International Journal on Artificial Intelligence Tools


# ROBUST CLASSIFIERS USING IMPRECISE PROBABILITY MODELS AND IMPORTANCE OF CLASSES

LEV V. UTKIN

*Department of Control, Automation and System Analysis*
*Saint-Peterburg State Forest Technical University*
*Saint-Petersburg, 194021, Russia*
*lev.utkin@gmail.com*

YULIA A. ZHUK

*Department of Information Technology and Systems*
*Saint-Peterburg State Forest Technical University*
*Saint-Petersburg, 194021, Russia*
*zhuk_yua@mail.ru*

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

A framework for constructing robust classification models is proposed in the paper. An assumption about importance of one of the classes in comparison with other classes is incorporated into the models. It often takes place in the real application, for example, in reliability, in medical diagnostic, etc. A main idea underlying the models is to consider a set of probability distributions on training examples produced by the imprecise probability models such as linear-vacuous mixture and constant odd-ratio contaminated models. Extreme points of the sets of probability distributions are a main tool for constructing the robust classifiers. It is shown that algorithms for computing optimal classification parameters are reduced to a finite number of weighted support vector machines with weights determined by the extreme points. Experimental results with synthetic and real data illustrate the proposed models.

*Keywords*: Machine learning; classification; support vector machine; imprecise probability model; minimax strategy; extreme points; quadratic programming.

## 1. Introduction

The binary classification problem can be formally written as follows. Given $n$ training data (examples, instances, patterns) $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\}$, in which $\mathbf{x}_i \in \mathbb{R}^m$ represents a feature vector involving $m$ features and $y_i \in \{-1, 1\}$ represents the class of the associated examples, the task of classification is to construct an accurate classifier $c : \mathbb{R}^m \to \{-1, 1\}$ that maximizes the probability that $c(\mathbf{x}_i) = y_i$ for $i = 1, ..., n$. Generally $\mathbf{x}_i$ may belong to an arbitrary set $\mathcal{X}$, but we consider the special case for simplicity $\mathcal{X} = \mathbb{R}^m$. A classification problem is usually characterized by an unknown probability distribution $p(\mathbf{x}, y)$ on $\mathbb{R}^m \times \{-1, +1\}$

defined by the training set or examples $\mathbf{x}_i$ and their corresponding class labels $y_i$. We will call classes with labels $y = -1$ and 1 as negative and positive.

The main problem is to find a decision function $g(\mathbf{x})$ which accurately predicts the class label $y$ of any example $\mathbf{x}$ that may or may not belong to the training set. In other words, we seek a function $g$ that minimizes the classification error, which is given by the probability that $g(\mathbf{x}) \neq y$. One of the possible approaches for solving the problem is the discriminant function approach which uses a real-valued function $f(\mathbf{x})$, called the discriminant function, whose sign determines the class label prediction: $g(\mathbf{x}) = \text{sgn}(f(\mathbf{x}))$. The discriminant function $f(\mathbf{x})$ may be parametrized with some parameters $w_0$, $\mathbf{w} = (w_1, ..., w_m)$, that are determined from the training examples by means of a learning algorithm. In particular, the function $f(\mathbf{x})$ may be linear, i.e. $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + w_0$. We also introduce the notation $x_i^{(k)}$ for the $k$-th element of the vector $\mathbf{x}_i$.

One of the commonly encountered problems in constructing an optimal classifier is the violation of one of the important assumptions accepted in classification that the training data and the testing data are sampled from the same or very similar sources. In real applications, the data might be taken from different processes. Moreover, another assumption underlying many classification models is that there exist a lot of training examples. This assumption is also violated in many applications. Simple examples are data sets in bioinformatics, for example, DNA microarrays. In order to take into account the above violated assumptions and to partially overcome the above difficulties, the robust classification models can be used.

