problem. In particular, consider an uncertainty set of the form

$$\left\{ (\bar{\mathbf{x}}, \mathbf{\Sigma}) : (\bar{\mathbf{x}} - \bar{\mathbf{x}}^0)' \mathbf{\Sigma}^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{x}}^0) \leq \delta^2, \|\mathbf{\Sigma} - \mathbf{\Sigma}^0\|_F \leq \rho \right\} \tag{10}$$

where $\bar{\mathbf{x}}^0$ and $\mathbf{\Sigma}^0$ are the nominal estimates of $\bar{\mathbf{x}}$ and $\mathbf{\Sigma}$, $\|\cdot\|_F$ denotes the Frobenious norm, and $\delta, \rho \geq 0$ are the corresponding

uncertainties. Then, if $\kappa(\alpha)+\delta<\zeta$, the problem is strictly feasible and $\mathbf{a}_1^*$ in (9) will be changed to

$$\mathbf{a}_1^*=\frac{(\boldsymbol{\Sigma}+\rho\mathbf{I})^{-1}\bar{\mathbf{x}}}{\zeta^2-(\kappa(\alpha)+\delta)\zeta}$$

while the value of b_1^* remains unchanged. Alternatively, robust covariance matrix estimation methods (such as [13] and [14]) can also be used.

III. Family of Single-Class MPMs

If the underlying data distribution is known, one no longer needs the infimum operator in (4). However, this is often impossible. Alternatively, nonparametric density estimation may be employed, though it is costly and difficult on high-dimensional data. Instead, the single-class MPM in Section II guarantees a certain probability mass for the normal region in the worst case. However, typically, the actual data distribution does not correspond to this worst case. Thus, this MPM is very conservative in classifying patterns as outliers and often misses a lot of the outliers in practice. To improve its outlier detection capability, obviously one should not move the hyperplane solution $(\mathbf{a}\mathbf{x}=b)$ in Fig. 1 further down, for otherwise even more patterns will be classified as normal. A simple remedy is thus to move the hyperplane up by adjusting the bias b. However, it is unclear why only the bias needs to be adjusted. Moreover, as $b\in(-\infty,+\infty)$, there is the question of how to set the bias in this infinite range.

A. Best Case Scenario: Aggressive Single-Class MPM

In this section, instead of considering the worst case scenario as in Section II, we consider the *best case* scenario. While the worst case corresponds to the infimum operator in (4), the best case corresponds to the supremum operator. This new single-class MPM will thus be more aggressive in labeling patterns as outliers, though also liable to having a higher false alarm rate. Interestingly, it can be shown that these two extreme versions of the single-class MPM differ only in the value of the bias. Thus, one can then move between these two extremes by simply varying the bias (Section III-B).

1) Formulation: Considering now the best case situation, the optimization problem in (4) becomes

$$\max_{\mathbf{a}\neq 0,b}\frac{b}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}}:\sup_{\mathbf{x}\sim(\bar{\mathbf{x}},\boldsymbol{\Sigma})}P(\mathbf{a}'\mathbf{x}\geq b)\geq\alpha. \tag{11}$$

As in the traditional single-class MPM, b has to be positive in order for $\mathcal{Q}$ not to contain the origin. Moreover, (11) is also homogeneous in $(\mathbf{a},b)$. Hence, without loss of generality, we can set $b=1$ and rewrite (11) as

$$\min_{\mathbf{a}\neq 0}\sqrt{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}:\sup_{\mathbf{x}\sim(\bar{\mathbf{x}},\boldsymbol{\Sigma})}P(\mathbf{a}'\mathbf{x}\geq 1)\geq\alpha. \tag{12}$$

There are the following two cases to consider.

Case 1: $\mathbf{a}'\bar{\mathbf{x}}<1$. From the proof of [11, Lemma 1], the left-hand side of the constraint in (12) is given by $\sup_{\mathbf{x}\sim(\bar{\mathbf{x}},\boldsymbol{\Sigma})}P(\mathbf{a}'\mathbf{x}\geq 1)=1/(1+d^2)$, where

$$d^2=\inf_{\mathbf{a}'\mathbf{x}\geq 1}(\mathbf{x}-\bar{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\bar{\mathbf{x}})=\frac{\max(1-\mathbf{a}'\bar{\mathbf{x}},0)^2}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}. \tag{13}$$

The constraint in (12) then becomes

$$\frac{1}{1+d^2}\geq\alpha\Leftrightarrow\frac{\max(1-\mathbf{a}'\bar{\mathbf{x}},0)^2}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}\leq\frac{1-\alpha}{\alpha}=\kappa^2(1-\alpha) \tag{14}$$

on using (13) and the definition of $\kappa(\cdot)$ in (6). As we are considering the case $\mathbf{a}'\bar{\mathbf{x}}<1$, $\max(1-\mathbf{a}'\bar{\mathbf{x}},0)=1-\mathbf{a}'\bar{\mathbf{x}}$ and (14) reduces to

$$1-\mathbf{a}'\bar{\mathbf{x}}\leq\kappa(1-\alpha)\sqrt{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}. \tag{15}$$

