International Journal of Machine Learning and Cybernetics manuscript No.
(will be inserted by the editor)

A framework for imprecise robust one-class classification models

Lev V. Utkin

Received: date / Accepted: date

Abstract A framework for constructing robust one-class classification models is proposed in the paper. It is based on Walley's imprecise extensions of contaminated models which produce a set of probability distributions of data points instead of a single empirical distribution. The minimax and minimin strategies are used to choose an optimal probability distribution from the set and to construct optimal separating functions. It is shown that an algorithm for computing optimal parameters is determined by extreme points of the probability set and is reduced to a finite number of standard SVM tasks with weighted data points. Important special cases of the models, including pari-mutuel, constant odd-ratio, contaminated models and Kolmogorov-Smirnov bounds are studied. Experimental results with synthetic and real data illustrate the proposed models.

Keywords Machine learning · Novelty detection · Classification · Minimax strategy · Support vector machine · Quadratic programming

1 Introduction

A special important problem of the statistical machine learning is the classification problem which can be regarded as a task of classifying some objects into classes (group) in accordance with their properties or features. However, for many real-world problems, the task is not to classify but to detect novel or abnormal instances [7,8,12,30,32]. Comprehensive and interesting reviews of the novelty detection approaches are provided by Markou and Singh [25], by Bartkowiak [1], by Khan and Madden [20], by Hodge and Austin [17]. Novelty detection is the identification of new or unknown data that a machine learning system is not aware of during training. In particular, it aims to detect anomalous observations [10,11,33]. It should be noted that a typical feature of novelty detection models is that only unlabeled samples are available. We make some assumptions on anomalies in order to distinguish between normal and anomalous future observations. One of the most common ways to define anomalies is by saying that anomalies are not concentrated [31]. The problem of statistical outlier detection is also closely related to that of novelty detection.

The first way to solve the novelty detection problem is to estimate the real-valued density of the data and then threshold it at some value. It is pointed out by many authors (see, for instance, [12]) that this way is likely to fail for sparse high-dimensional data. A better way is to model the support of the (unknown) data distribution directly from data, that is, to estimate a binary-valued function f that is positive in a region where the density is high, and negative elsewhere. This leads to a single-class learning formulation. The function allows us to specify the region in the input space where the data are explained by the model. Sample points outside this region can be regarded as anomalous observations.

Accepting novelty detection as the one-class classification (OCC), a lot of novelty detection models using kernel-based methods in the framework of the support vector machine (SVM) have been proposed. These models are called OCC SVMs. There are two main foundation approaches for constructing the OCC SVM. The

Lev V. Utkin
Department of Industrial Control and Automation, St.Petersburg State Forest Technical University, Institutski per. 5, 194021, St.Petersburg, Russia
Tel.: +7-812-6709262
Fax: +7-812-6709262
E-mail: lev.utkin@gmail.com

first approach is proposed by Tax and Duin [35,34]. This is one of the well-known novelty detection models, which can be regarded as an unsupervised learning problem. Tax and Duin [35,34] solve the OCC problem by distinguishing the positive class from all other possible data points. They find a hyper-sphere around the positive class data that contain almost all points in the data set with the minimum radius. This approach is called the Support Vector Data Description. In a nutshell, the approach considers the trade-off between the number of errors made on the training set (number of target objects rejected) and the size of the sphere (its radius). By adapting the kernel function, this approach becomes more flexible than just a sphere in the input space. Markou and Singh [25] pointed out that a drawback of the approach is that it often requires a large data set.

An alternative way to geometrically enclose a fraction of the training data is via a hyperplane and its relationship to the origin proposed by Scholkopf et al. [30,32]. Under this approach, a hyperplane is used to separate the training data from the origin with maximal margin, i.e., the objective is to separate off the region containing the data points from the surface region containing no data. Here the authors consider the trade-off between the number of errors made on the training set (number of target objects rejected) and the margin separation between the training points and the origin. This is achieved by constructing a hyperplane which is maximally distant from the origin with all data points lying on the opposite side from the origin. The data is mapped into the feature space corresponding to the kernel and is separated from the origin with maximum margin.

There are other interesting novelty detection or OCC models (see for instance, [5,8,21]), which are efficient in many applications. In particular, Campbell and Bennett [8] propose a linear programming approach to the OCC problem, which is much simpler than the standard quadratic programming in the OCC SVM. Kwok, Tsang and Zurada [21] propose the so-called single-class minimax probability machines which offer robust novelty detection with distribution-free worst case bounds on the probability that a pattern will fall inside the normal region. Bicego and Figueiredo [5] study a weighted OCC model. According to their approach, every data point has a weight indicating the importance assigned to each point of the training set. A question of getting the corresponding weights for every point is a main drawback of the approach.

In order to overcome some difficulties concerning the weighted OCC model, a general approach for constructing the imprecise robust OCC models is proposed in the paper. A main idea of the approach is to replace the empirical probability distribution which is exploited in the standard SVM [40, Chapter 10.] by another distribution produced by some imprecise inference models [41]. At first glance it would seem that the proposed approach is very similar to the weighted models, for example, [5]. Indeed, the resulting optimization problems sometimes coincide with those used in the weighted models. However, in contrast to the weights models, the weights of data points are probabilities of the points assigned in a specific way which is determined by an underlying imprecise probabilistic model.

The robustness of the proposed models is achieved by considering sets of probability distributions of data points produced by a number of imprecise probability models. In order to solve the classification problem under the set of probability distributions we select two probability distributions from the produced set. First, we select the “worst” distribution providing the largest value of the expected risk. It corresponds to the minimax (pessimistic) strategy in decision making and can be interpreted as an insurance against the worst case [29, Subsection 2.4.2.]. The second distribution minimizes the expected risk and corresponds to the minimin (optimistic) strategy. By using these probability distributions and the above assumptions, we construct the robust OCC model in the framework of OCC SVMs. It turns out that quadratic programming problems implementing the SVM are constructed in a simple way by using extreme points of the set of probabilities produced by imprecise probability models. In other words, the assigned weights in the proposed models coincide with one of the extreme points. This is a very important feature of the minimax and minimin strategies proved in the paper.

The following imprecise probability models are considered in the paper: the linear-vacuous mixture [41, Subsections 3.3.5 and 4.6.5.] which can be regarded as an extension of the well-known ε-contaminated (robust) model [18, Section 4.2.]; the pari-mutuel model, the constant odds-ratio model; bounded vacuous and bounded ε-contaminated robust models; Kolmogorov-Smirnov bounds. It is shown that the two last models produce the same OCC models. In fact, a framework for constructing robust imprecise OCC is proposed in the paper.

The paper is organized as follows. Section 2 presents the standard OCC problem proposed by Scholkopf et al. [30,32]. Robust models used in classification are considered in Section 3. The problem of the OCC under sets of probability distributions is stated in the same section. The OCC SVMs under sets of probability distributions using the minimax and minimin strategies are given in Section 4. It is shown in this section that the OCC model is determined by extreme points of a set of probability distributions. Extreme points of the imprecise linear-vacuous mixture, pari-mutuel and constant odds-ratio models are obtained in Section 5. Section 6 provides extreme points of the bounded ε-contaminated robust model and Kolmogorov-Smirnov bounds. Numerical experiments with synthetic and real data illustrating accuracy of the proposed models are given in Section 7. In Section 8, concluding remarks are made.

2 One-class classification

Suppose we have unlabeled training data $\mathbf{x}_1,...,\mathbf{x}_n \subset \mathcal{X}$, where n is the number of observations, $\mathcal{X}$ is some set, for instance, it is a compact subset of $\mathbb{R}^m$. Let $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$ be drawn i.i.d. from a distribution on $\mathcal{X}$. The sample space $\mathcal{D}$ is finite and discrete. According to papers [30,32], a well-known novelty detection or OCC model aims to construct a function f which takes the value $+1$ in a "small" region capturing most of the data points and -1 elsewhere. It can be done by mapping the data into the feature space corresponding to a kernel and by separating them from the origin with maximum margin.

Let ϕ be a feature map $\mathcal{X} \to G$ such that the data points are mapped into an alternative higher-dimensional feature space G. In other words, this is a map into an inner product space G such that the inner product in the image of ϕ can be computed by evaluating some simple kernel $K(\mathbf{x},\mathbf{y}) = (\phi(\mathbf{x}),\phi(\mathbf{y}))$, such as the Gaussian kernel

$$K(\mathbf{x},\mathbf{y}) = \exp\left(-\left\|\mathbf{x}-\mathbf{y}\right\|^2/\sigma^2\right).$$

σ is the kernel parameter determining the geometrical structure of the mapped samples in the kernel space. It is pointed out by [42] that the problem of selecting a proper parameter σ is very important in classification. When a very small σ is used ($\sigma \to 0$), $K(\mathbf{x},\mathbf{y}) \to 0$ for all $\mathbf{x} \neq \mathbf{y}$ and all mapped samples tend to be orthogonal to each other, despite their class labels. In this case, both between-class and within-class variations are very large. On the other hand, when a very large σ is chosen ($\sigma^2 \to \infty$), $K(\mathbf{x},\mathbf{y}) \to 1$ and all mapped samples converge to a single point. This obviously is not desired in a classification task. Therefore, a too large or too small σ will not result in more separable samples in G.