Robust models have been exploited in classification problems due to the opportunity to avoid some strong assumptions underlying the standard classification models. There are several different definitions of robustness in literature. An exhaustive and interesting review of robust models in machine learning was proposed Xu et al.[30]. As pointed out in this review, 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, works[4,5,9,12,13,14,19,31]) 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 $\delta_i$ centered at the "nominal" data point $\mathbf{x}_i$. This model has a very clear intuitive geometric interpretation[1]. 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 support vector machine (SVM) with the hinge loss function provides the following optimization problem with variables $\mathbf{w}$, $\xi = (\xi_1, ..., \xi_n)$, $w_0$,

$\triangle \mathbf{x}_i$ (see, for instance, a similar problem for the binary classification[3])

$$\min_{\mathbf{w},\xi,w_0,\triangle \mathbf{x}_i} \left( \|\mathbf{w}\|^2 /2 + C \sum_{i=1}^{n} \xi_i \right),$$

subject to

$$y_i \left( \langle \mathbf{w}, \phi(\mathbf{x}_i + \triangle \mathbf{x}_i) \rangle + w_0 \right) \geq 1 - \xi_i, \ \xi_i \geq 0,$$

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

Here $\phi$ is a feature map such that the data points are mapped into an alternative higher-dimensional feature space [17,27,29]; $C$ is a constant "cost" parameter.

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, [12]). This class of robust models for one-class classification problems has been studied by Utkin[21] and by Utkin and Zhuk[25]. According to these models, the probability distributions of examples or their weights are assumed to be different and may vary within some predefined set of probability distributions $\mathcal{P}$.

One of the possible sets of distributions $\mathcal{P}$ is produced by the $\varepsilon$-contaminated (robust) models[11]. 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 $\varepsilon$-contaminated model or linear-vacuous mixture [28] is a class $\mathcal{P}(\varepsilon, p)$ of probabilities such that for fixed $\varepsilon \in (0, 1)$ and $p_i$ there holds $\mathcal{P}(\varepsilon, p) = \{(1-\varepsilon)p_i + \varepsilon q_i\}$, where $q_i$ is arbitrary and $q_1 + ... + q_n = 1$. In other words, we take an arbitrary probability distribution $q = (q_1, ..., q_n)$ from the unit simplex. The rate $\varepsilon$ reflects how "close" we feel that $\pi$ must be to $p$ [2]. According to these models, for $0 < \varepsilon < 1$, $\mathcal{P}(\varepsilon, p)$ 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 the possible types of $\varepsilon$-contaminated models.

By accepting the $\varepsilon$-contaminated model or other imprecise statistical models producing $\mathcal{P}$, we assume that the probability of every example in the training set can move around within some subset which is defined by the lower and upper bounds defined above. In other words, while some robust models[1] assume that each point can move around within an Euclidean ball, the second type of robust models assumes that the probability of each point (but not a data point itself) can move around within some subset.

Let us rewrite the problem of minimizing the expected risk[27] in a general form taking into account that every example of the training set has some probability $h_i$ which is derived from the probability distribution $h = (h_1, ..., h_n)$ such that $h \in \mathcal{P}$

$$R(h, \mathbf{w}) = \sum_{i=1}^{n} l(\mathbf{w}, \phi(\mathbf{x}_i)) \cdot h_i. \tag{1}$$

Here $l$ is the loss function depending on parameters $\mathbf{w}$ and training data $\mathbf{x}_i$.

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

$$R_{\text{emp}}(\mathbf{w}) = n^{-1} \sum_{i=1}^{n} l(\mathbf{w}, \phi(\mathbf{x}_i)). \tag{2}$$

One of the possible ways for dealing with the set $\mathcal{P}$ 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 $\mathcal{P}$ such that the expected risk $R(h, \mathbf{w})$ achieves its maximum over $h$ 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 $\mathcal{P}$ 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[16]. This criterion of decision making can be regarded as the well-known $\Gamma$-minimax[2,10,20].

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

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

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

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

Of course, other strategies can be proposed, for instance, the minimin strategy. It can be regarded as a direct opposite to the minimax strategy. According to the minimin strategy, the expected risk is minimized over all probability distributions from the set $\mathcal{P}$ as well as over all values of parameters $\mathbf{w}$. The strategy can be called optimistic because it selects the "best" probability distribution from the set $\mathcal{P}$. Similarly to the minimax strategy, we can write $R(\mathbf{w}_{\text{opt}}) = \min_{\mathbf{w}} \min_{h \in \mathcal{P}} R(h, \mathbf{w})$.

Another strategy is the so-called cautious decision making which is a linear combination of the minimax and minimin strategies. It can be regarded as some intermediate strategy with a caution parameter. A method for cautious decision making under some assumptions concerning imprecise information about states of nature was proposed by Utkin and Augustin[22].