Then, the constrained optimization problem in (12) can be written as

$$\min_{\mathbf{a}\neq 0}\sqrt{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}:\mathbf{a}'\bar{\mathbf{x}}\geq 1-\kappa(1-\alpha)\sqrt{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}. \tag{16}$$

With the change of notations $\mathbf{a}\rightarrow\boldsymbol{\Sigma}^{-1/2}\mathbf{a}$ and $\bar{\mathbf{x}}\rightarrow\boldsymbol{\Sigma}^{1/2}\bar{\mathbf{x}}$, this further becomes

$$\min_{\mathbf{a}\neq 0}\|\mathbf{a}\|:\mathbf{a}'\bar{\mathbf{x}}\geq 1-\kappa(1-\alpha)\|\mathbf{a}\| \tag{17}$$

where $\|\cdot\|$ denotes the usual Euclidean norm. This can be easily solved by using the method of Lagrange multipliers, yielding

$$\frac{\mathbf{a}}{\|\mathbf{a}\|}=\frac{\gamma}{1-\gamma\kappa(1-\alpha)}\bar{\mathbf{x}}.$$

In other words, $\mathbf{a}=\lambda\bar{\mathbf{x}}$ for some $\lambda\in\mathbb{R}$. Obviously, λ has to be positive,[1] for otherwise when a pattern $\mathbf{z}$ moves from the origin (which must be an outlier according to the definition of the single-class MPMs) towards $\bar{\mathbf{x}}$, the value of $\mathbf{a}'\mathbf{z}-b$ will decrease, leading to the counterintuitive prediction that $\mathbf{z}$ is even more likely to be an outlier. Hence, with $\mathbf{a}=\lambda\bar{\mathbf{x}}$ where $\lambda>0$, (17) can be written as

$$\begin{aligned}\min_{\lambda>0}\lambda\|\bar{\mathbf{x}}\|:\lambda\|\bar{\mathbf{x}}\|^2&\geq 1-\kappa(1-\alpha)\lambda\|\bar{\mathbf{x}}\|\\&=\min_{\lambda>0}\lambda\|\bar{\mathbf{x}}\|:\lambda\left(\|\bar{\mathbf{x}}\|^2+\kappa(1-\alpha)\|\bar{\mathbf{x}}\|\right)\geq 1.\end{aligned}$$

Obviously, $\lambda=1/(\|\bar{\mathbf{x}}\|^2+\kappa(1-\alpha)\|\bar{\mathbf{x}}\|)$, and so

$$\mathbf{a}=\frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|^2+\kappa(1-\alpha)\|\bar{\mathbf{x}}\|}. \tag{18}$$

Case 2: $\mathbf{a}'\bar{\mathbf{x}}\geq 1$. In this case, we can just take $\mathbf{x}=\bar{\mathbf{x}}$ and $\sup_{\mathbf{x}\sim(\bar{\mathbf{x}},\boldsymbol{\Sigma})}P(\mathbf{a}'\mathbf{x}\geq 1)=1>\alpha$, and so the constraint in (12) is automatically satisfied. To find $\mathbf{a}$, we have to solve

$$\min_{\mathbf{a}\neq 0}\sqrt{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}:\mathbf{a}'\bar{\mathbf{x}}\geq 1. \tag{19}$$

[1]For $\lambda<0$, it can be shown that $\mathbf{a}=-\bar{\mathbf{x}}/(\kappa(1-\alpha)\|\bar{\mathbf{x}}\|-\|\bar{\mathbf{x}}\|^2)$ when $\kappa(1-\alpha)>\|\bar{\mathbf{x}}\|$, and has no feasible solution otherwise.

With the change of notations $\mathbf{a} \rightarrow \mathbf{\Sigma}^{-1/2}\mathbf{a}$ and $\bar{\mathbf{x}} \rightarrow \mathbf{\Sigma}^{1/2}\bar{\mathbf{x}}$, this becomes

$$\min_{\mathbf{a}\neq 0} ||\mathbf{a}|| : \mathbf{a}'\bar{\mathbf{x}} \geq 1. \tag{20}$$