It is shown [7] that the data points lie on the surface of a hypersphere in feature space since $\phi(\mathbf{x}) \cdot \phi(\mathbf{x}) = K(\mathbf{x},\mathbf{x}) = 1$ (translation invariant kernels). Now we have to find a hyperplane $f(\mathbf{x},\mathbf{w}) = \langle \mathbf{w},\phi(\mathbf{x}_i)\rangle - \rho = 0$ that separates the data from the origin with maximal margin, i.e., we want ρ to be as large as possible so that the volume of the halfspace $\langle \mathbf{w},\phi(\mathbf{x}_i)\rangle \geq \rho$ is minimized. Let us introduce the parameter $\nu \in [0;1]$ which is analogous to ν used for the ν-SVM [12,31]. Roughly speaking, it denotes the fraction of input data for which $\langle \mathbf{w},\phi(\mathbf{x}_i)\rangle \leq \rho$. To separate the data set from the origin, we solve the following quadratic program:

$$\min_{w,\xi,\rho} \frac{1}{2}\left\|\mathbf{w}\right\|^2 + \frac{1}{\nu n}\sum_{i=1}^n \xi_i - \rho,$$

subject to

$$\langle \mathbf{w},\phi(\mathbf{x}_i)\rangle \geq \rho - \xi_i,\ \xi_i \geq 0,\ i = 1,...,n.$$

Slack variables ξ_i are used to allow points to violate margin constraints.

Since nonzero slack variables ξ_i are penalized in the objective function, we can expect that if $\mathbf{w}$ and ρ solve this problem, then the decision function

$$f(\mathbf{x},\mathbf{w}) = \operatorname{sgn}\left(\langle \mathbf{w},\phi(\mathbf{x})\rangle - \rho\right)$$

will be positive for most examples $\mathbf{x}_i$ contained in the training set, while the SV type regularization term $\|\mathbf{w}\|$ will still be small. The actual trade-off between these two goals is controlled by ν.

Using multipliers $\alpha_i, \beta_i \geq 0$, we introduce a Lagrangian

$$L(\mathbf{w},\xi,\rho,\alpha,\beta) = \frac{1}{2}\left\|\mathbf{w}\right\|^2 + \frac{1}{\nu n}\sum_{i=1}^n \xi_i - \rho - \sum_{i=1}^n \alpha_i\left(\langle \mathbf{w},\phi(\mathbf{x}_i)\rangle - \rho + \xi_i\right) - \sum_{i=1}^n \beta_i\xi_i.$$

It is shown that the dual problem is of the form:

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j K(\mathbf{x}_i,\mathbf{x}_j),$$

subject to

$$0 \leq \alpha_i \leq \frac{1}{\nu n},\ \sum_{i=1}^n \alpha_i = 1.$$

The value of ρ can be obtained as

$$\rho = (\langle \mathbf{w},\phi(\mathbf{x}_j)\rangle) = \sum_{i=1}^n \alpha_i K(\mathbf{x}_i,\mathbf{x}_j).$$

After substituting the obtained solution into the expression for the decision function f, we get

$$f(\mathbf{x},\mathbf{w}) = \operatorname{sgn}\left(\sum_{i=1}^n \alpha_i K(\mathbf{x}_i,\mathbf{x}) - \rho\right).$$

3 Robust models and sets of probability distributions

Robust models have been exploited in classification problems due to the opportunity to avoid some strong assumptions underlying the standard classification models. As pointed out by Xu et al [45], the use of robust optimization in classification is not new. There are a lot of published results providing various robust classification and regression models (see, for instance, [6,9,15, 22–24,26,28,37,46]) in which box-type uncertainty sets are considered.

One of the popular robust classification models is based on the assumption that inputs are subject to an additive noise, i.e., $\mathbf{x}_i^* = \mathbf{x}_i + \triangle\mathbf{x}_i$, where noise $\triangle\mathbf{x}_i$ is governed by a certain distribution. The simplest way for dealing with noise is to consider a simple bounded uncertainty model $\|\triangle\mathbf{x}_i\| \leq \delta_i$ with uniform priors. According to this model, the data is uncertain, specifically, for every i, i-th "true" data point is only known to belong to the interior of an Euclidian ball of radius δ_i centered at the "nominal" data point $\mathbf{x}_i$. This model has a very clear intuitive geometric interpretation [2]. The maximally robust classifier in this case is the one that maximizes the radius of balls, i.e., it corresponds to the largest radius such that the corresponding balls around each data point are still perfectly separated. Applying the above ideas to the SVM with the hinge loss function provides the following optimization problem (see, for instance, a similar problem for binary classification problem [4])

$$\min_{w,\xi,\rho,\triangle\mathbf{x}_i} \frac{1}{2}\left\|\mathbf{w}\right\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i,$$

subject to

$$\langle \mathbf{w},\phi(\mathbf{x}_i + \triangle\mathbf{x}_i)\rangle \geq \rho - \xi_i,\ \xi_i \geq 0,$$

$$\left\|\triangle\mathbf{x}_i\right\|^2 \leq \delta_i,\ i = 1,...,n.$$

It is often assumed that each point can move around within a box ($\|\triangle\mathbf{x}_i\| \leq \delta_i$). In this case, we have linear constraints.

Another class of robust models is based on relaxing strong assumptions about a probability distribution of data points (see, for instance, [22]). We propose to use a model which can be partially regarded as a special case of these models and is based on using the framework of ε-*contaminated (robust) models* ([18, Section 4.2.]). It is constructed by eliciting a Bayesian prior distribution $p = (p_1, ..., p_n)$ over the sample space $\mathcal{D}$ as an estimate of the true prior distribution. The ε-contaminated model or linear-vacuous mixture [41, Subsection 2.9.2.] is a class $\mathcal{M}(\varepsilon, p)$ of probabilities $\pi = (\pi_1, ..., \pi_n)$ such that for fixed $\varepsilon \in (0, 1)$ and p_i there holds $\mathcal{M}(\varepsilon, p) = \{(1-\varepsilon)p_i + \varepsilon q_i\}$, where q_i is arbitrary and $q_1 + ... + q_n = 1$, i.e., $\pi_i = (1-\varepsilon)n^{-1} + \varepsilon q_i$. The rate ε reflects how "close" we feel that π must be to p [3, Subsection 3.5.3.]. In other words, we take an arbitrary probability distribution $q = (q_1, ..., q_n)$ from the unit simplex. According to these models, for $0 < \varepsilon < 1$, $\mathcal{M}(\varepsilon, p)$ is the set of all probabilities with the lower bound $(1-\varepsilon\}p_i$ and the upper bound $(1-\varepsilon\}p_i + \varepsilon$. Of course, the assumption that q is restricted by the unit simplex is one of possible types of ε-contaminated models. Generally, there are a lot of different assumptions which produce specific robust models and which will be studied below.

Let us rewrite the problem in a general form of minimizing the expected risk [40, Section 1.2.]

$$R(\mathbf{w},\rho) = \int_{\mathbb{R}^m} L(\mathbf{x})\mathrm{d}F_0(\mathbf{x}).$$

Here the loss function $L(\mathbf{x})$ can be represented as

$$L(\mathbf{x}) = \max\{0, \rho - \langle \mathbf{w}, \phi(\mathbf{x})\rangle\} - \rho\nu.$$

The standard SVM technique is to assume that F_0 is empirical (non-parametric) probability distribution whose use leads to the empirical expected risk

$$R_{\mathrm{emp}}(\mathbf{w},\rho) = \frac{1}{n}\sum_{i=1}^{n} L(\mathbf{w}, \phi(\mathbf{x}_i)). \tag{1}$$

The assumption of the empirical probability distribution means that every point $\mathbf{x}_i$ has the probability $p_i = 1/n$. This is a rather strong assumption when the number of points is not large. Its validity might give rise to doubt in this case [44]. Therefore, in order to relax the strong condition for probabilities of points, we apply the ε-contaminated model. According to the model, we replace the probability distribution $p = (1/n, ..., 1/n)$ by the set of probability distributions $\mathcal{M}(\varepsilon, p)$. In other words, there is an unknown precise "true" probability distribution in $\mathcal{M}(\varepsilon, p)$, but we do not know it and only know that it belongs to the set $\mathcal{M}(\varepsilon, p)$. While some robust models [2] assume that each point can move around within an Euclidean ball, the proposed robust model assumes that the probability $1/n$ of each point (but not a data point itself) can move around within a unit simplex under some restrictions. This is the main idea for constructing the robust OCC models below. The notation $\mathcal{M}(\varepsilon)$ is used below instead of $\mathcal{M}(\varepsilon, p)$ for short. Moreover, we will consider $\mathcal{M}(\varepsilon)$ as an arbitrary convex set of probability distributions without pointing out that it is produced by the ε-contaminated model.

One of the possible ways for dealing with the set $\mathcal{M}(\varepsilon)$ of probability distributions produced by the above constraints is to use the minimax (pessimistic) strategy. According to the minimax strategy, we select a probability distribution from the set $\mathcal{M}(\varepsilon)$ such that the expected risk $R(\mathbf{w},\rho)$ achieves its maximum for every fixed $\mathbf{w}$. It should be noted that the "optimal" probability distributions may be different for different values of parameters $\mathbf{w}$. The minimax strategy can be explained in a simple way. We do not know a precise probability distribution and every distribution from $\mathcal{M}(\varepsilon)$ can be selected. Therefore, we should take the "worst" distribution providing the largest value of the expected risk. The minimax criterion can be interpreted as an insurance against the worst case because it aims at minimizing the expected loss in the least favorable case [29, Subsection 2.4.2.]. This criterion of decision making can be regarded as the well-known Γ-minimax[1] [3,16,38].

