**Noname manuscript No.**
(will be inserted by the editor)

# Imprecise prior knowledge incorporating into one-class classification

**Lev V. Utkin · Yulia A. Zhuk**

Received: date / Accepted: date

**Abstract** An extension of Campbell and Bennett's novelty detection or one-class classification model incorporating prior knowledge is studied in the paper. The proposed extension relaxes the strong assumption of the empirical probability distribution over elements of a training set and deals with a set of probability distributions produced by prior knowledge about training data. The classification problem is solved by considering extreme points of the probability distribution set or by means of the conjugate duality technique. Special cases of prior knowledge are considered in detail, including the imprecise linear-vacuous mixture model and interval-valued moments of feature values. Numerical experiments show that the proposed models outperform Campbell and Bennett's model for many real and synthetic data.

**Keywords** Machine learning · One-class classification · Minimax strategy · Novelty detection · Linear programming · Imprecise statistical model · Extreme points.

## 1 Introduction

Many authors (see, for example, [25,29,31]) point out that prior knowledge about the problem domain at hand may significantly improve the performance of classifiers in many applications. Therefore, incorporating prior knowledge into classification models is an important problem. Prior knowledge can take various forms, ranging from knowledge about the importance of a class, the informativeness of features, the quality of samples, to the dependency of variables.

Lev V. Utkin and Yulia A. Zhuk
Department of Control, Automation and System Analysis
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

One of the pioneering works devoted to incorporating prior knowledge into classification models, in particular, into support vector machine (SVM) is by Scholkopf et al. [36]. The authors investigate two forms of prior knowledge: invariances under group transformations and prior knowledge about locality in images. The first form can be viewed as an important property of decision functions. A series of papers devoted to the invariance in kernel methods have been published [11,14,19] due to importance of this prior information. The second form of prior knowledge considers different amounts of information in image regions.

An interesting review of various methods for incorporating prior knowledge into SVMs was proposed by Lauer and Bloch [25]. The authors of the review point out that prior knowledge may be a key element allowing to increase the performance of classification in many applications. Lauer and Bloch [25] consider different types of prior knowledge encountered in pattern recognition by dividing them into the same two main groups: class-invariance and knowledge on the data. Another paper of Lauer and Bloch [26] explores the addition of constraints to the linear programming formulation of the support vector regression problem for the incorporation of prior knowledge.

Another simple and straightforward way to incorporate prior knowledge is to assign weights to examples of a training set. The weights can be assigned to classes in order to deal with unbalanced data (see, for example, [22,62]). The weights can also be assigned to every example of a training set in accordance with predefined rules. A number of papers have been devoted to incorporating this important information [5,17,43,58,60,61]. Authors of the work [56] propose a method to incorporate prior knowledge of a special form into SVM. This prior information is of the form: "for a randomly sampled unlabelled image, before seeing any additional evidence, its label is negative with high probability".

An important work was proposed by Mangasarian [31] where it is shown that prior expert knowledge can be incorporated into the classification problem by adding constraints in a linear program corresponding to prior knowledge. The obtained linear program is called by Mangasarian as a knowledge-based linear program. Authors of the work [16] introduce prior knowledge in the form of multiple polyhedral sets incorporated into a linear support vector machine classifier. The authors provide a typical example of Wisconsin breast cancer prognosis [27] illustrating how prior knowledge can be incorporated into linear inequalities. The prior knowledge utilized in this example consists of the prognosis rules provided by doctors [27] which depend on two features from the dataset: tumor size and lymph node status. It is clearly shown in [16] that the rules provided by doctors can be converted to linear inequalities and are a very important and useful prior information.

Prior knowledge in the form of simple advice rules can also greatly speed up convergence in learning algorithms. A method for incorporating this information into classification is provided by Kunapuli et al. [23]. The goal of the work [23] is to update the hypothesis taking into account the prior knowl-

edge in the form of soft polyhedral advice so as to make increasingly accurate predictions on subsequent examples.

Prior knowledge concerning a specific application may also be very useful in classification problems. For instance, prior information about different functional groups of genes which have been identified by previous studies to have biological functions can be successfully incorporated into classifiers such that genes from different groups have different prior distributions. Tai and Pan [44] show that this prior information can potentially improve both the predictive performance and interpretability of the resulting model. In order to improve learning performance, the authors of the work [30] consider two kinds of prior knowledge included in patent documents: patent's publishing date and a hierarchical structure of a patent classification system. The authors of the work [33] construct an explicitly heterogeneous kernel function by computing separate kernels for each data type for the gene functional classification. The resulting kernel incorporates prior knowledge about the heterogeneity of the data by accounting for higher-order correlations among features of one data type, but ignoring higher-order correlations across data types. This heterogeneous kernel leads to improved performance with respect to an SVM trained directly on the concatenated data. Dayanik et al. [13] incorporate domain knowledge into supervised learning of the text classification in a case of small and unsystematically collected training sets. The authors construct sets of words which are good predictors for a class and combine them with training examples in a Bayesian framework. Domain knowledge is used to specify a prior distribution for parameters of a logistic regression model. Xing et al. [59] extract interpretable features on time series as prior knowledge for early classification. Authors of the work [42] consider two sorts of prior knowledge in the handwritten Chinese character recognition. The first, solution knowledge, concerns the target of learning itself and is specific to the learning task at hand, for instance, the kernel function of an SVM. The second, domain knowledge, describes objects of the world, for example, in handwritten character recognition one may believe that the pixels in the input images arise from strokes of a writing implement.

Veillard et al. [53] consider prior knowledge which can be obtained as extra advice from experts of a particular problem. They propose a method for incorporating prior knowledge via an adaptation of the standard RBF kernel. Small et al. [39] study prior domain knowledge in the form of expert-provided ranked labeled features. The authors propose the so-called Constrained Weight Space SVM as an extension of the SVM by adding additional constraints to reflect this domain knowledge. Small et al. [39] show how experts beliefs of a special form can be directly encoded in the form of weight constraints.

Li et al. [28] present a robust conjugate duality theory for convex programming problems in the face of data uncertainty within the framework of robust optimization, extending the powerful conjugate duality technique. As an application, the authors of [28] derive a robust conjugate duality theorem for support vector machines. The idea of the authors of [28] to represent the

minimax optimization problem in the form of a single optimization problem by means of the duality technique will be used in the present paper.

Authors of the work [29] propose a kernel density classifier which integrates prior knowledge about measurement noise into system construction.

The intensive interest to the prior knowledge incorporation into classifier stimulates us to develop the corresponding models applied to one-class classification (OCC) or to novelty detection. The OCC or novelty detection is a special case of machine learning, which aims to detect novel or abnormal (anomalous) instances [6,7,9,10,12,35,38,41,45,46]. Comprehensive reviews of OCC models are provided by several authors, for example, by Markou and Singh [32], by Bartkowiak [2]. A typical feature of the models is that only unlabeled samples are available. One of the most common ways to define anomalies is by saying that anomalies are not concentrated [37].

One of the ways to solve the OCC problem is to estimate a binary-valued function $f$ that is positive in a region where a density of examples is high, and negative elsewhere. Sample points outside this region can be regarded as anomalous observations.

We mark out three main approaches for constructing the OCC models in the framework of SVM. The first approach is proposed by Tax and Duin [46, 45]. This is one of the well-known novelty detection models, which can be regarded as an unsupervised learning problem. According to this approach, the training of the one-class SVM consists in determining the smallest hypersphere containing training data. 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 at el. [35,38]. Under this approach, a hyperplane is used to separate the training data from the origin with the maximal margin, i.e., the objective is to separate off the region containing the data points from the surface region containing no data. The third approach which will be extended in the paper is the linear programming approach to OCC proposed by Campbell and Bennett [7]. The model proposed by Campbell and Bennett uses linear programming techniques.

It should be noted that there are other interesting novelty detection models (see for instance, [5,20,24]). However, we study the third approach [7] and modify or extend it in order to take into account prior knowledge about elements of a training set. Our main idea is to construct a set of robust imprecise one-classification models using the framework of imprecise probabilities [54]. We assume that the empirical probability distribution accepted in many classification models, in particular, in the model proposed by Campbell and Bennett [7], should be replaced by a set of probability distributions. Moreover, we restrict the set of probability distributions by distributions concentrated at data points, i.e., every probability distribution is defined over data points from the training set. In order to solve the classification problem, we select the "worst" probability distribution from the set, which provides the largest value of the expected risk. This choice corresponds to the minimax strategy in decision making and can be interpreted as an insurance against the worst case [34]. The next problem is how to construct the set of probability distributions from

prior knowledge. This problem is solved in the paper in a general form under condition that prior information is represented in the form of interval-valued expectations of some functions. This representation covers a lot of common estimates and judgments. For example, the information about intervals of probabilities can be represented by means of lower and upper expectations of the indicator function. Comparative judgments can be written as the expectation of difference of two functions which are defined by the judgments.