Using the method of Lagrange multipliers, we again obtain $\mathbf{a} = \lambda\bar{\mathbf{x}}$ for some $\lambda \in \mathbb{R}$. As in Case 1, λ has to be positive. Hence, with $\mathbf{a} = \lambda\bar{\mathbf{x}}$ where $\lambda > 0$, (20) can be written as

$$\min_{\lambda>0} \lambda||\bar{\mathbf{x}}|| : \lambda||\bar{\mathbf{x}}||^2 \geq 1$$

yielding $\lambda = 1/||\bar{\mathbf{x}}||^2$, and so

$$\mathbf{a} = \frac{\bar{\mathbf{x}}}{||\bar{\mathbf{x}}||^2}. \tag{21}$$

As $\kappa(1-\alpha) > 0$ for $\alpha \in (0,1)$, the value of the objective function $||\mathbf{a}||$ obtained by (18) is smaller than that obtained by (21). In other words, Case 1 yields the optimal solution. Transforming the solution in (18) back to the original coordinates, the optimal values of $\mathbf{a}$ and b (denoted $\mathbf{a}_2^*$ and b_2^*, respectively) are then

$$\mathbf{a}_2^* = \frac{\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}}{\zeta^2 + \kappa(1-\alpha)\zeta} \quad b_2^* = 1. \tag{22}$$

Prediction is still performed as for the traditional single-class MPM, i.e., a test pattern $\mathbf{z}$ will be predicted as an outlier if $\mathbf{a}_2^{*\prime}\mathbf{z} \leq b_2^*$.

2) Remarks: Recall that one can multiply $\mathbf{a}$ and b by the same constant without changing the hyperplane. Hence, this pair of the single-class MPMs share the same direction and differ only in the value of b. Moreover, it can be easily seen that they are identical when $\alpha = 0$, but then move in different directions as α increases. Moreover, as in the traditional MPM, uncertainties on $\bar{\mathbf{x}}$ and $\mathbf{\Sigma}$ in the form of (10) can also be easily incorporated. It can be shown that the optimal values of $\mathbf{a}$ and b in (22) are subsequently changed to

$$\mathbf{a}_2^* = \frac{(\mathbf{\Sigma} + \rho\mathbf{I})^{-1}\bar{\mathbf{x}}}{\zeta^2 + (\kappa(1-\alpha) + \delta)\,\zeta} \quad b_2^* = 1$$

where ζ in (7) is changed accordingly to $\zeta = \sqrt{\bar{\mathbf{x}}'(\mathbf{\Sigma} + \rho\mathbf{I})^{-1}\bar{\mathbf{x}}}$. It can also be kernelized in an analogous manner, thus effectively allowing the use of nonlinear boundaries to separate outliers from normal patterns.

Finally, consider the special case where $\mathbf{x}$ is indeed normally distributed as $\mathcal{N}(\bar{\mathbf{x}}, \mathbf{\Sigma})$. It can be shown, in a manner analogous to that for the two-class MPM in [11], that the only change for the original single-class MPM in Section II is to replace $\kappa(\alpha)$ in (9) by $\Phi^{-1}(\alpha)$, where $\Phi(z) = P(\mathcal{N}(0,1) \leq z)$ is the cumulative distribution function for a standard univariate normal distribution. Similarly, for the aggressive single-class MPM, one has to replace $\kappa(1-\alpha)$ by $\Phi^{-1}(1-\alpha)$. As $\kappa(\alpha) > \Phi^{-1}(\alpha)$ for $\alpha > 0$, the hyperplane $\mathbf{a}_1^{*\prime}\mathbf{x} = b_1^*$ is shifted in parallel away from the origin, while the hyperplane $\mathbf{a}_2^{*\prime}\mathbf{x} = b_2^*$ is shifted in parallel towards the origin. This, again, is in accordance with the fact that the traditional single-class MPM corresponds to the worst case situation, while the aggressive single-class MPM corresponds to the best case situation.

B. Moderate Single-Class MPM

The traditional single-class MPM is conservative in labeling patterns as outliers, and has a low false positive rate (FPR) but a high FNR. On the other hand, the new one is aggressive, and has a low FNR but high FPR. Nevertheless, patterns that are predicted as outliers by the conservative MPM are very likely to be true outliers, while those predicted as normal by the aggressive MPM are very likely to be truly normal. In classification problems, one sometimes has the option of rejecting patterns that are not unrecognizable. Analogously, we can obtain a more reliable novelty detector by combining these two single-class MPMs as follows.

1) If the test pattern $\mathbf{z}$ is predicted to be an outlier by the conservative MPM, predict that $\mathbf{z}$ is an outlier.
2) If $\mathbf{z}$ is predicted as normal by the aggressive MPM, predict that $\mathbf{z}$ is normal.
3) Otherwise ($\mathbf{z}$ is predicted as normal by the conservative MPM but as an outlier by the aggressive MPM), the confidence is low and $\mathbf{z}$ is rejected as being unrecognizable.

However, such a reject option may not be feasible in applications where the cost for rejects is high. Moreover, a large amount of patterns may also have to be rejected.