Let $h=(h_1,...,h_n)$ be a probability distribution which belongs to the set $\mathcal{M}(\varepsilon)$. The maximum value of the expected risk $R(\mathbf{w},\rho)$ is

$$\overline{R}(\mathbf{w},\rho)=\max_{h\in\mathcal{M}(\varepsilon)}R(\mathbf{w},\rho).$$

The minimax expected risk with respect to the minimax strategy is now of the form:

$$\overline{R}(\mathbf{w}_{\text{opt}},\rho_{\text{opt}})=\min_{\mathbf{w},\rho}\overline{R}(\mathbf{w},\rho)=\min_{\mathbf{w},\rho}\max_{h\in\mathcal{M}(\varepsilon)}R(\mathbf{w},\rho).$$

The upper bound for the expected risk can be found as a solution to the following programming problem:

$$\overline{R}(\mathbf{w},\rho)=\max_{h}\sum_{i=1}^{n}h_iL(\mathbf{w},\phi(\mathbf{x}_i))-\rho\nu,$$

subject to

$$h\in\mathcal{M}(\varepsilon). \tag{2}$$

The obtained optimization problem is linear with optimization variables $h_1,...,h_n$, but the objective function depends on $\mathbf{w}$. Therefore, it can not be directly solved by well-known methods. In order to overcome this difficulty, note, however, that all points h belong to a set of distributions $\mathcal{M}(\varepsilon)$ which has t extreme points denoted $H^{(k)}\in\operatorname{extr}(\mathcal{M}(\varepsilon))$, $k=1,...,t$, $H^{(k)}=\left(h_1^{(k)},...,h_n^{(k)}\right)$. According to some general results from linear programming theory, an optimal solution to the above problem is achieved at extreme points of the simplex, and the number of its extreme points is t.

This implies that there holds

$$\overline{R}(\mathbf{w},\rho)=\max_{k=1,...,t}\left(\sum_{i=1}^{n}h_i^{(k)}L(\mathbf{w},\phi(\mathbf{x}_i))\right)-\rho\nu. \tag{3}$$

The next task is to minimize the upper expected risk $\overline{R}(\mathbf{w},\rho)$ over the parameters $\mathbf{w}$ and ρ. This task will be solved in the framework of the SVM.

4 The minimax strategy and the SVM

Let us add the standard Tikhonov regularization term $\frac{1}{2}\langle\mathbf{w},\mathbf{w}\rangle$ (this is the most popular penalty or smoothness term) [36] to the objective function (3). The smoothness (Tikhonov) term can be regarded as a constraint which enforces uniqueness by penalizing functions with wild oscillation and effectively restricting the space of admissible solutions (we refer to [13] for a detailed analysis of regularization methods). Moreover, we introduce the following optimization variables:

$$\xi_i=\max\{0,\rho-\langle\mathbf{w},\phi(\mathbf{x}_i)\rangle\},\ G=\max_{k=1,...,t}\sum_{i=1}^{n}h_i^{(k)}\xi_i. \tag{4}$$

This leads to the quadratic programming problem

$$\overline{R}(\mathbf{w}_{\text{opt}},\rho_{\text{opt}})=\min_{\mathbf{w},\rho,G,\xi_i}\left(\frac{1}{2}\langle\mathbf{w},\mathbf{w}\rangle+G-\rho\nu\right), \tag{5}$$

subject to

$$\xi_i\geq\rho-\langle\mathbf{w},\phi(\mathbf{x}_i)\rangle,\ \xi_i\geq 0,\ i=1,...,n,$$

$$G\geq\sum_{i=1}^{n}h_i^{(k)}\xi_i,\ k=1,...,t.$$

Instead of minimizing the primary objective function (5), a dual objective function, the so-called Lagrangian, can be formed of which the saddle point is the optimum. The Lagrangian is

$$\begin{aligned}L=&\frac{1}{2}\langle\mathbf{w},\mathbf{w}\rangle+G-\rho\nu\\&-\sum_{i=1}^{n}\lambda_i\xi_i-\sum_{i=1}^{n}\varphi_i\left(\xi_i-\rho+\langle\mathbf{w},\phi(\mathbf{x}_i)\rangle\right)\\&-\sum_{k=1}^{t}\eta_k\left(G-\sum_{i=1}^{n}h_i^{(k)}\xi_i\right).\end{aligned}$$

Here $\lambda_i,\eta_i,\varphi_i$, $i=1,...,n$, are Lagrange multipliers. Hence, the dual variables have to satisfy positivity constraints $\eta_i\geq 0,\varphi_i\geq 0,\lambda_i\geq 0$ for all $i=1,...,n$. The saddle point can be found by setting the derivatives equal to zero

$$\partial L/\partial\rho=-\nu+\sum_{i=1}^{n}\varphi_i=0, \tag{6}$$

$$\partial L/\partial G=1-\sum_{k=1}^{t}\eta_k=0, \tag{7}$$

[1] This criterion is often given in terms of utilities. Therefore, it is usually called Γ-maximin.

$$\partial L/\partial \xi_i = -\lambda_i - \varphi_i + \sum_{k=1}^{t} \eta_k h_i^{(k)} = 0, \tag{8}$$

$$\partial L/\partial w_j = w_j - \sum_{i=1}^{n} \varphi_i \phi(x_j^{(i)}) = 0,\ j = 1, ..., m. \tag{9}$$

Here $x_j^{(i)}$ is the value of the j-th feature of the i-th example.

Using (6)-(8), we simplify the objective function as

$$L = \frac{1}{2}\left\langle \mathbf{w}, \mathbf{w} \right\rangle - \sum_{i=1}^{n} \varphi_i \cdot \left\langle \mathbf{w}, \phi(\mathbf{x}_i) \right\rangle .$$

Finally, substituting w_j from (9), we get the following dual optimization problem

$$\max_{\varphi_i, \eta_i} \left(-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \varphi_i \varphi_j K(\mathbf{x}_i, \mathbf{x}_j) \right), \tag{10}$$

subject to

$$0 \leq \varphi_i \leq \sum_{k=1}^{t} \eta_k h_i^{(k)},\ i = 1, ..., n, \tag{11}$$

$$\sum_{i=1}^{n} \varphi_i = \nu,\ \sum_{k=1}^{t} \eta_k = 1,\ \eta_i \geq 0,\ i = 1, ..., n. \tag{12}$$

The function $f(\mathbf{x})$ can be rewritten in terms of Lagrange multipliers as

$$f(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{n} \varphi_i K(\mathbf{x}_i, \mathbf{x}) - \rho.$$

Hence, we find the optimal value of ρ by taking $f(\mathbf{x}, \mathbf{w}) = 0$, i.e., there holds

$$\rho = \sum_{i=1}^{n} \varphi_i K(\mathbf{x}_i, \mathbf{x}_j).$$

Let us consider how the above problem can be modified in the "precise" case when we have a single precise nonparametric probability distribution. In this case, we can write $t = 1$, $\mathcal{M}(0) = (1/n, ..., 1/n)$. Hence, we get $\sum_{k=1}^{t} \eta_k h_i^{(k)} = 1/n$ and the constraints for φ_i become

$$0 \leq \varphi_i \leq 1/n,\ \sum_{i=1}^{n} \varphi_i = \nu.$$

This indeed gives the standard SVM. So, we get the SVM approach under the minimax strategy taking into account the introduced upper bounds.

In case of complete ignorance, we can write $\varepsilon = 1$ and the set $\mathcal{M}(1)$ is the unit simplex. This implies that $\sum_{k=1}^{t} \eta_k h_i^{(k)} = \eta_i$ for every $i = 1, ..., n$, and a single data point $\mathbf{x}_i$ with the largest loss function $L(\mathbf{w}, \phi(\mathbf{x}_i))$ completely defines the decision function f.

Proposition 1 *Optimization problem (10)-(12) can be represented as a set of t quadratic programming problems with the objective function*

$$\max_{\varphi_i} \left(-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \varphi_i \varphi_j K(\mathbf{x}_i, \mathbf{x}_j) \right), \tag{13}$$

and constraints

$$0 \leq \varphi_i \leq h_i^{(s)},\ i = 1, ..., n,\ \sum_{i=1}^{n} \varphi_i = \nu. \tag{14}$$

The optimal solution corresponds to the smallest value of the objective function (13).