In order to illustrate how the classification problem can be solved in special cases, we consider in detail, first, the $\varepsilon$-contaminated (robust) model [21] or the imprecise linear-vacuous mixture model [54] and, second, known interval-valued mean values and standard deviations of features or just their first two moments. The second special case is very interesting because it demonstrates a number of important and unexpected results. First of all, many experimental studies show that prior knowledge of the first two moments allows us to construct OCC models which outperform Campbell and Bennett model for most real data sets and for synthetic data. Second, we observe the following interesting property of the model. Suppose we do not have prior information, and we compute the precise mean values and standard deviations of features just by using the available training set. If we take some interval of mean values instead of the precise sample mean value such that the precise mean value lies in the interval, then the proposed robust imprecise model provides better results in comparison with Campbell and Bennett model. It is an unexpected result because we do not explicitly incorporate any prior information from without. The information results from the training set. Nevertheless, the proposed models outperforms Campbell and Bennett model in this case for many data sets. It is also important here that only imprecise information (for instance, intervals of mean values) leads to the outperforming classification characteristics.

Two approaches are used to solve the minimax optimization problems. The first approach is based on considering extreme points of the set of probability distributions. This approach is especially useful when the extreme points are a priori known for a statistical model producing the probability set. The second approach is based on the powerful conjugate duality technique. This approach is similar to those provided by Li et al. [28]. It is especially useful when the number of extreme points is very large or when their search is a hard computation problem. Moreover, an important advantage of the dual form of the corresponding optimization problem is that we do not need to assume that features are statistically independent because we consider only joint probability distributions even if prior knowledge deals with separate features. The assumption of statistical independence of features is criticized in the naïve Bayes classifier [57].

The paper is organized as follows. Section 2 presents the standard OCC problem proposed by Campbell and Bennett [7]. Section 3 considers how Campbell and Bennett's OCC model can be extended when we have prior knowledge. The section studies approaches to solve classification problems by means of extreme points of the set of probability distributions and by means of the duality technique. The robust contamination neighborhood model is

incorporated into the OCC problem in Section 4. Interval-valued mean values and standard deviations as prior knowledge and their incorporation into the classification model are studied in Section 5. Numerical experiments with synthetic and some real data illustrating accuracy of the proposed model are provided in Section 6. In Section 7, concluding remarks are made.

## 2 Campbell and Bennett's novelty detection linear model

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 $\mathbb{R}^l$. 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 [35,38], 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 the maximum margin.

Let $\phi$ 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 $\phi$ can be computed by evaluating some simple kernel $K(\mathbf{x}, \mathbf{y}) = (\phi(\mathbf{x}), \phi(\mathbf{y}))$, such as the Gaussian kernel

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

$\gamma$ is the kernel parameter determining the geometrical structure of the mapped samples in the kernel space. It is pointed out by [55] that the problem of selecting a proper parameter $\gamma$ is very important in classification. When a very small $\gamma$ is used ($\gamma \to 0$), $K(\mathbf{x}, \mathbf{y}) \to 0$ 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 $\gamma$ is chosen ($\gamma^2 \to \infty$), $K(\mathbf{x}, \mathbf{y}) \to 1$ 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 $\gamma$ will not result in more separable samples in $G$.

We consider the linear programming approach to novelty detection proposed by Campbell and Bennett [7]. The authors start from the hard margin case, when any training point $\mathbf{x}_j$ lying outside some predefined surface restricted the training points is viewed as abnormal. This surface is defined as the level set, $f(\mathbf{z}) = 0$, of some nonlinear function. In feature space, $f(\mathbf{z}) = \sum_i \varphi_i K(\mathbf{z}, \mathbf{x}_i) + b$, this corresponds to a hyperplane which is pulled onto the mapped data points with the restriction that the margin always remains positive or zero [7]. Here $\varphi = (\varphi_n, ..., \varphi_n)$ are parameters of the function $f$ in the feature space or Lagrange multipliers.

A criteria for constructing the optimal function $f(\mathbf{z})$ proposed by Campbell and Bennett is to minimize the mean value of the output of the function, i.e.,

$\sum_i f(\mathbf{x}_i)$. This is achieved by minimizing:

$$W(\varphi, b) = \sum_{i=1}^{n} \left( \sum_{j=1}^{n} \varphi_j K(\mathbf{x}_i, \mathbf{x}_j) + b \right),$$

subject to

$$\sum_{j=1}^{n} \varphi_j K(\mathbf{x}_i, \mathbf{x}_j) + b \geq 0, \ i = 1, ..., n,$$

$$\sum_{i=1}^{n} \varphi_i = 1, \ \varphi_i \geq 0. \tag{1}$$

The bias $b$ is just treated as an additional parameter in the minimization process. The added constraints on $\varphi$ restrict the class of models to be considered. As indicated by Campbell and Bennett [7], these constraints amount to a choice of scale for the weight vector normal to the hyperplane in feature space and hence do not impose a restriction on the model. Also, these constraints ensure that the problem is well-posed and that an optimal solution with $\varphi \neq 0$ exists. Other constraints on the class of functions are possible, e.g. $\|\varphi\|_1 = 1$ with no restriction on the sign of $\varphi_i$.

It is important to point out here that Campbell and Bennett propose to use the mean value of the output of the function. It follows from the form of $W(\varphi, b)$ that the empirical probability distribution $(1/n, ..., 1/n)$ is assumed to get the mean value $W(\varphi, b)$. The multiplier $1/n$ is omitted because it does not change the optimization variables $\varphi$ and $b$.

To handle noise and outliers a soft margin is introduced in analogy to the usual approach used with support vector machines [12,37,40,52]. In this case, the following function has to be minimized:

$$W(\varphi, b) = \sum_{i=1}^{n} \left( \sum_{j=1}^{n} \varphi_j K(\mathbf{x}_i, \mathbf{x}_j) + b \right) + \frac{1}{vn} \sum_{i=1}^{n} \xi_i,$$

subject to (1) and

$$\sum_{j=1}^{n} \varphi_j K(\mathbf{x}_i, \mathbf{x}_j) + b \geq -\xi_i, \ \xi_i \geq 0, \ i = 1, ..., n, \tag{2}$$

The parameter $v \in [0; 1]$ controls the extent of margin errors (smaller $v$ means fewer outliers are ignored: $v \to 0$ corresponds to the hard margin limit). It is a parameter which is analogous to $\nu$ for the $\nu$-SVM standard method [12]. Slack variables $\xi = (\xi_1, ..., \xi_n)$ are used to allow points to violate margin constraints.

We will shortly call Campbell and Bennett's novelty detection model below as C-B model.

## 3 Extension of C-B model and sets of probability distributions

One can observe that the mean value $W(\varphi, b)$ for the hard margin case as well as for the soft margin is derived under assumption that every data point $\mathbf{x}_i$ or its function $f(\mathbf{x}_i)$ has the probability $1/n$. This follows from the empirical expected risk definition provided by Vapnik [52]. When the training set is large, this assumption can be accepted. However, it might be violated for a small training set or when the portion of abnormal observations is rather large. Moreover, if we would have additional statistical information about training data, for instance, the mean value of a feature, we can violate the strong assumption of the empirical probability distribution by means of this information. An obvious way to incorporate prior knowledge into the classification model is to relax this strong assumption and to allow probability distributions to be arbitrary from some set of probability distributions $\mathcal{P}_0$ defined by prior information on $\mathcal{X}$. However, this set contains infinitely many different probability distributions and its use may lead to extremely hard computation problems. Moreover, the set may contain the distributions which are unrealistic in a considered applied problem. Efficient algorithms for solving the optimization problems which are constructed with using the set $\mathcal{P}_0$ are given in detail in the book [3]. Nevertheless, we propose to reduce the set $\mathcal{P}_0$ and to make it consisting of probability density functions concentrated on elements of the training set $\mathbf{x}_1, ..., \mathbf{x}_n$. The density at points which do not belong to the training set is assumed to be 0. Every distribution in this new set can be viewed as a discrete distribution defined on $\mathcal{D}$, i.e., we can write $p(\mathbf{x}_i) = p_i$. We denote the set of all discrete distributions $p = (p_1, ..., p_n)$ as $\mathcal{P}$. Under some initial conditions for the set $\mathcal{P}$, the distribution $(1/n, ..., 1/n)$ may belong to $\mathcal{P}$.

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. [63], the use of robust optimization in classification is not new. 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 assume that every "true" data point is only known to belong to the interior of an Euclidian ball centered at the "nominal" data point $\mathbf{x}_i$. This model has a very clear intuitive geometric interpretation [3]. The robust classifier in this case is the one that maximizes the radius of balls. In contrast to the above models, our idea of robust models is based on relaxing strong assumptions about a probability distribution of data points. While some robust models [3] assume that each point can move around within an Euclidean ball, the proposed robust models assume that the probability distribution of points (but not a data point itself) can move around within a unit simplex under some restrictions. These restrictions depend on the applied imprecise statistical models and on the prior knowledge incorporating into the model, and they produce the set $\mathcal{P}$.

So, we do not know a "true" probability distribution $p = (p_1, ..., p_n)$ of data points, but we know that it belongs to a set of distributions $\mathcal{P}$, i.e., $p \in \mathcal{P}$.