Note that the traditional and aggressive single-class MPMs are at different ends of the solution space. Moving up the hyperplane solution of the aggressive MPM ($\mathcal{H}_2 : \mathbf{a}_2^{*\prime}\mathbf{x} = b_2^*$) will classify even more patterns as outliers, while moving down that of the traditional MPM ($\mathcal{H}_1 : \mathbf{a}_1^{*\prime}\mathbf{x} = b_1^*$) will classify even more patterns as normal. Thus, both are clearly undesirable. Recall that $\mathcal{H}_1$ and $\mathcal{H}_2$ are parallel. Hence, an "optimal" solution can be obtained by simply varying the bias b in the finite range (b_1^*, b_2^*). However, this is still difficult unless one is willing to make assumptions on the data distribution. Alternatively, one may assume the presence of labeled normal patterns and outliers (that may be artificially generated [15]) and then, use resampling procedures such as cross validation to determine the optimal value of b. However, many novelty detection problems do not have the luxury of labeled data available, while the use of artificially generated data implicitly assumes the underlying data distribution. Besides, when labeled data is present, it is often better to address the novelty detection problem as a two-class, rather than one-class, classification problem [16].

As the actual data distribution is likely to correspond to neither the worst case nor the best case scenario, studying the average case situation will be more useful. However, without additional information such as some distributional assumption on the data, a formal analysis of the average case is impossible. In the following, we attempt to approximate the average case scenario. An intuitive compromise is struck by translating the hyperplane such that it is furthest away from both extremes, resulting in the hyperplane $\mathbf{a}_3^{*\prime}\mathbf{x} = b_3^*$ that is midway between $\mathcal{H}_1$ and $\mathcal{H}_2$. This can also be viewed as the Bayes point machine [17] located at the center of mass of the solution space $\{b : b \in (b_1^*, b_2^*)\}$. Moreover, recall that patterns predicted as

outliers by the conservative MPM are very likely to be true outliers, while those predicted as normal by the aggressive MPM are very likely to be truly normal. This "moderate" variant of the MPM is also the maximum-margin hyperplane that separates the confidently labeled outliers from the confidently labeled normal patterns. Recall that one can multiply $\mathbf{a}_3^*$ and b_3^* by the same arbitrary constant without changing the hyperplane; it can be shown that

$$\mathbf{a}_3^* = \mathbf{\Sigma}^{-1}\bar{\mathbf{x}}, \quad b_3^* = \zeta^2 + \frac{\kappa(1-\alpha) - \kappa(\alpha)}{2}\zeta. \tag{23}$$

Similar to its predecessors, this MPM can also be easily kernelized.[2] As will be demonstrated by the experiments in Section IV, empirically, it is a better compromise between false positives (FPs) and false negatives (FNs) than both of its conservative and aggressive predecessors.

IV. Experiments

In this section, we compare the performance of the various single-class MPMs on a number of real-world data sets. For simplicity of notations, we will denote the single-class MPMs in Sections II, III-A, and III-B as MPM1, MPM2, and MPM3, respectively.

A. Setup

Four real-world data sets[3] (Table I) from the University of California at Irvine (UCI) machine learning repository [18] are used in the experiments. All these are two-class problems. For each data set, we take the larger class as normal data and the other as outliers. We follow the setup in [8] by randomly sampling 80% of the normal patterns for training (no outlier is used). The remaining 20% of the normal patterns and all the outliers are used for testing. To reduce statistical variability, results here are based on averages over 100 random repetitions. Note that as only one class of patterns (the normal class) is used during training, standard resampling techniques such as cross validation cannot be used to determine the "optimal" bias.

To alleviate the possible problem of different scalings for different dimensions, we use the automatic relevance determination (ARD) kernel that automatically adapts different widths for each dimension [19]. It is defined as $k(\mathbf{x},\mathbf{y}) = \exp(-\sum_i((x_i - y_i)^2/\sigma_i^2))$, where x_is (and, similarly, for y_is) are the components of $\mathbf{x}$ and σ_i^2 is the 90% quantile of the value of $(x_i - y_i)^2$ over the training data.[4] As in [11], we add $\rho\mathbf{I}$ (where $\rho = 0.01$) to the plug-in estimate of the covariance matrix. The setting of α is typically unknown. In the experiments, we vary its value to study the performance of the various MPMs.

The following popular performance criteria will be adopted:

1) FP rate: FPR = FP/(TN + FP);
2) FN rate: FNR = FN/(TP + FN);
3) balanced loss: ℓ_{bal} = (FPR + FNR)/2;
4) F-value = 2· recall · precision/(recall + precision).

The two performance criteria, precision = (TP + 1)/(TP + FP + 1) and recall = TP/(TP + FN), have also been used. As expected, a low FPR empirically corresponds to high precision while low FNR corresponds to high recall, and vice versa. Hence, to reduce clutter, figures on precision and recall are not shown here.