Proof Let us write the Karush-Kuhn-Tucker complementarity conditions

$$\varphi_i \left(\xi_i - \rho + \left\langle \mathbf{w}, \phi(\mathbf{x}_i) \right\rangle \right) = 0,$$

$$\eta_k \left(G - \sum_{i=1}^{n} h_i^{(k)} \xi_i \right) = 0.$$

It follows from the second condition and from (4) that $G = \sum_{i=1}^{n} h_i^{(k)} \xi_i$ for a single value of k, say $k = s$, such that $s = \arg\max_{k=1,...,n} \sum_{i=1}^{n} h_i^{(k)} \xi_i$. First, we assume that values $\sum_{i=1}^{n} h_i^{(k)} \xi_i$ do not coincide for different k. This implies that $G \neq \sum_{i=1}^{n} h_i^{(k)} \xi_i$ and, therefore, $\eta_k = 0$ for all $k \neq s$. In other words, we have optimal vectors of variables η_k

$$(\eta_1, 0, ..., 0),\ (0, \eta_2, ..., 0), ..., (0, ..., 0, \eta_t).$$

It follows from (12) that $\eta_s = 1$, and we get new optimal vectors of variables η_k

$$(1, 0, ..., 0),\ (0, 1, ..., 0), ..., (0, ..., 0, 1).$$

Hence, we can write

$$\sum_{k=1}^{t} \eta_k h_i^{(k)} = h_i^{(s)}.$$

Returning to the constraints (11)-(12) and substituting the above optimal vectors of variables η_k into (11), we get constraints (14) for every $s = 1, ..., t$. Now we assume that there are two numbers k and l such that $G = \sum_{i=1}^{n} h_i^{(k)} \xi_i = \sum_{i=1}^{n} h_i^{(l)} \xi_i$. Then every optimal vector of variables η_i is of the form:

$$(0, .., \eta_k, 0, ..., \eta_l, ..., 0).$$

It follows from (12) that $\eta_k + \eta_l = 1$ and $\eta_k \leq 1$, $\eta_l \leq 1$. The right side of the constraint (11) for φ_i is $\eta_k h_i^{(k)} + \eta_l h_i^{(l)}$. In order to maximize the objective function L the values of φ_i should be taken as large as possible under condition $K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$, i.e., the sum $\eta_k h_i^{(k)} + \eta_l h_i^{(l)}$

has to be maximized. Its largest value is achieved when $\eta_k = 1$ or $\eta_l = 1$. This implies that we return to the first case when $G = \sum_{i=1}^{n} h_i^{(k)}\xi_i$ for a single value of k. The case of an arbitrary number of coinciding terms $\sum_{i=1}^{n} h_i^{(k)}\xi_i$ is proved in the same case.

Finally, we have to solve t quadratic programming problems with constraints (14) and with the same objective functions L. If we denote the optimal value of the objective function by a fixed value of s as L_s, then the optimal values of φ_i, $i = 1, ..., n$, correspond to the *smallest* value of objective function L_s, $s = 1, ..., t$.

If we have a precise probability distribution h of learning points, then $h_i^{(s)} = h_i$ for all $s = 1, ..., t$, all optimization problems are identical, and we get the standard weighted SVM [5,47]. It follows from (14) that a weight h_i close to zero forces the corresponding φ_i to also be close to zero, thus contributing very weakly to the definition of the decision function.

The minimax strategy is one of the possible decision strategies which can be used by dealing with sets of probability distributions. A direct opposite to this strategy is the minimin strategy. According to the minimin strategy, the expected risk R is minimized over all probability distributions from the set $\mathcal{M}(\varepsilon)$ as well as over all values of parameters $\mathbf{w}$, ρ. The strategy can be called optimistic because it selects the "best" probability distribution from the set $\mathcal{M}(\varepsilon)$. In spite of the fact that the minimin strategy is not robust, it is interesting to consider how the SVM based on this strategy differs from the "maximin" SVM.

Similarly to the minimax strategy, we can write

$$\underline{R}(\mathbf{w},\rho) = \min_{h\in\mathcal{M}(\varepsilon)} R(\mathbf{w},\rho).$$

The lower bound for the expected risk can be found as a solution to the following programming problem (see the quite similar derivation for the minimax strategy):

$$\underline{R}(\mathbf{w},\rho) = \min_{h\in\mathcal{M}(\varepsilon)} \left(\sum_{i=1}^{n} h_i L(\mathbf{w}, \phi(\mathbf{x}_i)) - \rho\nu \right),$$

subject to (2).

Hence, there holds

$$\underline{R}(\mathbf{w},\rho) = \min_{k=1,...,t} \left(\sum_{i=1}^{n} h_i^{(k)} L(\mathbf{w}, \phi(\mathbf{x}_i)) - \rho\nu \right).$$

The next task is to minimize the lower expected risk $\underline{R}(\mathbf{w},\rho)$ over the parameters $\mathbf{w}$ and ρ. In order to solve this task, we have to solve t quadratic optimization problems of the form:

$$\underline{R}_k(\mathbf{w}_{\text{opt}},\rho_{\text{opt}}) = \min_{\mathbf{w},\rho,\xi_i} \left(\frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle + \sum_{i=1}^{n} h_i^{(k)}\xi_i - \rho\nu \right),$$

subject to

$$\xi_i \geq \rho - \langle \mathbf{w}, \phi(\mathbf{x}) \rangle, \ \xi_i \geq 0, \ i = 1, ..., n.$$

The optimal values of the optimization variables $\mathbf{w}$,ρ, ξ_i, $i = 1, ..., n$, correspond to the *smallest* value of objective function $\underline{R}_k$, $k = 1, ..., t$. The smallest value of the expected risk $\underline{R}_k$ corresponds to the *largest* value of the Lagrangian. Hence, the following dual optimization problem can be derived:

$$L_k = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \varphi_i\varphi_j K(\mathbf{x}_i, \mathbf{x}_j)$$

subject to

$$0 \leq \varphi_k \leq h_i^{(k)}, \ i = 1, ..., n, \ \sum_{i=1}^{n} \varphi_i = \nu.$$

The optimal values of φ_i, $i = 1, ..., n$, correspond to the *largest* value of the objective function L_k, $k = 1, ..., t$, where $t = n$.

One can see from the above that we get quite the same quadratic optimization problems as for the minimax strategy. However, in contrast to the minimax strategy, we search for the largest value of the Lagrangian.

5 Important special imprecise neighbourhood models

We briefly consider in this section some typical cases of the initial information about the set $\mathcal{M}$ which can be used in classification. First of all, we analyze imprecise neighbourhood models. It has been indicated by Walley [41, Subsection 4.6.5.], neighbourhood models are most appropriate when there is substantial evidence which supports an additive probability p but there is incomplete confidence in p. In practice it is difficult to assess additive probabilities and we prefer models which do not require them. Rather than making precise assessments to determine p and adding imprecision by forming a neighbourhood, it is easier to directly elicit a model through imprecise assessments.

We consider how to apply neighbourhood models to the OCC problem, i.e., we show how the constraints to the dual quadratic programming problem have to be changed. The objective function remains without changes.

Every imprecise model is characterized by a common parameter ε. Usually ε measures the probability that the elicited p is incorrect. There are other interpretations of the parameter ε. For example, it is the size of possible errors in p or the amount of information on which the model is based.

5.1 The linear-vacuous mixture

The linear-vacuous mixture or ε-contaminated models have been studied in Section 3. They produce the set $\mathcal{M}(\varepsilon)$ of probabilities.

Proposition 2 *The set $\mathcal{M}(\varepsilon)$ of probabilities under condition $p = (1/n, ..., 1/n)$ has n extreme points. The extreme points consist of $n-1$ elements $(1-\varepsilon)n^{-1}$ and one element $(1-\varepsilon)n^{-1}+\varepsilon$, i.e., they are of the form:*

$$\left(\frac{1-\varepsilon}{n}, ..., \frac{1-\varepsilon}{n}+\varepsilon, ..., \frac{1-\varepsilon}{n}\right).$$

Proof It follows from the fact that the unit simplex has extreme points consisting of $n-1$ elements 0 and one element 1.

In sum, we get n optimization problems (13)-(14) such that constraints (14) are of the form:

$$0 \leq \varphi_i \leq \frac{1-\varepsilon}{n},\ i \neq s,$$
$$0 \leq \varphi_s \leq \frac{1-\varepsilon}{n}+\varepsilon.$$

5.2 Pari-mutuel models

Different types of models can be constructed in the framework of neighbourhood. One of the models is Walley's imprecise pari-mutuel model [41, Subsection 3.3.5.] for which the set of probability distributions is defined as follows:

$$\mathcal{M}_P(\varepsilon) = \{\pi : \pi_i \leq (1+\varepsilon)p_i\}.$$

The above class is such that the probabilities of events do not exceed a constant multiple of p_i. The set $\mathcal{M}(\varepsilon)$ can be obtained from $\mathcal{M}_P(\varepsilon)$ by reflecting $\mathcal{M}(\varepsilon)$ about the point π. Lower and upper probabilities of the i-th point are $\max\{(1+\varepsilon)p_i-\varepsilon, 0\}$ and $(1+\varepsilon)p_i$, respectively. The difference between the upper and lower probabilities is ε. It is constant over all events for which p_i is not too close to 0 or 1.

Let us consider the extreme points of the set $\mathcal{M}_P(\varepsilon)$ under condition that the empirical distribution is $p = (1/n, ..., 1/n)$.

Proposition 3 *The set $\mathcal{M}_P(\varepsilon)$ under condition $p = (1/n, ..., 1/n)$ has n extreme points. The extreme points consist of $n-1$ elements $(1+\varepsilon)n^{-1}$ and one element $(1+\varepsilon)n^{-1}-\varepsilon$, i.e., they are of the form:*

$$\left(\frac{1+\varepsilon}{n}, ..., \frac{1+\varepsilon}{n}-\varepsilon, ..., \frac{1+\varepsilon}{n}\right).$$

Proof It follows from the definition of extreme points that an extreme point is an intersection of n linearly independent (bounding) hyperplanes. This implies that a subset of n of the hyperplanes are selected and their equations are simultaneously solved in order to find an extreme point. One of the equations is $\pi_1 + ... + \pi_n = 1$. Hence, we select $n-1$ equations of the form $\pi_i = (1+\varepsilon)/n$ and solve them. The following proof is obvious.