Then we define a set of mean values of the function $f$ such that every element of the set is the following expected value:

$$\mathbb{E}_p f = \sum_{i=1}^{n} p_i f(\mathbf{x}_i).$$

The same can be written for the slack variables. Finally, we suppose that the objective function $W(\varphi, b)$ depends on the probability distribution $p$ (denoted $W_p(\varphi, b)$), and it can be written as follows:

$$W_p(\varphi, b) = \sum_{i=1}^{n} p_i \left( \sum_{j=1}^{n} \varphi_j K(\mathbf{x}_i, \mathbf{x}_j) + b \right) + \frac{1}{v} \sum_{i=1}^{n} p_i \xi_i$$
$$\sum_{i=1}^{n} p_i \left( \sum_{j=1}^{n} \varphi_j K(\mathbf{x}_i, \mathbf{x}_j) + b + \frac{1}{v} \xi_i \right).$$

Note that we do not know the probability distribution $p$. This implies that there is a set of $W_p(\varphi, b)$ defined by the set $\mathcal{P}$. The set of expectations can be studied in frameworks of robust models [21] or imprecise models [54].

We assume below that the set $\mathcal{P}$ is convex, i.e., it is produced by finitely many linear constraints. Then the set of $W_p(\varphi, b)$ has some lower $\underline{W}(\varphi, b)$ and upper $\overline{W}(\varphi, b)$ bounds due to convexity of the set $\mathcal{P}$. Here $\underline{W}(\varphi, b) = \min_{p \in \mathcal{P}} W_p(\varphi, b)$ and $\overline{W}(\varphi, b) = \max_{p \in \mathcal{P}} W_p(\varphi, b)$. The upper bound determines the well-known minimax (pessimistic) strategy. According to the minimax strategy, a probability distribution is selected from the set $\mathcal{P}$ such that the expected value $W_p(\varphi, b)$ achieves its maximum $\overline{W}(\varphi, b)$ for every fixed $\varphi, b$. It should be noted that the "optimal" probability distributions may be different for different values of parameters $\varphi, b$. The minimax strategy can be explained in a simple way. We do not know a precise probability distribution and every distribution from $\mathcal{P}$ 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 [34]. This criterion of decision making can be regarded as the well-known $\Gamma$-minimax [4, 18, 47]. The criterion is often given in terms of utilities. In this case, it is called $\Gamma$-maximin.

Finally, we write the optimization problem corresponding to the minimax strategy as follows:

$$\overline{W}(\varphi, b) = \min_{\varphi, b, \xi} \max_{p \in \mathcal{P}} \sum_{i=1}^{n} p_i V(\varphi, b, \mathbf{x}_i), \tag{3}$$

subject to (1), (2) and $p \in \mathcal{P}$.

Here

$$V(\varphi, b, \mathbf{x}_i) = \sum_{j=1}^{n} \varphi_j K(\mathbf{x}_i, \mathbf{x}_j) + b + \frac{1}{v} \xi_i.$$

The solution of the problem strictly depends on the set $\mathcal{P}$. We can consider two approaches to its solving.

### 3.1 Dual form for solving the problem

Let us consider the problem

$$W_{\mathcal{P}} = \max_{p \in \mathcal{P}} \sum_{i=1}^{n} p_i V(\varphi, b, \mathbf{x}_i)$$

in detail. This is the primal form of the linear programming problem with variables $p$. Due to its linearity, we can write its dual form. For doing that, we suppose the set $\mathcal{P}$ is produced by a set of $r$ constraints of the form:

$$\underline{a}_k \leq \mathbb{E}_p h \leq \overline{a}_k, \;\; k = 1, ..., r.$$

Here $h$ is some function of variables $\mathbf{x}_1, ..., \mathbf{x}_n$, $\underline{a}_k$ and $\overline{a}_k$ are lower and upper bounds for the expectation $\mathbb{E}_p h$ of a function $h$. For example, if we restrict the probability or the weight of the $k$-th point by interval $[\underline{a}_k, \overline{a}_k]$, then $h(\mathbf{x}) = \mathbf{1}_k(\mathbf{x}_i)$, $i = 1, ..., n$, is the indicator function taking the value 1 if $k = i$. If we know the mean value of the $k$-th feature, then $h(x^{(k)}) = x$, where $x^{(k)}$ is the variable corresponding to the $k$-th feature. If it is known that the probability of the $k$-th example is larger than the probability of the $j$-th example, then this prior information is represented as the lower bound for the expectation of function $h(\mathbf{x}) = \mathbf{1}_k(\mathbf{x}_i) - \mathbf{1}_j(\mathbf{x}_i)$, $i = 1, ..., n$, such that the lower bound for the expectation is non-negative.

It should be noted that the lower and upper bounds may coincide in some cases.

Denote vectors $\mathbf{H}_i = (h_1(\mathbf{x}_i), ..., h_r(\mathbf{x}_i))$, $\underline{\mathbf{A}} = (\underline{a}_1, ..., \underline{a}_r)$, $\overline{\mathbf{A}} = (\overline{a}_1, ..., \overline{a}_r)$. Let us introduce optimization variables $\mathbf{C} = (c_1, ..., c_r)^{\mathrm{T}}$, $\mathbf{D} = (d_1, ..., d_r)^{\mathrm{T}}$. The dual problem can be written as follows:

$$W_{\mathcal{P}} = \min_{c, \mathbf{C}, \mathbf{D}} \left\{ c + \overline{\mathbf{A}}\mathbf{C} - \underline{\mathbf{A}}\mathbf{D} \right\},$$

subject to $c \in \mathbb{R}$, $\mathbf{C}, \mathbf{D} \in \mathbb{R}^r_+$, and

$$c + \mathbf{H}_i \left( \mathbf{C} - \mathbf{D} \right) \geq V(\varphi, b, \mathbf{x}_i), \;\; i = 1, ..., n.$$

Here $c$, $\mathbf{C}, \mathbf{D}$ are optimization variables. The variable $c$ in the dual form corresponds to the constraint $\sum_{i=1}^{n} p_i = 1$ in the primal form. The variables $\mathbf{C}$ correspond to the constraints $\mathbb{E}_p h \leq \overline{a}_k$ in the primal form, and the variables $\mathbf{D}$ correspond to the constraints $\underline{a}_k \leq \mathbb{E}_p h$.

Now we have two optimization problems with the same objective function $c + \overline{\mathbf{A}}\mathbf{C} - \underline{\mathbf{A}}\mathbf{D}$ to be minimized. This implies that minimization over $\varphi, b$ can simply be combined with the dual optimization problem as follows:

$$\min_{c, \mathbf{C}, \mathbf{D}, \varphi, b, \xi} \left\{ c + \overline{\mathbf{A}}\mathbf{C} - \underline{\mathbf{A}}\mathbf{D} \right\}, \tag{4}$$

subject to $c \in \mathbb{R}$, $\mathbf{C}, \mathbf{D} \in \mathbb{R}^r_+$, (1), (2) and

$$c + \mathbf{H}_i (\mathbf{C} - \mathbf{D}) - V(\varphi, b, \mathbf{x}_i) \geq 0, \ i = 1, ..., n. \tag{5}$$

We obtain the linear optimization problem with variables $c$, $\mathbf{C}$, $\mathbf{D}$, $\varphi$, $b$, $\xi$.

When we know precise values $a_k$ of expectations $\mathbb{E}_p h_k$, then the problem can be simplified as

$$\min_{c, \mathbf{C}, \varphi, b, \xi} \{c + \mathbf{AC}\}, \tag{6}$$

subject to $c \in \mathbb{R}$, $\mathbf{C} \in \mathbb{R}^r$, (1), (2) and

$$c + \mathbf{H}_i \mathbf{C} - V(\varphi, b, \mathbf{x}_i) \geq 0, \ i = 1, ..., n. \tag{7}$$

One of the important advantages of the dual form of the classification model is that we do not need to assume that features are statistically independent because we consider only joint probability distributions even if prior knowledge concerns separate features.

### 3.2 Extreme points for solving the problem

The approach using extreme points has been studied by Augustin [1] and successfully applied to many problems (see, for instance, [51]). It can be very useful when we can easy find extreme points of the set $\mathcal{P}$ denoted as $\mathcal{E}(\mathcal{P})$. According to this approach, a new variable

$$G = \max_{p \in \mathcal{P}} \sum_{i=1}^{n} p_i V_i(\varphi, b) \tag{8}$$

is introduced. Then problem (3) can be rewritten as

$$\min_{\varphi, b, \xi, G} G, \tag{9}$$

subject to (1), (2) and

$$G \geq \sum_{i=1}^{n} p_i V(\varphi, b, \mathbf{x}_i), \ \forall p \in \mathcal{P}. \tag{10}$$

The above optimization problem contains infinitely many constraints (all probability distributions in $\mathcal{P}$). In order to overcome this difficulty, note, however, that the set of distributions $\mathcal{P}$ can be viewed as a simplex in a finite dimensional space. According to some general results from linear programming theory, an optimal solution to the above problem is achieved at extreme points $\mathcal{E}(\mathcal{P})$ of the simplex, and the number of its extreme points is finite. This implies that the infinite set of constraints is reduced to a set with some finite number of constraints, and standard routines for linear programming can be used to determine the optimal solution.