Recall that outliers are treated as positives while normal patterns as negatives. The balanced loss can be regarded as an aggregate performance measure for FPR and FNR, while the F-value is another aggregate for precision and recall. As can be seen from Table I, the sizes of the normal patterns and outliers are very different in the test set, and the balanced loss has been shown to be particularly suitable in such an imbalanced setting [20]. A good novelty detector should attain low values on FPR, FNR, and ℓ_{bal}, however, high values on precision, recall, and F-value. The receiver operating characteristic (ROC) graph is also a common tool for performance evaluation [21]. However, as the three MPMs differ only in the bias, they share the same ROC and so ROC analysis cannot be applied here.

TABLE I
Data Sets Used in the Experiments

date set	training # normal patterns	testing # normal patterns	testing # outliers	# attributes
adult	968	242	395	123
breast cancer	356	88	239	9
ionosphere	180	45	126	33
wdbc	286	71	212	30

B. Experimental Results

Results on the data sets are shown in Fig. 2. As α increases, note from (9), (22), and (23) that the biases of all three MPMs decrease and move closer to the origin. Consequently, all their FPRs decrease while their FNRs increase. Besides, MPM3 is almost identical to MPM2 when α is small. In particular, when $\alpha \to 0$, both b_2^* and b_3^* approach $+\infty$, as can be seen from (22) and (23), and, consequently, all the patterns are classified as outliers by MPM2 and MPM3. Hence, a small value of α should not be used for MPM2 and MPM3. As a rule of thumb, we recommend the setting of $0.6 \leq \alpha \leq 0.8$. This is not overly restrictive as we assume that outliers form the minority class (and so $\alpha > 0.5$). Moreover, as MPM1 (and, consequently, MPM3) becomes infeasible, when (8) is violated and $\kappa(\alpha)$ in (8) is monotonically increasing with α, α cannot be too large.

As can be seen from Fig. 2, the various MPMs have different strengths and weaknesses and its appropriateness depends on the specific situations. If the goal is to obtain low FPR and high precision, MPM1 is the most desirable. This is then followed by MPM3, and the last one is MPM2. On the other hand, for good FNR and recall, the rankings are reversed: MPM2 now becomes the best, which is then followed by MPM3 and the worst is MPM1. The results in Fig. 2 also confirm our previous discussion that MPM1 is reluctant to label patterns as outliers. On the other hand, MPM2 is too aggressive in detecting outliers. Hence, many patterns are labeled as outliers and the FPR obtained is high (often close to one even when α is large). Clearly,

[2]Given a kernel function k, define $\mathbf{K} = [k(\mathbf{x}_i,\mathbf{x}_j)]_{ij} \in \mathbb{R}^{N\times N}$ the kernel matrix, $\boldsymbol{\ell} = [\ell_1,\ldots,\ell_N]'$, where $\ell_i = (1/N)\sum_{j=1}^N k(\mathbf{x}_j,\mathbf{x}_i)$ and $\mathbf{L} = (1/\sqrt{N})(\mathbf{K} - \mathbf{1}_N\boldsymbol{\ell}')$, where $\mathbf{1}_N = [1,\ldots,1]' \in \mathbb{R}^N$. It can be shown that the hyperplane solution is $\sum_{i=1}^N \gamma_i^* k(\mathbf{x}_i,\mathbf{x}) + b^*$, where $\boldsymbol{\gamma}^* = [\gamma_1^*,\ldots,\gamma_N^*]' = (\mathbf{L}'\mathbf{L})^{-1}\boldsymbol{\ell}$ and $b^* = \boldsymbol{\ell}'(\mathbf{L}'\mathbf{L})^{-1}\boldsymbol{\ell} + ((\kappa(1-\alpha) - \kappa(\alpha))/2)\sqrt{\boldsymbol{\ell}'(\mathbf{L}'\mathbf{L})^{-1}\boldsymbol{\ell}}$.

[3]Experiments have been performed on more data sets. However, because of the lack of space, only results on these four representative data sets are reported here.

[4]As in [19], this value is estimated from a 20% random sample of the training data.