The use of the considered strategies, especially, the minimax strategy states a very important question. What is the sense of the proposed models? This question concerns also the robust models with the additive noise[30]. Suppose that we have observations of the reliability system behavior, which can be divided into two classes: working and failed. What do we get by constructing the separating function $f$ in accordance with one of the robust approaches? We get the robust function $f$.

Unfortunately, we do not have a clear answer what the robust function $f$ means from the reliability point of view. Does it increase the failed region or the working region? This question is incorrect if we use one of the above robust models. It is important to use robust classification models. But it is more important to take into account the meaning of the considered applied problem for constructing the robust classifiers. It is also important to have a clear understanding of the constructed robust model.

Considering the above, we propose another robust minimax or pessimistic strategy which is also interpreted as an insurance against the worst case. However, the worst case here is defined by the applied problem. Let us return to the example with the system reliability analysis. The pessimistic strategy for this example is to assume the failed region as large as possible. Let us consider another example. We have a data set which is composed of biomedical measurements from healthy people and people with some disease. The pessimistic strategy is to prefer to attribute a person just in case to people with the disease. At least this person might take additional medical investigation. We again see from the example that the "disease" region should be as large as possible.

It should be also noted that the meaning of the robust one-class classification model is usually clear in contrast to the binary classification problems. The loss function takes non-zero positive values only at points which correspond to abnormal observations. The probabilities of abnormal observations increase in this case. Therefore, the minimax strategy increases the region of abnormal observations that is understandable in many applications. The minimin strategy as an optimistic one reduces this region because it minimizes probabilities of abnormal observations.

The main idea underlying the proposed robust models is to consider two expected loss measures. In the framework of the minimax strategy and the first standard expected loss measure $R(h, \mathbf{w})$ is minimized over parameters $\mathbf{w}$ under condition that the second loss measure, say $Q(h, \mathbf{w})$, is maximized. The measure $Q$ may be a part of the measure $R$ such that it deals with some subset of observations or training examples. The proposed approach is close to the well-known trade-off between measures of sensitivity and specificity. We indirectly control the sensitivity-specificity ratio by increasing the sensitivity measure in order to reduce the risk of incorrect identification, for example, of sick people. However, in contrast to models taking into account these statistical measures, we combine the robustness and the trade-off between measures of sensitivity and specificity to provide a meaning of the robustness in binary classification under condition of the small amount of training data.

Only two imprecise probability models are considered in the paper: the linear-vacuous mixture, the constant odds-ratio model[28]. These models are selected due to the following reasons. First, these models are well known and have a simple explanation. Second, these models cover the most important and typical cases for constructing the robust classifiers.

The paper is organized as follows. Section 2 presents the formal problem state-

ment and a draft algorithm of the robust classification. A way for getting a detailed algorithm of the robust classification with using extreme points of sets of probability distributions are given in Section 3. Incorporating the linear-vacuous mixture and the constant odds-ratio models into the robust classifiers is considered in Section 4. Sections 5 and 6 provide the modified SVM algorithms taking into account the above imprecise probability models. Numerical experiments with synthetic and real data illustrating performance of the proposed models and their comparison with the standard SVM model are given in Section 7.

## 2. The formal problem statement

Without loss of generality, we assume that training examples with numbers $J_0 = (1, ..., n_0)$ belong to the negative class with $y = -1$ and examples with numbers $J_1 = (n_0 + 1, ..., n)$ belong to the positive class with $y = 1$. We suppose that the expected risk measure has the form of (1). Let us introduce the "partial" expected loss

$$Q(h, \mathbf{w}) = \sum_{i=1}^{n_0} l(\mathbf{w}, \mathbf{x}_i) \cdot h_i.$$

We omit the notation $\phi$ in some places for short.

The expected loss $Q$ characterizes the "negative" region and it is a part of the expected loss measure $R(h, \mathbf{w})$. Its maximization over probabilities $h$ leads to increasing this region and to reducing the "positive" region. If we return to the applied examples with the reliability behavior, then increase of $Q$ means the pessimistic strategy. We prefer in this case to increase this measure as an insurance against the worst case which is a system failure. The same concerns the applied example with the disease. We pessimistically prefer to suppose that a person has the disease in order to avoid a fatal mistake.