The pari-mutuel model has a restriction on the values of ε. Its value has to satisfy the condition $(1+\varepsilon)/n - \varepsilon \geq 0$. Hence, there holds $\varepsilon \leq 1/(n-1)$. If this condition is not satisfied, then $\mathcal{M}_P(\varepsilon)$ is empty.

In sum, we get n optimization problems (13)-(14) such that constraints (14) are of the form:

$$0 \leq \varphi_i \leq \frac{1+\varepsilon}{n},\ i \neq s,$$
$$0 \leq \varphi_s \leq \frac{1+\varepsilon}{n}-\varepsilon.$$

5.3 Constant odds-ratio models

Another type of neighbourhood models is the so-called constant odds-ratio (π, ε) model. The set of probability distribution for the model is defined as

$$\mathcal{M}_C(\varepsilon) = \{\pi : \pi_i/\pi_j \geq (1-\varepsilon)p_i/p_j\}.$$

Here i and j are arbitrary numbers of points from 1 to n, $\varepsilon \in [0,1)$. The probability π_j can not be 0 because if $\pi_j = 0$ and $p_i/p_j > 0\ \forall i, j$, then we can always find the non-zero probability π_k such that $\pi_j/\pi_k = 0$ by $k \neq j$ and π_j/π_k does not satisfy the model condition.

The upper and lower probabilities of the i-th point are written as

$$\overline{\pi}_i = \frac{\pi_i}{1-\varepsilon(1-\pi_i)},\ \underline{\pi}_i = \frac{(1-\varepsilon)\pi_i}{1-\varepsilon(1-\pi_i)}.$$

Walley [41, Subsection 2.9.4.] indicates that the constant odds-ratio model is convenient for statistical applications because its form is not changed by conditioning or statistical updating. If the constant odds-ratio (π, ε) model is used to represent prior beliefs about a statistical parameter, and statistical data are obtained, then posterior beliefs are represented by the constant odds-ratio (π_1, ε) model, where π_1 is the posterior of π defined by Bayes' rule.

Let us consider the extreme points of the set $\mathcal{M}_C(\varepsilon)$ under condition $p = (1/n, ..., 1/n)$.

Proposition 4 *The set $\mathcal{M}_C(\varepsilon)$ under condition $p = (1/n, ..., 1/n)$ has $2n$ extreme points. The first n points are*

$$\left(\frac{1-\varepsilon}{n(1-\varepsilon)+\varepsilon}, ..., \frac{1}{n(1-\varepsilon)+\varepsilon}, ..., \frac{1-\varepsilon}{n(1-\varepsilon)+\varepsilon}\right),$$

(15)

where the element $(n(1-\varepsilon)+\varepsilon)^{-1}$ in the k-th extreme point is located on the k-th position. The second n points are

$$\left(\frac{1}{n-\varepsilon},...,\frac{1-\varepsilon}{n-\varepsilon},...,\frac{1}{n-\varepsilon}\right), \tag{16}$$

where the element $(1-\varepsilon)/(n-\varepsilon)$ in the k-th extreme point is located on the k-th position.

Proof It follows from the definition of the set $\mathcal{M}_C(\varepsilon)$ that it is produced by the set of inequalities

$$\pi_i \geq (1-\varepsilon)\pi_j,\ \forall i,j=1,...,n,$$

and one equality $\pi_1+...+\pi_n=1$. Every extreme point of $\mathcal{M}_C(\varepsilon)$ is determined by considering $n-1$ equalities $\pi_i=(1-\varepsilon)\pi_j$. Without loss of generality, we take $n-1$ equalities of the form

$$\pi_i=(1-\varepsilon)\pi_1,\ i=2,...,n.$$

By substituting the above into the condition for the sum of probabilities, we get

$$\pi_1+(n-1)(1-\varepsilon)\pi_1=1.$$

Hence there holds

$$\pi_1=\frac{1}{n(1-\varepsilon)+\varepsilon},\ \pi_i=\frac{1-\varepsilon}{n(1-\varepsilon)+\varepsilon},\ i=2,...,n.$$

It is easy to prove that the extreme points are of the form (15) and their number is n. Let us take another subset of equalities for looking for extreme points

$$\pi_1=(1-\varepsilon)\pi_i,\ i=2,...,n.$$

By substituting the above into the condition for the sum of probabilities, we get

$$\pi_1\left(1+\frac{n-1}{1-\varepsilon}\right)=1.$$

Hence there holds

$$\pi_1=\frac{1-\varepsilon}{n-\varepsilon},\ \pi_i=\frac{1}{n-\varepsilon},\ i=2,...,n.$$

In sum, the next n extreme points are of the form (16). Let us prove that there are no more extreme points. Suppose that there are different extreme points. Then constraints produced every extreme point has to contain the following equalities

$$\pi_k=(1-\varepsilon)\pi_j,\quad \pi_j=(1-\varepsilon)\pi_l.$$

Here $k\neq j$ and $j\neq l$. It follows from the above conditions that

$$\pi_k=(1-\varepsilon)^2\pi_l.$$

At the same time, the corresponding point belongs to the set $\mathcal{M}_C(\varepsilon)$ if the condition $\pi_k\geq(1-\varepsilon)\pi_l$ is fulfilled. However, it is not fulfilled because

$$\pi_k=(1-\varepsilon)^2\pi_l\leq(1-\varepsilon)\pi_l.$$

We get the contradiction. Therefore, we have $2n$ extreme points (15) and (16).

It should be noted that we have only two extreme points when $n=2$ because the points obtained by means of (15) and (16) coincide.

In sum, we get n optimization problems (13)-(14) such that constraints (14) for $s=1,...,n$ are of the form:

$$0\leq\varphi_i\leq\frac{1-\varepsilon}{n(1-\varepsilon)+\varepsilon},\ i\neq s,$$

$$0\leq\varphi_s\leq\frac{1}{n(1-\varepsilon)+\varepsilon},$$

constraints (14) for $s=n+1,...,2n$ are of the form:

$$0\leq\varphi_i\leq\frac{1}{n-\varepsilon},\ i\neq s-n,$$

$$0\leq\varphi_{s-n}\leq\frac{1-\varepsilon}{n-\varepsilon}.$$

By analyzing the above models (linear-vacuous mixture, pari-mutuel, constant odds-ratio), we can see that extreme points have $n-1$ identical elements and a single different element. This implies that we choose one point from the training set and assign a smaller or larger weight to this point. Of course, weights of other points are also changed. But they remain identical.

6 Bounded contaminated model and Kolmogorov-Smirnov bounds

We consider three interesting uncertainty models in this section, which are quite different, but two of them provide the same extreme points for their use in the weighted SVM. In other words, these uncertainty models produce the same classification models.

6.1 Upper bounded vacuous and bounded ε-contaminated robust models

The set $\mathcal{M}(1)$ in the linear-vacuous mixture by $\varepsilon\to 1$ can be regarded as a "vacuous" model because it is not informative and is of the largest size. As a result, almost all data points in the training set by the minimax or minimin strategies and by $\varepsilon\to 1$ have probabilities 0 except for a single point. In order to reduce the set $\mathcal{M}(1)$ we consider a simple generalization of the vacuous model by introducing upper bounds for the probabilities $h_1, ..., h_n$, namely, we state that $h_i \leq \vartheta$, $i = 1, ..., n$, where ϑ is some number between $1/n$ and 1. We denote the reduced set of probability distributions as $\mathcal{M}_U(\vartheta)$. The lower possible bound for ϑ is $1/n$ because the set $\mathcal{M}_U(\vartheta)$ becomes empty when $\vartheta < 1/n$.

In sum, we can state that the set $\mathcal{M}_U(\vartheta)$ is produced by the following system of $2n$ inequalities and one equality:

$$0 \leq h_i \leq \vartheta, \ i = 1, ..., n, \ h_1 + ... + h_n = 1. \tag{17}$$

It is not difficult to find all extreme points of the set $\mathcal{M}_U(\vartheta)$, which strongly depend on the value ϑ.

Proposition 5 *Let k be an integer from 1 to $n-1$ such that the following condition is fulfilled:*

$$\frac{1}{n-k+1} < \vartheta \leq \frac{1}{n-k}. \tag{18}$$

Then the set $\mathcal{M}_U(\vartheta)$ has $t = k \cdot \binom{n}{k}$ extreme points such that every extreme point consists of exactly $n-k$ elements ϑ, $k-1$ elements 0 and one element $1-(n-k)\vartheta$.

Proof It follows from the system of $2n$ inequalities (17) that every extreme point satisfies $n-1$ equalities $h_i = 0$ (their number is, say, s) and $h_i = \vartheta$ (their number is $n-s-1$). The next question is how to find the value s. Without loss of generality, we suppose that $h_i = \vartheta$ for $i = 1, ..., n-s-1$, and $h_i = 0$ for $i = n-s, ..., n-1$. Hence, we can write

$$\sum_{i=1}^{n-s-1} h_i = \vartheta(n-s-1) \leq 1.$$

Condition (18) implies that

$$\frac{n-s-1}{n-k+1} < \vartheta(n-s-1) \leq 1.$$

Hence, there holds $n-s-1 \leq n-k$. We also can write

$$h_n + \sum_{i=1}^{n-s-1} h_i \leq \vartheta + \vartheta(n-s-1) \leq \frac{n-s}{n-k}.$$

Hence, there holds $n-s-1 \geq n-k-1$. So, we have two variants of the extreme points: $n-s-1 = n-k$ and $n-s-1 = n-k-1$. Let us prove that the second variant is not valid. Indeed, the probability h_n in this case is $1-(n-k)\vartheta + \vartheta > \vartheta$. We get the contradiction because $h_n \leq \vartheta$. Hence, we get $n-s-1 = n-k$. The proof of all extreme points is obvious now.