Moreover, the above optimization problem can be represented as a set of $t$ optimization problems where $t$ is the number of extreme points in $\mathcal{P}$. We assume that there are no identical sums $\sum_{i=1}^{n} p_i V(\varphi, b, \mathbf{x}_i)$ by different $p$ from $\mathcal{E}(\mathcal{P})$. If this assumption is valid, then there is a probability distribution $p$ from $\mathcal{E}(\mathcal{P})$, say $p^{(k)}$, such that there holds

$$G = \sum_{i=1}^{n} p_i^{(k)} V(\varphi, b, \mathbf{x}_i).$$

Here $k$ is the number of an extreme point for which the above equality is valid. This equality follows from the definition of the variable $G$ given by (8), which can be viewed as a linear optimization problem with variables $p$ restricted by the same set $\mathcal{P}$ with extreme points $\mathcal{E}(\mathcal{P})$.

This implies that problem (9)-(10) can be rewritten as the set of $t$ linear optimization problems

$$W_k(\varphi, b) = \min_{\varphi, b, \xi} \sum_{i=1}^{n} p_i^{(k)} V(\varphi, b, \mathbf{x}_i), \tag{11}$$

subject to (1), (2).

Here $k = 1, ..., n$. Note that the largest value of $G$ is chosen from all possible its values. Therefore, the optimal values of variables $\varphi, b, \xi$ correspond to the largest value of $W_k(\varphi, b)$.

It should be noted that the second approach requires to search all extreme points. This is computationally hard task in some cases and may lead to solution of a huge number of linear optimization problems. The first approach allows us to avoid this difficulty and to solve only one linear programming problem.

Below, we will consider two important special cases of prior knowledge.

## 4 A robust contamination neighborhood model

One of the well-known classes of robust models is based on relaxing strong assumptions about a probability distribution of data points. It calls the $\varepsilon$-contaminated (robust) model ([21]). The model is constructed by eliciting a Bayesian prior distribution $p = (1/n, ..., 1/n)$ as an estimate of the true prior distribution. The $\varepsilon$-contaminated model is a class of probabilities which for fixed $\varepsilon \in (0, 1)$ and $p_i$ is the set $\mathcal{P}(\varepsilon) = \{(1 - \varepsilon)/n + \varepsilon q_i\}$, where $q = (q_1, ..., q_n)$ satisfies the conditions $q_1 + ... + q_n = 1$, $q_i \geq 0$, $i = 1, ..., n$. The rate $\varepsilon$ reflect the amount of uncertainty in $p$ [4]. The choice of the probability distributions $q$ totally defines the contaminated model. Walley [54] proposed the imprecise linear-vacuous mixture model for which the distribution $q$ is assumed to be arbitrary, i.e., it can be arbitrary from the unit simplex denoted by $S(1, n)$. According to this model, for $0 < \varepsilon < 1$, $\mathcal{P}(\varepsilon)$ is the set of all probabilities with the lower bound $(1-\varepsilon)/n$ and the upper bound $(1-\varepsilon)/n+\varepsilon$. Of course, the assumption that $q$ is restricted by the unit simplex $S(1, n)$ is

one of possible types of the $\varepsilon$-contaminated model. Generally, there are a lot of different assumptions which produce specific robust models.

### 4.1 Extreme points

Let us find extreme points of the set $\mathcal{P}(\varepsilon)$ produced by the imprecise linear-vacuous mixture model. Extreme points of the simplex $S(1,n)$ are of the form:

$$(1,0,...,0),\ (0,1,...,0),...,\ (0,0,...,1). \tag{12}$$

Hence, we get extreme points of the set $\mathcal{P}(\varepsilon)$ as

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

Then we can rewrite constraints (10) by taking into account the set of the above extreme points

$$G \geq \frac{1-\varepsilon}{n}\sum_{i=1}^{n} V(\varphi,b,\mathbf{x}_i)+\varepsilon V(\varphi,b,\mathbf{x}_k),\ \ k=1,...,n. \tag{13}$$

So, we obtain the linear programming problem with the objective function (9) and $n$ constraints (13), $3n+1$ constraints (1), (2). It is obvious that the above problem can be replaced by $n$ linear programming problems

$$Z_k(\varphi,b)=\min_{\varphi,b,\xi}\left(\frac{1-\varepsilon}{n}\sum_{i=1}^{n} V(\varphi,b,\mathbf{x}_i)+\varepsilon V(\varphi,b,\mathbf{x}_k)\right),$$

subject to (1), (2).

From all optimal values $Z_k(\varphi,b)$, $k=1,...,n$, we take the largest one.

### 4.2 Dual problem

It should be noted that the same problem can be solved by means of the dual form approach. Indeed, the set $\mathcal{P}(\varepsilon)$ is produced by constraints $p_1+...+p_n=1$ and

$$(1-\varepsilon\}/n \leq p_i \leq (1-\varepsilon\}/n+\varepsilon.$$

Hence, $h_k(i)=\mathbf{1}_k(i)$, $\underline{a}_k=(1-\varepsilon)/n$, $\overline{a}_k=(1-\varepsilon)/n+\varepsilon$, $k=1,...,n$. Here $\mathbf{1}_k(i)$ is the indicator function taking the value 1 if $k=i$. The dual problem is of the form:

$$\overline{W}(\varphi,b)=\min_{c,\mathbf{C},\mathbf{D},\varphi,b,\xi}\left\{c+\left(\frac{1-\varepsilon}{n}+\varepsilon\right)\sum_{k=1}^{n} c_k-\frac{1-\varepsilon}{n}\sum_{k=1}^{n} d_k\right\}$$

subject to $c_k \geq 0$, $d_k \geq 0$, $k = 1, ..., n$, (1), (2) and

$$c + c_k - d_k \geq V(\varphi, b, \mathbf{x}_k), \ k = 1, ..., n.$$

Here the number of constraints can be decreased by introducing new variables $\mathbf{G} = \mathbf{C} - \mathbf{D}$. Then there holds

$$\overline{W}(\varphi, b) = \min_{c, \mathbf{C}, \mathbf{G}, \varphi, b, \xi} \left\{ c + \varepsilon \sum_{k=1}^{n} c_k + \frac{1-\varepsilon}{n} \sum_{k=1}^{n} g_k \right\}$$

subject to $c_k \geq 0$, $k = 1, ..., n$, (1), (2) and

$$c + g_k \geq V(\varphi, b, \mathbf{x}_k), \ k = 1, ..., n.$$

It can be seen from the above problem that $c_k = 0$ for all $k = 1, ..., n$. This implies that

$$\overline{W}(\varphi, b) = \min_{c, \mathbf{G}, \varphi, b, \xi} \left\{ c + \frac{1-\varepsilon}{n} \sum_{k=1}^{n} g_k \right\}$$

subject to (1), (2) and

$$c + g_k \geq V(\varphi, b, \mathbf{x}_k), \ k = 1, ..., n.$$

Finally, we have two algorithms for computing optimal parameters $\varphi_i$, $i = 1, ..., n$, and $b$ of the linear OCC model. Every algorithm has virtues and shortcomings. The algorithm using the duality technique requires to solve a single optimization problem with $4n + 1$ constraints and $2n + 2$ variables. The algorithm based on extreme points requires to solve $n$ linear programming problems with $3n + 1$ constraints and $2n + 1$ variables.

## 5 Known mean values and second moments

So far we have studied the case when extreme points of $\mathcal{P}$ can simply be found. In this case, both approaches (extreme points and the dual form) are efficient and can be implemented. However, there are a lot of judgments which restrict the set $\mathcal{P}$ in such a way that it is very difficult to find and to enumerate its extreme points. In this case, we should use only the dual form approach. Below we consider one of such cases when we know mean values and second moments of some features. It is interesting to note that this information allows us to construct the lower and upper bounds for the cumulative distribution function by using the Chebychev-Cantelli inequality [8] which is also known as the one-sided version of Chebyshev's inequality. However, we apply another way to solving the classification problem, which is simple and efficient. Moreover, the proposed way does not use the condition of feature independence, which is often criticized in Naive Bayes classification models.

So, we suppose now that the mean values of some features and their second moments are known. Denote the sets of $l$ feature indices whose mean values

$m_k$ are known by $\mathcal{J}$ and whose standard deviations $\sigma_k^2$ or second moments $\sigma_k^2 + m_k^2$ are known by $\mathcal{K}$, respectively, i.e., there hold

$$\mathcal{J} = \{k \in \{1, ..., l\} : m_k \text{ is known}\},$$

$$\mathcal{K} = \{k \in \{1, ..., l\} : \sigma_k \text{ is known}\}.$$

In particular, $\mathcal{J} = \mathcal{K}$ when we simultaneously know the first and the second moments of features.

Note that the moments of the $k$-th feature can be represented in the form of expectations with respect to the distribution $p$ as follows:

$$m_k = \mathbb{E}_p x^{(k)} = \sum_{i=1}^{n} \mathbf{x}_i^{(k)} p_i,$$

$$\sigma_k^2 = \mathbb{E}_p \left(x^{(k)}\right)^2 - m_k^2.$$

Then there holds

$$\mathbb{E}_p \left(x^{(k)}\right)^2 = \sigma_k^2 + m_k^2.$$

We may have imprecise information about $m_k$ in the form of interval $[\underline{m}_k, \overline{m}_k]$. The same can be said about the interval of standard deviations $[\underline{\sigma}_k, \overline{\sigma}_k]$. In this case, the lower and upper second moments are

$$\underline{\mathbb{E}}_p \left(x^{(k)}\right)^2 = \underline{\sigma}_k^2 + \underline{m}_k^2, \ \overline{\mathbb{E}}_p \left(x^{(k)}\right)^2 = \overline{\sigma}_k^2 + \overline{m}_k^2.$$

In spite of a certain incorrectness of the above bounds for the interval of second moments, we will use them because they include the correct bounds, i.e., they can be regarded as an approximation of the interval-valued second moment.