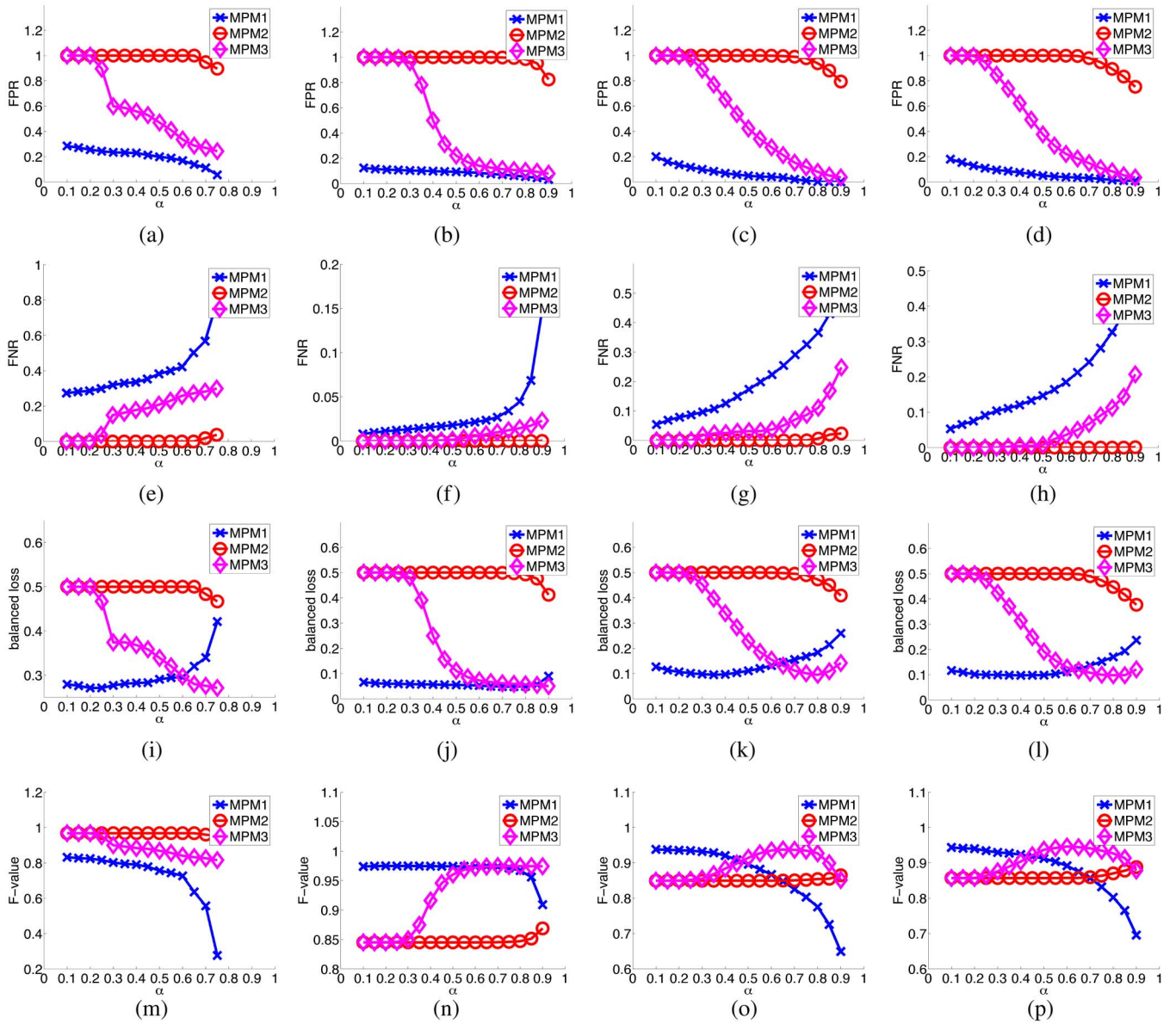

Fig. 2. Novelty detection results. First row: FPR. Second row: FNR. Third row: balanced loss. Fourth row: F-value. (a), (e), (i), and (m) Adult. (b), (f), (j), and (n) Breast cancer. (c), (g), (k), and (o) Ionosphere. (d), (h), (l), and (p) wdbc.

this reflects that the underlying data distributions of these data sets correspond to neither the worst case nor the best case situations as considered by the MPM models. On the other hand, the MPM3, which corresponds to the "average case," obtains a better tradeoff between FPR and FNR than both of its conservative and aggressive predecessors, especially when α is set in the recommended range of [0.6, 0.8].

As for the aggregated performance measures of the balanced loss and F-value, the specific rankings of the MPMs again depend on the value of α. As discussed earlier, a small value of α leads to the poor performance for both the MPM2 and MPM3. However, with $\alpha \in [0.6, 0.8]$ as suggested, the empirical performance of MPM3 is almost always close to the best performance achievable by the other MPMs over the whole range of α. In particular, MPM3 is clearly superior on the ionosphere and wdbc data sets.

For comparison, we also perform experiments with the one-class SVM [6]. To reduce clutter, we only show the graphs for the F-value and balanced loss in Fig. 3. As can be seen, MPM3 with $\alpha \in [0.6, 0.8]$ often leads to a lower balanced loss and a higher/comparable F-value than the one-class SVM.