For example, if $n = 3$ and $\vartheta = 0.4$, then $k = 1$, and we have three extreme points of the form $(\vartheta, \vartheta, 1-2\vartheta)$, i.e.,

$$(0.4, 0.4, 0.2), \ (0.4, 0.2, 0.4), \ (0.2, 0.4, 0.4).$$

Note that points of the form $(\vartheta, 1-\vartheta, 0)$ do not belong to the set $\mathcal{M}_U(\vartheta)$ in this case because $1-\vartheta \geq \vartheta$ and the condition $h_i \leq \vartheta$ is violated.

So, we have to solve $k \cdot \binom{n}{k}$ quadratic programming problems (13)-(14) for computing $f(\mathbf{x}, \mathbf{w})$ with respect to the minimax and minimin strategies. Let M_r be a subset of $n-k$ indices $1, ..., n$. Problem (13)-(14) is solved for every $r = 1, ..., n$ and every subset $M_r \subset \{1, 2, ..., n\} \backslash \{r\}$. Constraints (14) by fixed r and M_r are of the form:

$$\begin{aligned} 0 \leq \varphi_r &\leq 1-(n-k)\vartheta, \\ 0 \leq \varphi_i &\leq \vartheta, \ i \in M_r, \\ \varphi_j &= 0, \ j \notin M_r, \ j \neq r. \end{aligned}$$

Let us return to the ε-contaminated robust model and consider the question how the set $\mathcal{M}(\varepsilon)$ is changed in the case of bounded set $\mathcal{M}_U(\vartheta)$. Every extreme point in the modified (restricted) set $\mathcal{M}(\varepsilon, \vartheta)$ has exactly $n-k$ elements $(1-\varepsilon)n^{-1} + \varepsilon\vartheta$, $k-1$ elements $(1-\varepsilon)n^{-1}$ and one element $(1-\varepsilon)n^{-1} + \varepsilon(1-(n-k)\vartheta)$.

6.2 Lower bounded vacuous models

Another generalization of the vacuous model is obtained by introducing lower bounds for the probabilities $h_1, ..., h_n$, namely, we state that $h_i \geq \theta$, $i = 1, ..., n$, where θ is some number between 0 and $1/n$. We denote the reduced set of probability distributions as $\mathcal{M}_L(\theta)$. The upper possible bound for θ is $1/n$ because the set $\mathcal{M}_L(\theta)$ becomes empty when $\theta > 1/n$.

In sum, we can state that the set $\mathcal{M}_L(\theta)$ is produced by the following system of $2n$ inequalities and one equality:

$$\theta \leq h_i \leq 1, \ i = 1, ..., n, \ h_1 + ... + h_n = 1. \tag{19}$$

Proposition 6 *The set $\mathcal{M}_L(\theta)$ has $t = n$ extreme points such that every extreme point consists of exactly $n-1$ elements θ and one element $1-(n-1)\theta$.*

Proof It follows from the system of $2n$ inequalities (19) that every extreme point satisfies to $n-1$ equalities $h_i = 1$ (their number is s) and $h_i = \theta$ (their number is $n-s-1$). Suppose $s = 1$ or larger. Then $h_j = 0$ if $j \neq i$. But $h_j \geq \theta$. This implies that $s = 0$, and we have $n-1$ equalities $h_i = \theta$. The proof of all extreme points is obvious now.

In sum, we get n optimization problems (13)-(14) such that constraints (14) are of the form:

$$0 \leq \varphi_i \leq \theta, \ i \neq s,$$
$$0 \leq \varphi_s \leq 1-(n-1)\theta.$$

6.3 Kolmogorov-Smirnov bounds

One of the ways for taking into account the amount of statistical data and for constructing bounds for the set of probability distributions is using the Kolmogorov-Smirnov confidence limits for the empirical cumulative distribution function $F_n(\mathbf{x})$ (see [43, Section 1.1.] for details). It is a quite different model which does not have anything common with the neighbourhood models at first glance. However, we will see an interesting link between the models.

If we assume that $F(\mathbf{x})$ is some unknown true probability distribution of points from the training set, then we can choose a critical value of the test statistic $d_{n,1-\gamma}$ such that a band of width $\pm d_{n,1-\gamma}$ around $F_n(\mathbf{x})$ will entirely contain $F(\mathbf{x})$ with probability $1-\gamma$, which is to be interpreted as a confidence statement in the frequentist statistical framework. The ways for computing $d_{n,1-\gamma}$ by given n and γ can be found in the book [19, Subsection 8.9.3.].

By taking into account that the bounds are cumulative distribution functions, we write the following bounds $\underline{F}_n(\mathbf{x})$ and $\overline{F}_n(\mathbf{x})$ for some unknown distribution function $F(\mathbf{x})$:

$$\underline{F}_n(\mathbf{x}) \leq F(\mathbf{x}) \leq \overline{F}_n(\mathbf{x}), \tag{20}$$

where

$$\underline{F}_n(\mathbf{x}) = \max(F_n(\mathbf{x}) - d_{n,1-\gamma}), 0),$$
$$\overline{F}_n(\mathbf{x}) = \min(F_n(\mathbf{x}) + d_{n,1-\gamma}, 1).$$

It can be seen from the above inequality that the left tail of the upper probability distribution is $d_{n,1-\gamma}$. The right tail of the lower probability distribution is $1-d_{n,1-\gamma}$.

It has been shown by Utkin and Coolen [39] that the largest value of the expected risk is achieved at the probability distribution which coincides with the lower Kolmogorov-Smirnov confidence limit. Moreover, it has been shown that this distribution has $k-1$ jumps of size 0, one jump of size $k/n-d_{n,1-\gamma}$ and the other jumps all of size $1/n$, where k is determined from the condition

$$d_{n,1-\gamma}n < k \leq d_{n,1-\gamma}n + 1.$$

It is assumed here that there are no coinciding points, i.e., $x_j \neq x_i$ for every $j \neq i$. This assumption can be relaxed for the case of coincident points. In this case, the number of jumps is reduced.

It is important to point out here that Kolmogorov-Smirnov bounds use an assumption that there is a jumps $1-d_{n,1-\gamma}$ which is located at boundary points of $\mathcal{X}$ far from all data points (see [39] for details). Therefore, in order to normalize the sizes of jumps at points $\mathbf{x}_1, ..., \mathbf{x}_n$, every size has to be divided into $1-d_{n,1-\gamma}$. As a result, we get one jump of size

$$\frac{k-nd_{n,1-\gamma}}{n(1-d_{n,1-\gamma})},$$

$n-k$ jumps of size

$$\frac{1}{n(1-d_{n,1-\gamma})},$$

and $k-1$ jumps of size 0.

Let us denote

$$\vartheta = \frac{1}{n(1-d_{n,1-\gamma})}.$$

Then we get one jump equal to $1-(n-k)\vartheta$, $n-k$ jumps equal to ϑ, $k-1$ jumps of size 0. Moreover, the value ϑ fulfils the following condition:

$$\frac{1}{n-k+1} < \vartheta \leq \frac{1}{n-k}.$$

It follows from the above that we have obtained extreme points of the upper bounded vacuous model. In other words, the set of probability distributions produced by Kolmogorov-Smirnov bounds and by the upper bounded vacuous model coincide. This is a very important feature which provides a way for interpretation of the bound ϑ in terms of critical values of the test statistic $d_{n,1-\gamma}$. The above can be regarded as a proof of the following proposition.

Proposition 7 *The upper bounded vacuous model and Kolmogorov-Smirnov bounds produce the same classification models.*

Quadratic programming problems (13)-(14) are constructed and solved in the same way as in the case of the bounded vacuous model.

The upper bounded vacuous model and the model using Kolmogorov-Smirnov bounds have another interesting property. Every extreme point has $k-1$ zero-valued elements. This implies that weights of $k-1$ points from the training set are 0, i.e., these points are not used in computing the optimal parameters. Of course, every point from the training set is used, but it is used by taking other extreme points. The number of the zero-valued points strongly depends on the total number of points n in the training set and on the parameter ϑ or the confidence probability $1-\gamma$.

7 Experiments

The models proposed in this paper are illustrated via several examples, all computations have been performed using the statistical software R. We investigate the performance of the proposed method and compare it with the standard SVM approach by considering the accuracy (ACC), which is the proportion of correctly classified cases on a sample of data and is often used to quantify the predictive performance of classification methods. ACC is an estimate of a classifier's probability of a correct response, and it is an important statistical measures of the performance an OCC test. ACC can formally be written as

$$ACC = \frac{1}{N}\sum_{i=1}^{N}\left(I(y_i \cdot f(\mathbf{x}_i, \mathbf{w}) \geq 0)\right),$$

where y_i is the label of the i-th test example $\mathbf{x}_i$; $I(\cdot)$ is the indicator function.