The proposed robust classification problem can be formulated as follows:

$$\mathbf{w}_{\rm opt} = \arg\min_{\mathbf{w}} R(h^*, \mathbf{w})$$

such that

$$h^* = \arg\max_{h \in \mathcal{P}} Q(h, \mathbf{w}_{\rm opt}).$$

In other words, we have to find the optimal probability distribution maximizing the function $Q$ and then to find the optimal parameters $\mathbf{w}$ minimizing the function $R$ by the optimal probability distribution $h^*$. We meet two obstacles by solving the above classification problem. First, the optimal function $h^*$ depends on $\mathbf{w}$, i.e., we have the optimal function $h^*$ for every $\mathbf{w}$. Second, the second optimization problem $\max_{h \in \mathcal{P}} Q(h, \mathbf{w}_{\rm opt})$ defines only a part of elements of $h$, namely, $h^0 = (h_1, ..., h_{n_0})$. Elements $h^1 = (h_{n_0+1}, ..., h_n)$ remain to be arbitrary and restricted only by the set $\mathcal{P}$. This peculiarity leads to the ambiguity of the second problem $\min_{\mathbf{w}} R(h^*, \mathbf{w})$ because we have infinitely many optimal distributions $h^*$.

In order to correctly state the classification problem, we introduce the second expected loss function $S(h^1, \mathbf{w})$ such that

$$S(h^1, \mathbf{w}) = \sum_{i=n_0+1}^{n} l(\mathbf{w}, \mathbf{x}_i) \cdot h_i.$$

This function depends only on the second part $h^1$ of the probability distribution $h$. So, we can write $R(h, \mathbf{w}) = Q(h^0, \mathbf{w}) + S(h^1, \mathbf{w})$.

If we maximize $Q(h^0, \mathbf{w})$ over $h^0$, then it is quite reasonable to minimize $S(h^1, \mathbf{w})$ over $h^1$ under condition that $h = (h^0, h^1) \in \mathcal{P}$. It should be noted that the optimistic strategy is used for dealing with the positive class. The main reason for choosing it is that the pessimistic or minimax strategy is robust, but it may not provide an optimal solution in the sense of the smallest value of risk measure $R$. We do not need to provide the robustness for the positive class. Therefore, in order to compensate the nonoptimality of the robust strategy used for the negative class, we apply the optimistic strategy to the positive class.

A rough algorithm can be formulated as follows.

(1) For every $\mathbf{w}$, we find the optimal vector $h^0_{\mathbf{w}}$ maximizing $Q(h^0, \mathbf{w})$. Here the index $\mathbf{w}$ indicates that the optimal vector $h^0$ depends on parameters $\mathbf{w}$.
(2) For every $\mathbf{w}$ and a fixed vector $h^0_{\mathbf{w}}$, we find the optimal vector $h^1_{\mathbf{w}}$ minimizing $S(h^1, \mathbf{w})$ such that $(h^0_{\mathbf{w}}, h^1_{\mathbf{w}}) \in \mathcal{P}$.
(3) By having optimal probability distribution $(h^0_{\mathbf{w}}, h^1_{\mathbf{w}}) \in \mathcal{P}$, we find the optimal parameters $\mathbf{w}_{\text{opt}}$ minimizing $R((h^0_{\mathbf{w}}, h^1_{\mathbf{w}}), \mathbf{w})$.

## 3. An algorithm for solving the classification problem

In fact, the above sequence of steps can not be regarded as an algorithm because we have infinitely many vectors $\mathbf{w}$ and every optimal probability distribution depends on this vector. This sequence can be regarded as the formal classification problem statement. Therefore, we propose the following approach for solving the classification problem below.

Let us introduce two loss functions $l^0$ and $l^1$ such that $l^0(\mathbf{w}, \mathbf{x}_i) = l(\mathbf{w}, \mathbf{x}_i)$ and $l^1(\mathbf{w}, \mathbf{x}_i) = 0$ if $i \in J_0$, and $l^0(\mathbf{w}, \mathbf{x}_i) = 0$ and $l^1(\mathbf{w}, \mathbf{x}_i) = l(\mathbf{w}, \mathbf{x}_i)$ if $i \notin J_0$. Then there holds

$$\max_{h^0} \sum_{i=1}^{n_0} l(\mathbf{w}, \mathbf{x}_i) \cdot h^0_i = \max_{h \in \mathcal{P}} \sum_{i=1}^{n} l^0(\mathbf{w}, \mathbf{x}_i) \cdot h_i.$$