V. Conclusion

In this paper, we propose a simple and intuitive method for performing novelty detection. This is based on a more aggressive version of the single-class MPM that guarantees a certain probability mass for the normal region in the best case situation. By using a moderate version that is midway between this pair of single-class MPMs, we obtain good FNR and precision (which is typical of a conservative novelty detector) as well as good FPR and recall (typical of an aggressive novelty detector). Experiments on the real-world data sets show encouraging results,

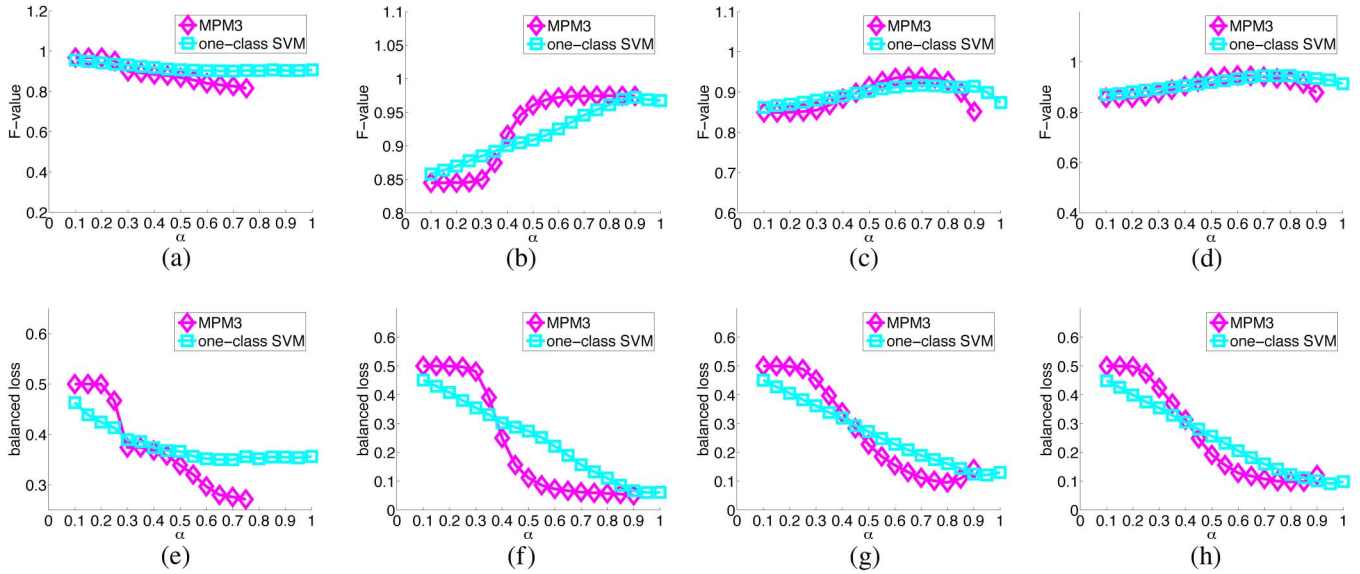

Fig. 3. Comparing the MPM3 with one-class SVM. Top row: F-value. Bottom row: balanced loss. (a) and (e) Adult. (b) and (f) Breast cancer. (c) and (g) Ionosphere. (d) and (h) wdbc.

especially in terms of the aggregated performance measures of balanced loss and F-value.

In general, FPs and FNs may have different costs. In the future, we will explore the computation of probabilistic outputs for this moderate version of the single-class MPM, which can then be used for cost-sensitive novelty detection. Moreover, the effect of using robust estimators for the mean and covariance matrix (such as [13] and [14]) will also be studied.