The labels y_i are unknown for the classifier. However, in order to evaluate it, testing examples are divided into two classes whose labels are -1 for abnormal examples and 1 for other examples.

We will denote the accuracy measure for models using the linear-vacuous mixture as ACC_{lv}, the constant odds-ratio model as ACC_{or} and for the standard SVM as ACC_{st}.

All experiments use a standard Gaussian radial basis function (GRBF) kernel with the kernel parameter σ. Different values for the parameter σ have been tested, choosing those leading to the best results.

We consider the performance of our method with synthetic data having two features x_1 and x_2. The training set consisting of two subsets is generated in accordance with the normal probability distributions such that $N_1 = (1-\varepsilon_0)N$ examples (the first subset) are generated with mean values $\mathbf{m}_1 = (4,4)$ and $N_2 = \varepsilon_0 N$ examples (the second subset) have mean values $\mathbf{m}_2 = (12,12)$. The standard deviation is $s = 2$ for both subsets and both features. Here ε_0 is a portion of abnormal examples in training set.

The parameters ν and ε_0 are 0.1 and 0.1, respectively. We take $N = 30$, $\varepsilon = 0.3$, $\sigma = 100$. The corresponding accuracy measures for the minimax strategy are $ACC_{\text{lv}} = 0.608$, $ACC_{\text{or}} = 0.596$ and $ACC_{\text{st}} = 0.596$. One can see from the results that the linear-vacuous mixture provides better accuracy in comparison with other models. The accuracy measures for the minimin strategy are $ACC_{\text{lv}} = 0.574$, $ACC_{\text{or}} = 0.57$ and $ACC_{\text{st}} = 0.592$. Here the standard approach gives better results.

Now we increase the parameter $\varepsilon = 0.4$. The corresponding accuracy measures for the minimax strategy are $ACC_{\text{lv}} = 0.564$, $ACC_{\text{or}} = 0.582$ and $ACC_{\text{st}} = 0.576$. The accuracy measures for the minimin strategy are $ACC_{\text{lv}} = 0.602$, $ACC_{\text{or}} = 0.578$ and $ACC_{\text{st}} = 0.598$. The obtained results show that the constant odds-ratio model is better by the minimax strategy. However, the linear-vacuous mixture provides better accuracy in case of the minimin strategy.

The contours $f(\mathbf{x}, \mathbf{w}) = 0$ and generated data points with the above parameters are shown Fig. 1 where pictures (a) and (b) correspond to the minimax and the minimin strategies by $\varepsilon = 0.3$, pictures (c) and (d) correspond to the minimax and the minimin strategies by $\varepsilon = 0.4$, respectively. Three models are compared: the linear-vacuous mixture (thick curve), the constant odds-ratio (thin curve) and the standard SVM by $\varepsilon = 0$ (dashed curve).

The next experiment is to study how the robust models work when there are no outliers. In other words, we take $\varepsilon_0 = 0$. The contours $f(\mathbf{x}, \mathbf{w}) = 0$ and generated data points by $\varepsilon_0 = 0$ are shown Fig. 2 where pictures (a) and (b) correspond to the minimax and the minimin strategies by $\varepsilon = 0.3$, pictures (c) and (d) correspond to the minimax and the minimin strategies by $\varepsilon = 0.4$, respectively. We again compare three models: the linear-vacuous mixture (thick curve), the constant odds-ratio (thin curve) and the standard SVM by $\varepsilon = 0$ (dashed curve).

If $\varepsilon = 0.3$ (other parameters are without changes), then we get for the minimax strategy: $ACC_{\text{lv}} = 0.75$, $ACC_{\text{or}} = 0.75$ and $ACC_{\text{st}} = 0.753$. The corresponding accuracy measures for the minimin strategy are $ACC_{\text{lv}} = 0.734$, $ACC_{\text{or}} = 0.725$ and $ACC_{\text{st}} = 0.723$.

If $\varepsilon = 0.4$ (other parameters are without changes), then we get for the minimax strategy: $ACC_{\text{lv}} = 0.749$, $ACC_{\text{or}} = 0.749$ and $ACC_{\text{st}} = 0.739$. The corresponding accuracy measures for the minimin strategy are $ACC_{\text{lv}} = 0.689$, $ACC_{\text{or}} = 0.756$ and $ACC_{\text{st}} = 0.736$.

As a further example analyzed models are applied to the well-known "Iris" data set from the UCI Machine Learning Repository [14]. The data set contains 3 classes (Iris Setosa, Iris Versicolour, Iris Virginica) of 50 instances each. The number of features is 4 (sepal length in cm, sepal width in cm, petal length in cm, petal width in cm). It is supposed that data from the Iris Setosa class are abnormal. The number of training data n is 30. For the experiment, n points are randomly selected such that $(1-\varepsilon_0)n$ points are taken from the set of positively labelled examples and $\varepsilon_0 n$ points are from negatively labelled examples. Here $\varepsilon_0 = 50/150 \simeq 0.333$. The parameters for modelling are $v = 0.2$, $\sigma = 12.9$. The accuracy measures for the minimax strategy are $ACC_{\text{lv}} = 0.933$, $ACC_{\text{or}} = 0.847$ and $ACC_{\text{st}} = 0.74$. If we take $n = 50$ and $\sigma = 31.6$, then the accuracy

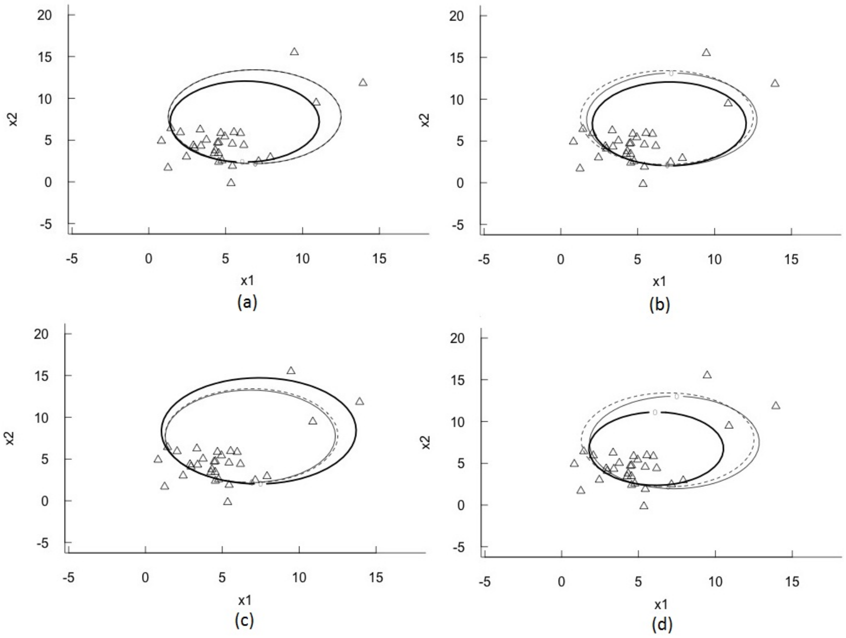

Fig. 1 The contours $f(\mathbf{x},\mathbf{w}) = 0$ by $\varepsilon_0 = 0.1$

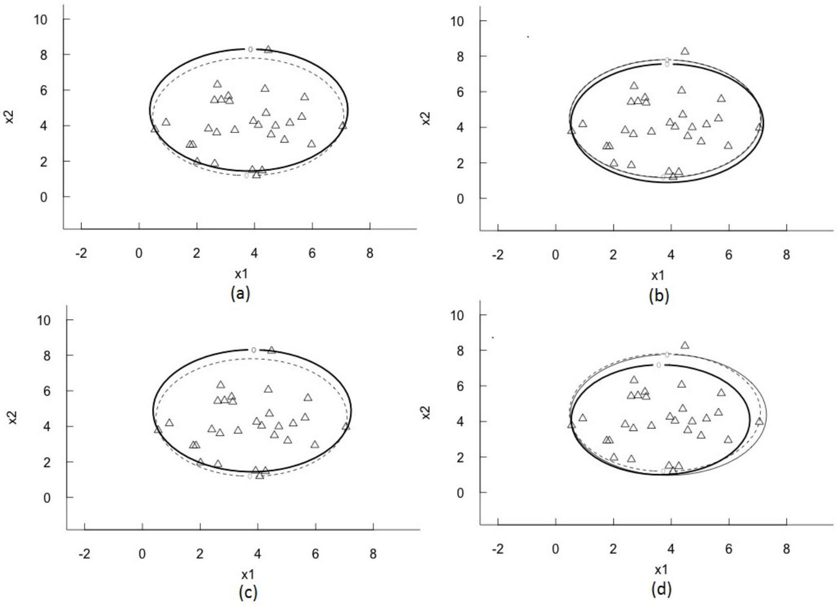

Fig. 2 The contours $f(\mathbf{x},\mathbf{w}) = 0$ by $\varepsilon_0 = 0$

measures for the minimax strategy are $ACC_{\text{lv}} = 0.913$, $ACC_{\text{or}} = 0.88$ and $ACC_{\text{st}} = 0.88$. One can see that the quality of different models equalizes with increase of the size of the training set, i.e., the impact of the contaminated parameter ε is reduced.