Hence, we can write the following linear programming problem (see the dual optimization problem (6)-(7)):

$$\min_{c, c_k, a_k, c_k^*, a_k^*, \varphi, b, \xi} \left\{ c + \sum_{k \in \mathcal{J}} \left(\overline{m}_k c_k - \underline{m}_k c_k^*\right) + \sum_{k \in \mathcal{K}} \left(\overline{\sigma}_k^2 + \overline{m}_k^2\right) a_k - \sum_{k \in \mathcal{K}} \left(\underline{\sigma}_k^2 + \underline{m}_k^2\right) a_k^* \right\},$$

subject to $c \in \mathbb{R}$, $c_k, a_k, c_k^*, a_k^* \in \mathbb{R}_+$, (1), (2) and

$$c + \sum_{k \in \mathcal{J}} \mathbf{x}_i^{(k)} (c_k - c_k^*) + \sum_{k \in \mathcal{K}} \left(\mathbf{x}_i^{(k)}\right)^2 (a_k - a_k^*) - V(\varphi, b, \mathbf{x}_i) \geq 0, \ i = 1, ..., n.$$

In particular, if we have precise values of $m_k$ and $\sigma_k^2$, then there holds:

$$\min_{c,\mathbf{C},\mathbf{D},\varphi,b,\xi} \left\{ c + \sum_{k\in\mathcal{J}} m_k c_k + \sum_{k\in\mathcal{K}} \left(\sigma_k^2 + m_k^2\right) a_k \right\},$$

subject to $c, c_k, a_k \in \mathbb{R}$, (1), (2) and

$$c + \sum_{k\in\mathcal{J}} \mathbf{x}_i^{(k)} c_k + \sum_{k\in\mathcal{K}} \left(\mathbf{x}_i^{(k)}\right)^2 a_k - V(\varphi, b, \mathbf{x}_i) \geq 0,\ i = 1, ..., n.$$

## 6 Numerical 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 model and compare it with C-B model 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 measure of the classifier performance. ACC can formally be written as

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

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 proposed method as $ACC_{\text{imp}}$, C-B model as $ACC_{\text{CB}}$. For some experiments, we compare the proposed model with the well known model provided by Scholkopf at el. [38], whose accuracy measure is denoted as $ACC_{\text{S}}$.

All experiments use a standard Gaussian radial basis function (RBF) kernel with the kernel parameter $\gamma$. Different values for the parameter $\gamma$ 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 $\mathbf{m}(1) = (m_1(1), m_2(1))$, and $N_2 = \varepsilon_0 N$ examples (the second subset) have mean values $\mathbf{m}(2) = (m_1(2), m_2(2))$. The standard deviation is $\sigma = 2$ for both subsets and both features. Here $\varepsilon_0$ is a portion of abnormal examples in the training set, $m_1(i)$, $m_2(i)$ are mean values of the first and the second features. The parameters $\nu$ and $\varepsilon_0$ are 0.2 and 0.8, respectively.

For all experiments with synthetic data, we use the additional information in the form of the lower $\underline{m}$ and upper $\overline{m}$ mean values and the precise $\sigma$ standard

**Table 1** The classification performance by different values of $n$

| $n$ | $ACC_{\text{imp}}$ | $ACC_{\text{CB}}$ |
|---|---|---|
| 10 | 0.814 | 0.732 |
| 20 | 0.862 | 0.73 |
| 30 | 0.871 | 0.717 |
| 40 | 0.86 | 0.696 |
| 50 | 0.907 | 0.727 |
| 60 | 0.897 | 0.727 |
| 70 | 0.909 | 0.713 |
| 80 | 0.906 | 0.718 |
| 90 | 0.924 | 0.714 |
| 100 | 0.913 | 0.73 |

deviation of features. It is important to point out here that the random observations are generated with respect to the normal probability distributions with mean values $\mathbf{m}(1) = (4, 4)$, $\mathbf{m}(2) = (12, 12)$ and standard deviations $\sigma = 2$. It is supposed that this information is just for generating random observations, and it is unknown for the classifier. But the lower and upper mean values and standard deviations for the whole sample as prior additional information for the proposed classification model is provided before classifying. This information does not imply that $\mathbf{m}(1)$, $\mathbf{m}(2)$ must be in the interval $[\underline{m}, \overline{m}]$. By means of values $\underline{m}$, $\overline{m}$ and the sample standard deviation we try to "guess" the values $\mathbf{m}(1)$, $\mathbf{m}(2)$, $\sigma$, respectively, but not to copy them.

First, we investigate how the accuracy measures depend on the number of examples in the training set. We accept $\underline{m} = 3.5$ and $\overline{m} = 4$. The parameter $\gamma$ is 32. One can see from Table 1 and from Fig. 1 that the proposed model outperforms the C-B novelty detection model for all $n$ from 10 to 100. The solid curve with triangle markers in Fig. 1 corresponds to the proposed model and the dashed curve with round markers corresponds to the C-B model. This implies that the assumption of the empirical probability distribution of examples accepted in the C-B novelty detection provides inferior accuracy in comparison with the knowledge of interval-value mean values and standard deviation.

Contours $f(\mathbf{x}) = 0$ and generated data points with the above parameters shown in Fig. 2 illustrate how numbers of examples in a training set impact on the form of contours. Pictures (a), (b), (c), (d), (e), (f) in Fig. 2 correspond to the proposed model (thick curve) and the C-B model (thin curve) by $n = 10$, 20, 30, 40, 50, 60, respectively. It is clearly seen from pictures that prior information about mean values and standard deviations leads to contours which cover mainly normal examples. In contrast to the proposed model, C-B model provides worse results. One can see from Fig. 2 that contours corresponding to C-B model cover many abnormal points.

Second, we investigate how the accuracy measure for the proposed model depends on the interval $[\underline{m}, \overline{m}]$. The simulation results are shown in Table 2. We use index 1 for the accuracy measure of the proposed model (see the third column in Table 2). The parameter $\gamma$ is 32. We have to also note that

$ACC_{\rm CB} = 0.687$ and $ACC_{\rm S} = 0.622$. The above measures do not depend on the interval $[\underline{m}, \overline{m}]$ and on knowledge of mean values. It can be seen from the table that there are some values of the lower and upper mean values as well as the width $\overline{m} - \underline{m}$ of the interval for which the accuracy measure $ACC_{\rm imp1}$ is the largest one. It is 0.868 for intervals with upper bound 4 and for the width of the interval larger than 1. It can be explained as follows. Note that the value 4 of the mean value is taken for generating the random normal examples. The abnormal examples having the mean value 12 bias the mean value of the whole training set. As a result, the standard C-B model "covers" some number of abnormal examples taking them into account. By reducing the lower mean value (by taking the lower bound to be 3, 2, 1, 0), we try to correct the probability distribution functions produced by the given information in the form of mean values and standard deviations. Contours $f(\mathbf{x}) = 0$ and generated data points with the above parameters shown in Fig. 3 illustrate how the interval of mean values $[\underline{m}, \overline{m}]$ impact on the form of contours. Pictures (a), (b), (c), (d) in Fig. 3 correspond to the proposed model (thick curve) and the C-B model (thin curve) by $[\underline{m}, \overline{m}] = [4; 4]$, $[3; 4]$, $[2; 4]$, $[2; 5]$, respectively. One can see from pictures that too precise information (interval $[4; 4]$) may lead to the worse results. The same can be said about too imprecise information (interval $[2; 5]$). This is a very interesting observation. On the one hand, we should not use the precise mean values. Only imprecise estimates of mean values lead to outperforming results. On the other hand, the large imprecision of mean values may lead to the overcautious solution which reduces the classifier quality. However, the proper choice of the interval of mean values provides the largest accuracy measure.

So far we have evaluated the proposed model by using the same normal probability distribution of abnormal observations. Now we consider how the accuracy measure for the proposed model depends on the interval of mean values when abnormal observations are governed by the uniform probability distributions $U(a, b)$ where $a$ and $b$ are the lower and upper bounds of the distribution's support. We take $a = -8$ and $b = 16$ for both features such that their mean values $(4, 4)$ coincide with the mean values of normal examples. The sample standard deviation of all examples is about 3 for both features. The parameter $\gamma$ is 22. The simulation results are shown in Table 2 (the fourth column, $ACC_{\rm imp2}$). We use index 2 for the accuracy measure of the proposed model. The corresponding accuracy measures for C-B model and for the model proposed by Scholkopf at el. [38] are $ACC_{\rm CB} = 0.875$ and $ACC_{\rm S} = 0.748$, respectively. One can see from Table 2 that the proposed model outperforms C-B model almost for all interval-valued mean values. However, the difference between $ACC_{\rm imp2}$ and $ACC_{\rm CB}$ is not so large as in the previous example because, by taking the uniform probability distributions of the abnormal examples with the mean values $(4, 4)$ of features, we in fact reduce the value of $\varepsilon_0$.