REFERENCES

[1] B. Schölkopf and A. Smola, *Learning With Kernels*. Cambridge, MA: MIT Press, 2002.
[2] M. Markou and S. Singh, "Novelty detection: a review, part I: Statistical approaches," *Signal Process.*, vol. 83, no. 12, pp. 2481–2497, 2003.
[3] S. Marsland, "Novelty detection in learning systems," *Neural Comput. Surv.*, vol. 3, pp. 157–195, 2003.
[4] P. Hayton, B. Schölkopf, L. Tarassenko, and P. Anuzis, "Support vector novelty detection applied to jet engine vibration spectra," in *Advances in Neural Information Processing Systems 13*, T. Leen, T. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2001.
[5] D. Tax and R. Duin, "Support vector domain description," *Pattern Recognit. Lett.*, vol. 20, no. 14, pp. 1191–1199, 1999.
[6] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.
[7] C. Campbell and K. Bennet, "A linear programming approach to novelty detection," in *Advances in Neural Information Processing Systems 14*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2002.
[8] G. Lanckriet, L. El Ghaoui, and M. Jordan, "Robust novelty detection with single-class MPM," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003.
[9] D. Bertsimas and I. Popescu, "Optimal inequalities in probability theory: a convex optimization approach," *SIAM J. Optim.*, vol. 15, no. 3, pp. 780–804, 2005.
[10] G. Lanckriet, L. El Ghaoui, and M. Jordan, "Robust novelty detection with single-class MPM," in *IMA Workshop Semidefinite Programm. Robust Optim.*, 2003.
[11] G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. Jordan, "A robust minimax approach to classification," *J. Mach. Learn. Res.*, vol. 3, pp. 555–582, 2002.
[12] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.
[13] F. Alqallaf, K. Konis, R. Martin, and R. Zamar, "Scalable robust covariance and correlation estimates for data mining," in *Proc. 8th Int. Conf. Knowl. Discovery Data Mining (ACM SIGKDD)*, Edmonton, AB, Canada, 2002, pp. 14–23.
[14] N. Wang and A. Raftery, "Nearest neighbor variance estimation (NNVE): Robust covariance estimation via nearest neighbor cleaning (with discussion)," *J. Amer. Statist. Assoc.*, vol. 97, no. 460, pp. 994–1019, 2002.
[15] D. Tax and R. Duin, "Uniform object generation for optimizing oneclass classifiers," *J. Mach. Learn. Res.*, vol. 2, pp. 155–173, 2001.
[16] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection," *J. Mach. Learn. Res.*, vol. 6, pp. 211–232, 2005.
[17] R. Herbrich, T. Graepel, and C. Campbell, "Bayes point machines," *J. Mach. Learn. Res.*, vol. 1, pp. 245–279, 2001.
[18] C. Blake, E. Keogh, and C. Merz, "UCI repository of machine learning databases," Dept. Inf. Comput. Sci., Univ. California, Irvine, 1998 [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html
[19] C. Ong, A. Smola, and R. Williamson, "Learning the kernel with hyperkernels," *J. Mach. Learn.*, vol. 6, pp. 1043–1071, 2005.
[20] B. Schölkopf, J. Weston, E. Eskin, C. Leslie, and W. Noble, "A kernel approach for learning from almost orthogonal patterns," in *Proc. 13th Eur. Conf. Mach. Learn.*, Helsinki, Finland, Aug. 2002, pp. 511–528.
[21] A. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.

James T. Kwok received the Ph.D. degree in computer science from the Hong Kong University of Science and Technology (HKUST), Hong Kong, in 1996.

He then joined the Department of Computer Science, Hong Kong Baptist University, Hong Kong, as an Assistant Professor. He returned to HKUST in 2000 and is now an Associate Professor in the Department of Computer Science. His research interests include kernel methods, machine learning, pattern recognition, and artificial neural networks.

Dr. Kwok is an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS and the *Journal of Neurocomputing*.

Ivor Wai-Hung Tsang received the B.Eng. and M.Phil. degrees in computer science, in 2001 and 2003, respectively, from the Hong Kong University of Science and Technology (HKUST), Hong Kong, where he is currently working toward the Ph.D. degree.

He was awarded the IEEE TRANSACTIONS ON NEURAL NETWORKS Outstanding 2004 Paper Award, the Microsoft Fellowship in 2005, the Best Paper Award from the IEEE Hong Kong Chapter of Signal Processing Postgraduate Forum in 2006, and also an Honor Outstanding Student in 2001. His scientific interests includes machine learning and kernel methods.

Jacek M. Zurada (M'82–SM'83–F'96) received the M.S. and Ph.D. degrees in electrical engineering from the Technical University of Gdansk, Gdansk, Poland.

He is the Samuel T. Fife Alumni Professor, and Chair of the Electrical and Computer Engineering Department, University of Louisville, Louisville, KY. He was the coeditor of *Knowledge-Based Neurocomputing* (MIT Press: 2000), the author of Introduction to *Artificial Neural Systems* (PWS: 1992), contributor to the 1994 and 1995 volumes of *Progress in Neural Networks* (Ablex), and coeditor of *Computational Intelligence: Imitating Life* (IEEE Press: 1994). He is the author or coauthor of more than 250 journal and conference papers in the area of neural networks, computational intelligence, data mining, image processing, and very large scale integration (VLSI) circuits. He has delivered numerous invited plenary conference presentations and seminars throughout the world. In March 2003, he was conferred the Title of Professor by the President of Poland, Aleksander Kwasniewski, the Honorary Professorship of Hebei University in China, and, since 2005, he has been serving as a Foreign Member of the Polish Academy of Sciences.

Dr. Zurada was an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—PART I: REGULAR PAPERS and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—PART II: EXPRESS BRIEFS. From 2001 to 2003, he was a member of the Editorial Board of the IEEE PROCEEDINGS. From 1998 to 2003, he was the Editor-in-Chief of the IEEE TRANSACTIONS ON NEURAL NETWORKS. He is an Associate Editor of *Neurocomputing*. He has received a number of awards for distinction in research and teaching, including the 1993 Presidential Award for Research, Scholarship and Creative Activity. In 2001, he received the University of Louisville President's Distinguished Service Award for Service to the Profession. He is the Past President and a Distinguished Speaker of IEEE Computational Intelligence Society.