8 Conclusion

Robust OCC models have been proposed in the paper, which are based on using the imprecise statistical models instead of the empirical probability distribution accepted in the standard SVM. Classification parameters of every model are obtained by solving a finite set of simple quadratic programming problems whose solution does not meet any difficulties. The simplicity of the OCC models follows from the fact that the optimal solution is achieved at one of the extreme points of a set of probability distributions produced by the underlying imprecise statistical model. The minimax and minimin strategies are used for choosing a single "optimal" probability distribution from the set of distributions. These strategies have a clear explanation and justification in the framework of decision theory. However, they are "extreme" in the sense that they provide too pessimistic and optimistic decisions. Therefore, decision strategies different from the above can also be applied to the OCC and investigated. It should be noted that the minimin strategy is not robust. However, it is introduced as the second "extreme" case which provides some optimistic decision. Moreover, both strategies can be a basis for the so-called "cautious" strategy which is the linear combination of minimax and minimin strategies with a "caution" parameter. The "cautious" strategy is similar to Hurwicz criterion in decision theory. It is also interesting to study the mean value of expected losses over all extreme points or the linear mixture of the "extreme" strategies. This is an important direction for future research.

The algorithm for computing the optimal parameters of every OCC model can be easily implemented with standard functions of the statistical software package R. Experimental results with synthetic and some real data reported have shown that the proposed robust OCC models are comparable with the standard approach proposed by Scholkopf et al. [30,32] for some initial parameters. At the same time, the paper does not contain a detailed experimental study of the proposed models because its main aim is to provide a framework for constructing the robust OCC models. Some experimental results are given in the paper for illustrating opportunities of the proposed models and their possible outperforming the standard approach.

One of the advantages of the proposed OCC models is that they reflect the possible violation of too strong assumptions about uniformity of the probability mass function of data points accepted in the standard approach during testing. This violation is taken into account by considering the set of probability distributions.

Another important direction for future work is to apply the imprecise statistical models to the OCC model proposed by Tax and Duin [35,34], to the model proposed by Campbell and Bennett [8], which uses linear programming techniques. It is interesting to apply the proposed imprecise models to the logistic regression models [27].

Finally, it is also worth noticing that the proposed models can easily be extended on the case of binary or multi-class classification.

Acknowledgement

I would like to express my appreciation to the anonymous referees whose very valuable comments have improved the paper.

References

1. Bartkowiak, A.: Anomaly, novelty, one-class classification: A comprehensive introduction. International Journal of Computer Information Systems and Industrial Management Applications **3**, 61–71 (2011)
2. Ben-Tal, A., Ghaoui, L., Nemirovski, A.: Robust optimization. Princeton University Press, Princeton, New Jersey (2009)
3. Berger, J.: Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York (1985)
4. Bi, J., Zhang, T.: Support vector classification with input data uncertainty. In: Saul, L., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems, vol. 17, pp. 161–168. MIT Press, Cambridge, MA (2004)
5. Bicego, M., Figueiredo, M.: Soft clustering using weighted one-class support vector machines. Pattern Recognition **42**, 27–32 (2009)
6. Bouveyron, C., Girard, S.: Robust supervised classification with mixture models: Learning from data with uncertain labels. Pattern Recognition **42**(11), 2649 – 2658 (2009)
7. Campbell, C.: Kernel methods: a survey of current techniques. Neurocomputing **48**(1-4), 63–84 (2002)
8. Campbell, C., Bennett, K.: A linear programming approach to novelty detection. In: Leen, T., Dietterich, T., Tresp, V. (eds.) Advances in Neural Information Processing Systems, vol. 13, pp. 395–401. MIT Press (2001)
9. Cerioli, A., Riani, M., Atkinson, A.: Robust classification with categorical variables. In: Rizzi, A., Vichi, M. (eds.) Compstat 2006 - Proceedings in Computational Statistics, pp. 507–519. Physica-Verlag HD (2006)
10. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. Tech. Rep. TR 07-017, University of Minnesota, Minneapolis, MN, USA (2007)
11. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys **41**, 1–58 (2009)
12. Cherkassky, V., Mulier, F.: Learning from Data: Concepts, Theory, and Methods. Wiley-IEEE Press, UK (2007)
13. Evgeniou, T., Poggio, T., Pontil, M., Verri, A.: Regularization and statistical learning theory for data analysis. Computational Statistics & Data Analysis **38**(4), 421 – 432 (2002)
14. Frank, A., Asuncion, A.: UCI machine learning repository (2010). URL http://archive.ics.uci.edu/ml
15. Ghaoui, L., Lanckriet, G., Natsoulis, G.: Robust classification with interval data. Tech. Rep. Report No. UCB/CSD-03-1279, University of California, Berkeley, California 94720 (2003)
16. Gilboa, I., Schmeidler, D.: Maxmin expected utility with non-unique prior. Journal of Mathematical Economics **18**(2), 141–153 (1989)
17. Hodge, V., Austin, J.: A survey of outlier detection methodologies. Artificial Intelligence Review **22**(2), 85–126 (2004)

18. Huber, P.: Robust Statistics. Wiley, New York (1981)
19. Johnson, N., Leone, F.: Statistics and experimental design in engineering and the physical sciences, vol. 1. Wiley, New York (1964)
20. Khan, S., Madden, M.: A survey of recent trends in one class classification. In: Coyle, L., Freyne, J. (eds.) Artificial Intelligence and Cognitive Science, *Lecture Notes in Computer Science*, vol. 6206, pp. 188–197. Springer Berlin / Heidelberg (2010)
21. Kwok, J., Tsang, I.H., Zurada, J.: A class of single-class minimax probability machines for novelty detection. IEEE Transactions on Neural Networks **18**(3), 778–785 (2007)
22. Lanckriet, G., Ghaoui, L., Bhattacharyya, C., Jordan, M.: A robust minimax approach to classification. Journal of Machine Learning Research **3**, 555–582 (2002)
23. Lanckriet, G., Ghaoui, L., Jordan, M.: Robust novelty detection with single-class mpm. In: Becker, S., Thrun, S., Obermayer, K. (eds.) Advances in Neural Information Processing Systems, vol. 15, pp. 905–912. MIT Press, Cambridge, MA (2003)
24. Liu, Z., Wu, Q., Zhang, Y., Chen, C.L.P.: Adaptive least squares support vector machines filter for hand tremor canceling in microsurgery. International Journal of Machine Learning and Cybernetics **2**(1), 37–47 (2011)
25. Markou, M., Singh, S.: Novelty detection: a reviewpart 1: statistical approaches. Signal Processing **83**(12), 2481–2497 (2003)
26. Maulik, U., Chakraborty, D.: A novel semisupervised SVM for pixel classification of remote sensing imagery. International Journal of Machine Learning and Cybernetics **3**(3), 247–258 (2012)
27. Musa, A.B.: Comparative study on classification performance between support vector machine and logistic regression. International Journal of Machine Learning and Cybernetics (2012), DOI: 10.1007/s13042-012-0068-x
28. Provost, F., Fawcett, T.: Robust classification for imprecise environments. Machine Learning **42**(3), 203–231 (2001)
29. Robert, C.: The Bayesian Choice. Springer, New York (1994)
30. Scholkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., Williamson, R.: Estimating the support of a high-dimensional distribution. Neural Computation **13**(7), 1443–1471 (2001)
31. Scholkopf, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, Cambridge, Massachusetts (2002)
32. Scholkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., Platt, J.: Support vector method for novelty detection. In: Advances in Neural Information Processing Systems, pp. 526–532 (2000)
33. Steinwart, I., Hush, D., Scovel, C.: A classification framework for anomaly detection. Journal of Machine Learning Research **6**, 211–232 (2005)
34. Tax, D., Duin, R.: Support vector domain description. Pattern Recognition Letters **20**, 1191–1199 (1999)
35. Tax, D., Duin, R.: Support vector data description. Machine Learning **54**, 45–66 (2004)
36. Tikhonov, A., Arsenin, V.: Solution of Ill-Posed Problems. W.H. Winston, Washington DC (1977)
37. Trafalis, T., Gilbert, R.: Robust support vector machines for classification and computational issues. Optimization Methods and Software **22**(1), 187–198 (2007)
38. Troffaes, M.: Decision making under uncertainty using imprecise probabilities. International Journal of Approximate Reasoning **45**(1), 17–29 (2007)
39. Utkin, L., Coolen, F.: On reliability growth models using Kolmogorov-Smirnov bounds. International Journal of Performability Engineering **7**(1), 5–19 (2011)
40. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
41. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London (1991)
42. Wang, J., Lu, H., Plataniotis, K., Lu, J.: Gaussian kernel optimization for pattern classification. Pattern Recognition **42**(7), 1237 – 1247 (2009)
43. Wasserman, L.: All of Nonparametric Statistics. Springer, New York (2006)
44. Xiao, J.-Z., Wang, H.-R., Yang, X.-C., Gao, Z.: Multiple faults diagnosis in motion system based on SVM. International Journal of Machine Learning and Cybernetics **3**(1), 77–82 (2012)
45. Xu, H., Caramanis, C., Mannor, S.: Robustness and regularization of support vector machines. The Journal of Machine Learning Research **10**, 1485–1510 (2009)
46. Xu, L., Crammer, K., Schuurmans, D.: Robust support vector machine training via convex outlier ablation. In: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), vol. 21, pp. 536–542. AAAI Press; MIT Press, Boston, Massachusetts (2006)
47. Yang, X., Song, Q., Wang, Y.: A weighted support vector machine for data classification. International Journal of Pattern Recognition and Artificial Intelligence **21**(5), 961–976 (2007)