The proposed model has been evaluated and investigated by the following publicly available data sets: Iris, Mammographic masses, Parkinsons, Indian Liver Patient Dataset, Lung cancer, Breast tissue. All data sets are from the

**Table 2** The classification performance by different lower and upper mean values of features

| $\underline{m}$ | $\overline{m}$ | $ACC_{\text{imp1}}$ | $ACC_{\text{imp2}}$ |
|---|---|---|---|
| 4 | 4 | 0.632 | 0.806 |
| 3.5 | 4 | 0.867 | 0.898 |
| 3 | 4 | 0.868 | 0.898 |
| 2 | 4 | 0.868 | 0.898 |
| 1 | 4 | 0.868 | 0.898 |
| 0 | 4 | 0.868 | 0.898 |
| 4 | 5 | 0.838 | 0.894 |
| 3 | 5 | 0.838 | 0.894 |
| 2 | 5 | 0.838 | 0.894 |
| 2 | 6 | 0.801 | 0.890 |

UCI Machine Learning Repository [15]. The following is a brief introduction about these datasets, while more detailed information can be found from, respectively, the data resources.

Iris data set contains 3 classes (Iris Setosa, Iris Versicolour, Iris Virginica) of 50 instances each. The number of features is 4 (sepal length, sepal width, petal length, petal width). It is supposed that data points from the Iris Setosa class are abnormal, i.e., the number of abnormal examples is 30.

Mammographic masses (MM) data set contains 961 instances characterized by 4 predictive features (age, shape, margin, density). Classes correspond to severity (benign and malignant). We attribute the malignant instances to abnormal observations because their number 445 smaller than the number of benign instances

Parkinsons data set is composed of a range of biomedical voice measurements from healthy people and people with Parkinsons disease. It contains 195 instances characterized by 23 predictive features, including the average vocal fundamental frequency, maximum and minimum vocal fundamental frequencies, several measures of variation, etc. We attribute 48 instances corresponding to healthy people to abnormal observations.

Indian Liver Patient Dataset (ILPD) contains 416 liver patient records and 167 non-liver patient records characterized by 10 predictive features (Age of the patient, Gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphotase, Alamine Aminotransferase, Aspartate Aminotransferase, Total Protiens, Albumin, Ratio Albumin and Globulin Ratio). Non-liver patient records are viewed as abnormal.

Lung cancer data set describes 3 types of pathological lung cancers. It contains 32 instances having 57 predictive features. All predictive attributes are nominal, taking on integer values from 0 to 3. 9 instances corresponding to the first type of pathological lung cancers are viewed as abnormal.

Breast tissue data set contains 106 instances characterized by 9 numerical-valued features. Every instance corresponds to one of the six classes of the freshly excised tissue studied using electrical impedance measurements: carcinoma, fibro-adenoma, mastopathy, glandular, connective, adipose. The first

**Table 3** Initial data for comparison of the proposed model with C-B model

| | $\gamma$ | $v$ | $\underline{m}_1$ | $\overline{m}_1$ | $\underline{m}_2$ | $\overline{m}_2$ | $\sigma_1$ | $\sigma_2$ |
|---|---|---|---|---|---|---|---|---|
| Iris | 7 | 0.3 | 5 | 7 | 2 | 4 | 0.83 | 0.43 |
| MM | 3.2 | 0.4 | 3 | 7 | 50 | 60 | 1.78 | 14.4 |
| Parkinsons | 32 | 0.2 | 150 | 160 | 190 | 200 | 41.4 | 91.5 |
| ILPD | 16 | 0.01 | 40 | 50 | 0.2 | 0.3 | 16.2 | 0.43 |
| Lung cancer | 10 | 0.1 | 0.03 | 0.04 | 2 | 3 | 0.18 | 0.55 |
| Breast tissue | 71 | 0.05 | 700 | 800 | 0.1 | 0.2 | 754 | 0.07 |

**Table 4** The classification performance for different data sets

| | $ACC_{\text{imp}}$ | $ACC_{\text{CB}}$ |
|---|---|---|
| Iris | 0.853 | 0.66 |
| MM | 0.563 | 0.537 |
| Parkinsons | 0.738 | 0.641 |
| ILPD | 0.719 | 0.659 |
| Lung cancer | 0.594 | 0.719 |
| Breast tissue | 0.66 | 0.632 |

two classes (carcinoma, fibro-adenoma) are contain 36 instances and are labeled as negative ($y = -1$).

Parameters for evaluating the models are given in Table 3. They are provided in order to have an opportunity to repeat the experiments. Parameters $\gamma$ and $v$ have been changed in order to maximize the accuracy measure. For the first and the second features, we calculate their sample means and sample standard deviations, i.e., it is assumed that $\mathcal{J} = 2$. Then we take the lower and upper mean values of every feature such that the sample mean lies in the interval $[\underline{m}_i, \overline{m}_i]$, $i = 1, 2$. The standard deviations are taken precise such that they coincide with the sample standard deviations. In spite of the fact that the number of features in every data set is larger than 2, we assume that there is information in the form of interval-valued expectations and standard deviations only about two features for short. The corresponding accuracy measures for the above data sets are shown in Table 4. One can see that the accuracy measure of the proposed model outperforms the same measure of C-B model almost for all data sets except for Lung cancer data set. It is interesting to note that the balance of $ACC_{\text{imp}}$ and $ACC_{\text{CB}}$ for Lung cancer data set does not change for various values of parameters. The largest difference between accuracy measures of two models takes place for Iris data set.

Another part of experiments is devoted to analysis of Walley's linear-vacuous mixture model. According to this model, the corresponding classifier deals with the set $\mathcal{P}(\varepsilon)$ of probability distributions produced by the linear-vacuous mixture model or the $\varepsilon$-contaminated model. The information we have is the contamination parameter $\varepsilon$. We investigate how the accuracy measures of the classifiers depend on parameters $\varepsilon$, $\varepsilon_0$ and $n$. We are interesting to find out whether there is a relationship between the portion of contaminated examples $\varepsilon_0$ and the parameter $\varepsilon$ or not. We have to mention here that $\varepsilon_0$ is a parameter of generated data, and $\varepsilon$ is the classification model parameter.

**Table 5** The classification performance by different values of $n$, $\varepsilon_0$ and by $\varepsilon = 0.2$

| | $n = 20$ | | $n = 30$ | | $n = 40$ | | $n = 50$ | |
|---|---|---|---|---|---|---|---|---|
| $\varepsilon_0$ | Imp | CB | Imp | CB | Imp | CB | Imp | CB |
| 0.1 | 0.779 | 0.777 | 0.796 | 0.785 | 0.801 | 0.789 | 0.800 | 0.790 |
| 0.2 | 0.745 | 0.739 | 0.709 | 0.703 | 0.729 | 0.720 | 0.726 | 0.716 |
| 0.3 | 0.660 | 0.655 | 0.656 | 0.650 | 0.667 | 0.660 | 0.643 | 0.651 |
| 0.4 | 0.576 | 0.581 | 0.573 | 0.576 | 0.579 | 0.587 | 0.596 | 0.593 |

**Table 6** The classification performance by different values of $n$, $\varepsilon_0$ and by $\varepsilon = 0.2$

| | $n = 20$ | | $n = 30$ | | $n = 40$ | | $n = 50$ | |
|---|---|---|---|---|---|---|---|---|
| $\varepsilon_0$ | Imp | CB | Imp | CB | Imp | CB | Imp | CB |
| 0.1 | 0.776 | 0.777 | 0.786 | 0.785 | 0.779 | 0.789 | 0.795 | 0.790 |
| 0.2 | 0.743 | 0.739 | 0.727 | 0.703 | 0.718 | 0.720 | 0.735 | 0.716 |
| 0.3 | 0.660 | 0.655 | 0.652 | 0.650 | 0.639 | 0.660 | 0.652 | 0.651 |
| 0.4 | 0.576 | 0.581 | 0.566 | 0.576 | 0.588 | 0.587 | 0.599 | 0.593 |

Again we exploit synthetic data having two features $x_1$ and $x_2$ and the parameters which have been used in experiments with the known mean values and standard deviations, i.e., $\mathbf{m}(1) = (4, 4)$, $\mathbf{m}(2) = (12, 12)$ and $\sigma = 2$. The parameter $\gamma$ is 32. Typical contours $f(\mathbf{x}) = 0$ and $n = 40$ generated data points with the above parameters are shown Fig.4 where pictures (a), (b), (c), (d), (e), (f) correspond to the proposed model (thick curve) and the C-B model (dashed curve) by $\varepsilon = 0.1$, 0.2, 0.3, 0.4, 0.5, 0.6, respectively.

The accuracy measures of the proposed model and C-B model by $\varepsilon = 0.2$ are shown in Table 5. We can see from the table that the proposed model provides better results in comparison with C-B model for most values of $\varepsilon_0$ and $n$. At the same time, the largest difference between accuracy measures $ACC_{\text{imp}} - ACC_{\text{C-B}}$ takes place for $\varepsilon_0 = 0.1$ or 0.2. Moreover, one can see that the accuracy measure of C-B model is larger than the same measure of the proposed model in some cases of large values of $\varepsilon_0$, for example, $\varepsilon_0 = 0.4$. The accuracy measures of the proposed model and C-B model by $\varepsilon = 0.4$ are shown in Table 6. One can see from the table that results provided by the proposed model become better for $\varepsilon_0 = 0.3$ or 0.4. We can conclude from these results that the quality of the proposed model depends on the relationship between $\varepsilon_0$ and $\varepsilon$. The largest accuracy is achieved when the values of $\varepsilon_0$ and $\varepsilon$ are close to each other.

## 7 Conclusion

Robust OCC models have been proposed in the paper, which are based on incorporating prior knowledge into Campbell and Bennett linear model using the imprecise statistical models instead of the empirical probability distribution accepted in the standard SVM. The models are represented in two forms. The first form uses extreme points of the set of probability distributions produced by an imprecise statistical model, for example, by the imprecise linear-vacuous

mixture model [54]. The second form is based on the powerful conjugate duality technique [28]. Both the forms have virtues and shortcomings which depend on the application problem. Classification parameters of every model are obtained by solving a finite set of simple linear programming problems (models based on extreme points) or a single linear problem (models based on the duality technique). The algorithm for computing the optimal parameters of every OCC model can be easily implemented with standard functions of the statistical software package R. The proposed approaches for constructing OCC models are rather general because they allow us to construct various algorithms by applying different imprecise statistical models or different prior knowledge.

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. This violation is taken into account by considering the set of probability distributions. This implies that the proposed models can be successfully applied to noisy data and to problems when there is a small training set.

The most interesting case of prior knowledge in the form of the interval-valued first and second moments is based on using the duality technique. Experimental studies have illustrated that the proposed model with such the prior information outperforms C-B model in most cases. At the same time, it is difficult to call the information incorporated into proposed model in the experiments (Section 6) by prior information because it is computed from the available training set. Indeed, we took the lower $\underline{m}_i$ and upper $\overline{m}_i$ mean values of every feature in examples with real data sets such that its sample mean lies in the interval $[\underline{m}_i, \overline{m}_i]$. The standard deviations were taken precise coinciding with the sample standard deviations. The fact of outperforming of the proposed model and the successful experimental studies open ways to develop a series of classification models using "internal" information as prior knowledge.

Another important direction for future work is to study and develop robust classification models based on incorporating prior knowledge different from those considered in the paper, for example, imprecise comparative judgements, imprecise second-order uncertainty models [48–50]. Finally, it is also worth noticing that the proposed models can easily be extended on the case of binary or multi-class classification.

## Acknowledgement

We would like to express our appreciation to the anonymous referees and the editor whose very valuable comments have improved the paper.

## References

1. Augustin, T.: Expected utility within a generalized concept of probability - a comprehensive framework for decision making under ambiguity. Statistical Papers **43**, 5–22

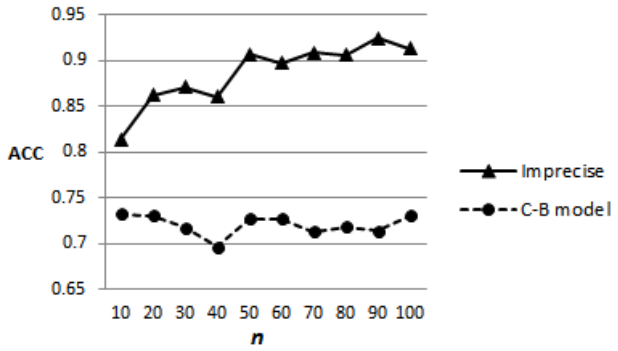


**Fig. 1** Accuracy measures for two models as a function of $n$

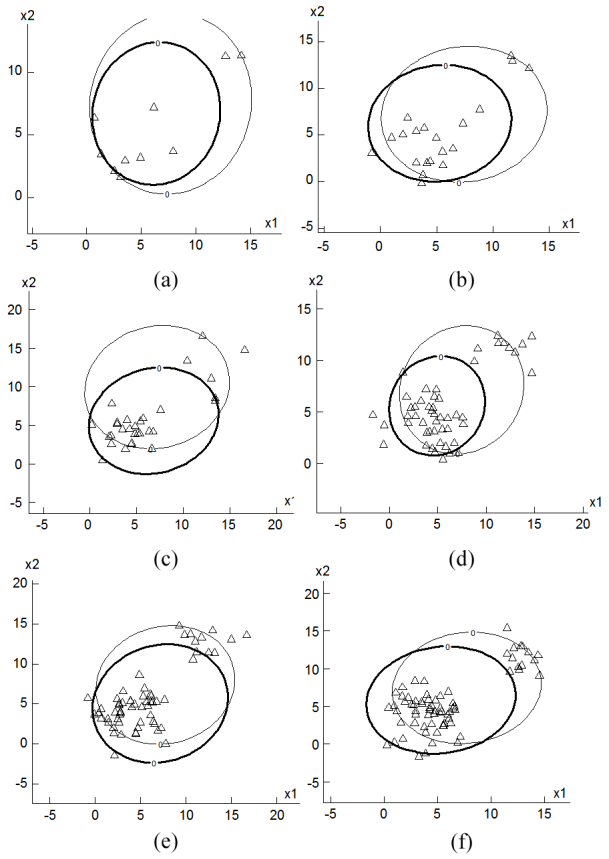


**Fig. 2** Contours $f(\mathbf{x}) = 0$ for two models by different values of $n$

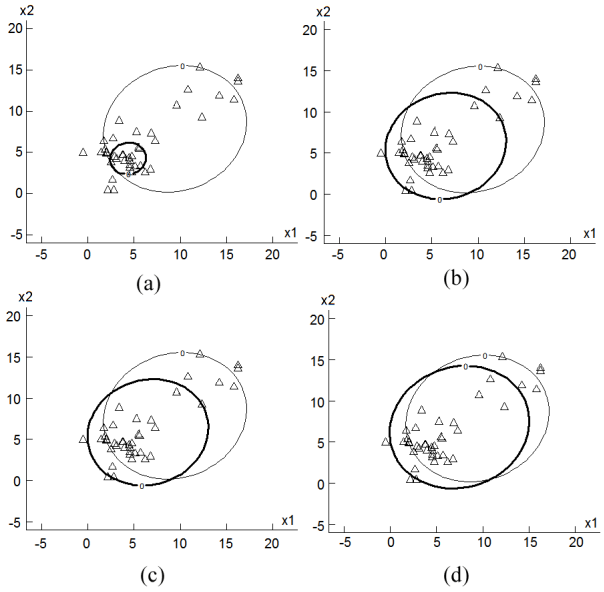


**Fig. 3** Contours $f(\mathbf{x}) = 0$ for two models by different intervals $[\underline{m}, \overline{m}]$

(2002)
2. Bartkowiak, A.: Anomaly, novelty, one-class classification: A comprehensive introduction. International Journal of Computer Information Systems and Industrial Management Applications **3**, 61–71 (2011)
3. Ben-Tal, A., El Ghaoui, L., Nemirovski, A.: Robust Optimization. Princeton University Press, Princeton and Oxford (2009)
4. Berger, J.: Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York (1985)
5. Bicego, M., Figueiredo, M.: Soft clustering using weighted one-class support vector machines. Pattern Recognition **42**, 27–32 (2009)
6. Campbell, C.: Kernel methods: a survey of current techniques. Neurocomputing **48**(1-4), 63–84 (2002)
7. Campbell, C., Bennett, K.: A linear programming approach to novelty detection. In: T. Leen, T. Dietterich, V. Tresp (eds.) Advances in Neural Information Processing Systems, vol. 13, pp. 395–401. MIT Press (2001)
8. Cantelli, F.: Intorno ad un teorema fundamentale della teoria del rischio. Boll. Assoc. Attuar. Ital. (Milan) pp. 1–23 (1910)
9. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. Tech. Rep. TR 07-017, University of Minnesota, Minneapolis, MN, USA (2007)
10. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys **41**, 1–58 (2009)
11. Chapelle, O., Scholkopf, B.: Incorporating invariances in non-linear support vector machines. In: T. Dietterich, S. Becker, Z. Ghahraman (eds.) Advances in Neural Information Processing Systems, pp. 609–616. MIT Press, Cambridge, MA, USA (2001)
12. Cherkassky, V., Mulier, F.: Learning from Data: Concepts, Theory, and Methods. Wiley-IEEE Press, UK (2007)
13. Dayanik, A., Lewis, D., Madigan, D., Menkov, V., Genkin, A.: Constructing informative prior distributions from domain knowledge in text classification. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 493–500. ACM, New York, NY, USA (2006)
14. Decoste, D., Schölkopf, B.: Training invariant support vector machines. Machine Learning **46**(1-3), 161–190 (2002)

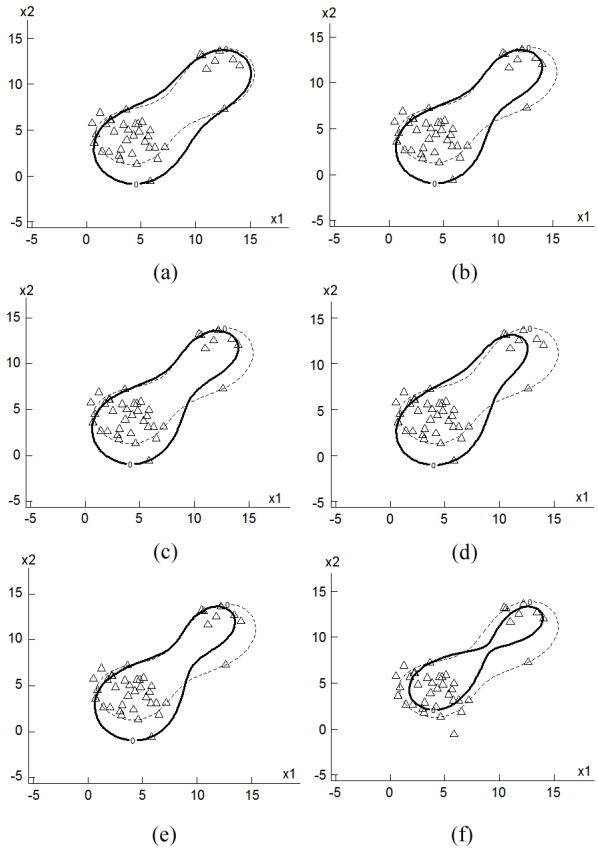


**Fig. 4** Contours $f(\mathbf{x}) = 0$ for two models by different values of $\varepsilon$

15. Frank, A., Asuncion, A.: UCI machine learning repository (2010). URL http://archive.ics.uci.edu/ml
16. Fung, G., Mangasarian, O., Shavlik, J.: Knowledge-based support vector machine classifiers. In: S. Becker, S. Thrun, K. Obermayer (eds.) Advances in Neural Information Processing Systems, pp. 521–528. MIT Press, Cambridge, MA, USA (2002)
17. Gao, Y., Gao, F.: Edited adaboost by weighted knn. Neurocomputing **73**(16-18), 3079–3088 (2010)
18. Gilboa, I., Schmeidler, D.: Maxmin expected utility with non-unique prior. Journal of Mathematical Economics **18**(2), 141–153 (1989)
19. Haasdonk, B., Vossen, A., Burkhardt, H.: Invariance in kernel methods by haar-integration kernels. In: H. Kalviainen, J. Parkkinen, A. Kaarna (eds.) Image Analysis, *Lecture Notes in Computer Science*, vol. 3540, pp. 841–851. Springer Berlin Heidelberg (2005)
20. Hodge, V., Austin, J.: A survey of outlier detection methodologies. Artificial Intelligence Review **22**(2), 85–126 (2004)
21. Huber, P.: Robust Statistics. Wiley, New York (1981)
22. Joachims, T.: Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Kluwer Academic Publishers, Norwell, MA, USA (2002)
23. Kunapuli, G., Bennett, K., Shabbeer, A., Maclin, R., Shavlik, J.: Online knowledge-based support vector machines. In: Machine Learning and Knowledge Discovery in

Databases, *Lecture Notes in Computer Science*, vol. 6322, pp. 145–161. Springer Berlin / Heidelberg (2010)
24. 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)
25. Lauer, F., Bloch, G.: Incorporating prior knowledge in support vector machines for classification: A review. Neurocomputing **71**(7-9), 1578–1594 (2008)
26. Lauer, F., Bloch, G.: Incorporating prior knowledge in support vector regression. Machine Learning **70**(1), 89–118 (2008)
27. Lee, Y.J., Mangasarian, O., Wolberg, W.: Survival-time classification of breast cancer patients. Computational Optimization and Applications **25**(1-3), 151–166 (2003)
28. Li, G., Jeyakumar, V., Lee, G.: Robust conjugate duality for convex optimization under uncertainty with application to data classification. Nonlinear Analysis: Theory, Methods & Applications **74**(6), 2327–2341 (2011)
29. Li, Y., de Ridder, D., Duin, R., Reinders, M.: Integration of prior knowledge of measurement noise in kernel density classification. Pattern Recognition **41**, 320–330 (2008)
30. Lu, B., Wang, X., Utiyama, M.: Incorporating prior knowledge into learning by dividing training data. Frontiers of Computer Science in China **3**(1), 109–122 (2009)
31. Mangasarian, O.: Knowledge-based linear programming. SIAM J. on Optimization **15**(2), 375–382 (2005)
32. Markou, M., Singh, S.: Novelty detection: a review—part 1: statistical approaches. Signal Processing **83**(12), 2481–2497 (2003)
33. Pavlidis, P., Weston, J., Cai, J., Grundy, W.N.: Gene functional classification from heterogeneous data. In: Proceedings of the fifth annual international conference on Computational biology, pp. 249–255. ACM, New York, NY, USA (2001)
34. Robert, C.: The Bayesian Choice. Springer, New York (1994)
35. 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)
36. Scholkopf, B., Simard, P., Smola, A., Vapnik, V.: Prior knowledge in support vector kernels. In: Advances in neural information processing systems. Proceedings of the 1997 conference., vol. 10, pp. 640–646. MIT Press, Cambridge (1998)
37. Scholkopf, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, Cambridge, Massachusetts (2002)
38. 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)
39. Small, K., Wallace, B., Brodley, C., Trikalinos, T.: The constrained weight space svm: learning with ranked features. In: Proc. of the 28th International Conference on Machine Learning (ICML)., pp. 865–872. Omnipress, Bellevue, WA, USA (2011)
40. Smola, A., Scholkopf, B.: A tutorial on support vector regression. Statistics and Computing **14**, 199–222 (2004)
41. Steinwart, I., Hush, D., Scovel, C.: A classification framework for anomaly detection. Journal of Machine Learning Research **6**, 211–232 (2005)
42. Sun, Q., Wang, L.L., Lim, S., DeJong, G.: Robustness through prior knowledge: using explanation-based learning to distinguish handwritten Chinese characters. International Journal of Document Analysis and Recognition **10**(3-4), 175–186 (2007). DOI 10.1007/s10032-007-0053-1. URL http://dx.doi.org/10.1007/s10032-007-0053-1
43. Sun, Z., Zhang, Z.K., Wang, H.G.: Incorporating prior knowledge into kernel based regression. Acta Automatica Sinica **34**(12), 1515 – 1521 (2008)
44. Tai, F., Pan, W.: Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. Bioinformatics **23**(14), 1775–1782 (2007)
45. Tax, D., Duin, R.: Support vector domain description. Pattern Recognition Letters **20**, 1191–1199 (1999)
46. Tax, D., Duin, R.: Support vector data description. Machine Learning **54**, 45–66 (2004)
47. Troffaes, M.: Decision making under uncertainty using imprecise probabilities. International Journal of Approximate Reasoning **45**(1), 17–29 (2007)

48. Utkin, L.: Imprecise calculation with the qualitative information about probability distributions. In: P. Grzegorzewski, O. Hryniewicz, M. Gil (eds.) Soft Methods in Probability, Statistics and Data Analysis, pp. 164–169. Phisica-Verlag, Heidelberg, New York (2002)
49. Utkin, L.: Imprecise second-order hierarchical uncertainty model. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **11**(3), 301–317 (2003)
50. Utkin, L.: Second-order uncertainty calculations by using the imprecise Dirichlet model. Intelligent Data Analysis **11**(3), 225 – 244 (2007)
51. Utkin, L., Augustin, T.: Decision making under incomplete data using the imprecise Dirichlet model. International Journal of Approximate Reasoning **44**(3), 322–338 (2007)
52. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
53. Veillard, A., Racoceanu, D., Bressan, S.: Incorporating prior-knowledge in support vector machines by kernel adaptation. In: Proceedings of the IEEE 23rd International Conference on Tools with Artificial Intelligence, pp. 591–596. IEEE Computer Society, Washington, DC, USA (2011)
54. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London (1991)
55. Wang, J., Lu, H., Plataniotis, K., Lu, J.: Gaussian kernel optimization for pattern classification. Pattern Recognition **42**(7), 1237 – 1247 (2009)
56. Wang, L., Xue, P., Chan, K.L.: Incorporating prior knowledge into SVM for image retrieval. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04), vol. 2, pp. 981–984. IEEE Computer Society, Los Alamitos, CA, USA (2004)
57. Wu, X., Kumar, V., Ross, Q., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B., Yu, P., Zhou, Z.H., Steinbach, M., Hand, D., Steinberg, D.: Top 10 algorithms in data mining. Knowledge and Information Systems **14**(1), 1–37 (2008)
58. Wu, X., Srihari, R.: Incorporating prior knowledge with weighted margin support vector machines. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 326–333. ACM, New York, NY, USA (2004)
59. Xing, Z., Pei, J., Yu., P., Wang, K.: Extracting interpretable features for early classification on time series. In: Proceedings of the Eleventh SIAM International Conference on Data Mining, pp. 247–258. Omnipress (2011)
60. 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)
61. Zadrozny, B., Langford, J., Abe, N.: Cost-sensitive learning by cost proportionale example weighting. In: Proceedings of the Third IEEE International Conference on Data Mining, pp. 435–442. Melbourne, FL (2003)
62. Zhao, Z., Zhong, P., Zhao, Y.: Learning svm with weighted maximum margin criterion for classification of imbalanced data. Mathematical and Computer Modelling **54**(3-4), 1093 – 1099 (2011)
63. Xu, H., Caramanis, C., Mannor, S.: Robustness and regularization of support vector machines. The Journal of Machine Learning Research **10**, 1485–1510 (2009)