Note that the above optimization problem is linear. This implies that the optimal solution of the problem can be found among the set of extreme points of the set $\mathcal{P}$ denoted as $\mathcal{E}(\mathcal{P})$. So, the above problem can be rewritten as follows:

$$\max_{h \in \mathcal{E}(\mathcal{P})} \sum_{i=1}^{n} l^0(\mathbf{w}, \mathbf{x}_i) \cdot h_i.$$

The same can be written for the loss function $l^1$, i.e., there holds

$$\min_{h^1} \sum_{i=n_0+1}^{n} l(\mathbf{w}, \mathbf{x}_i) \cdot h_i^1 = \min_{(h^0,h^1)\in\mathcal{E}(\mathcal{P})} \sum_{i=1}^{n} l^1(\mathbf{w}, \mathbf{x}_i) \cdot h_i.$$

It is important in the second optimization problem that the objective function is maximized not over all probability distributions $h \in \mathcal{E}(\mathcal{P})$, but over distributions with a fixed optimal part $h^0$. This implies that the second optimization problem is solved by using only part of extreme points having the fixed part $h^0$. It should be noted that for many well-known imprecise statistical models producing the set $\mathcal{P}$, there exists a single extreme point from $\mathcal{E}(\mathcal{P})$ satisfying the above condition. In this case, the whole classification problem is significantly simplified.

Finally, we can write the classification problem as

$$\left( \max_{h\in\mathcal{E}(\mathcal{P})} \sum_{i=1}^{n} l^0(\mathbf{w}, \mathbf{x}_i) \cdot h_i + \min_{(h^0,h^1)\in\mathcal{E}(\mathcal{P})} \sum_{i=1}^{n} l^1(\mathbf{w}, \mathbf{x}_i) \cdot h_i \right) \to \min_{\mathbf{w}}$$

Denote the set of unrepeated vectors $h^0$ from $\mathcal{E}(\mathcal{P})$ as $\mathcal{E}_0$ and the subset of vectors $h^1$ from $\mathcal{E}(\mathcal{P})$ such that the distribution $(h^0, h^1)$ has the fixed vector $h^0$ as $\mathcal{E}_1(h^0)$. Let us introduce the optimization variable $G$ such that

$$G = \max_{h^0\in\mathcal{E}_0} \sum_{i=1}^{n_0} l(\mathbf{w}, \mathbf{x}_i) \cdot h_i^0.$$

Then we can rewrite the optimization problem as follows:

$$\left( G + \sum_{i=n_0+1}^{n} l(\mathbf{w}, \mathbf{x}_i) \cdot h_i^1 \right) \to \min_{\mathbf{w},\ h^1\in\mathcal{E}_1(h^0)}$$

subject to

$$G \geq \sum_{i=1}^{n_0} l(\mathbf{w}, \mathbf{x}_i) \cdot h_i^0,\ h^0 \in \mathcal{E}_0.$$

The main difficulty here is that the vector $h^1$ depends on the vector $h^0$. Moreover, it is defined by the optimal vector $h^0$ maximizing $Q(h^0, \mathbf{w})$. Therefore, in order to solve the classification problem in a general case, the simplest way is to enumerate all different vectors $h^1$ from $\mathcal{E}(\mathcal{P})$ and to solve the following optimization problem for every $h^1$ from $\mathcal{E}(\mathcal{P})$:

$$R(h^1, \mathbf{w}_{\rm opt}(h^1)) = \left( G + \sum_{i=n_0+1}^{n} l(\mathbf{w}, \mathbf{x}_i) \cdot h_i^1 \right) \to \min_{\mathbf{w}} \tag{3}$$

subject to

$$G \geq \sum_{i=1}^{n_0} l(\mathbf{w}, \mathbf{x}_i) \cdot h_i^0,\ h^0 \in \mathcal{E}_0. \tag{4}$$

For every problem, we at first find the optimal vector $h^0$ by looking for an equality among inequalities (4) in constraints. Then we check whether the vector $h^1$ used for constructing the linear programming problem jointly with the optimal vector $h^0$ make up an extreme point from the set $\mathcal{E}(\mathcal{P})$. If it satisfies this condition, then the obtained solution (the vector $\mathbf{w}$) is a candidate for the optimal solution of the whole problem. Among all candidates for the optimal solution, we select one that provides the smallest value of the expected loss measure $R(\mathbf{w}_{\rm opt})$.

The proposed algorithm requires a large amount of the computation time when the number of extreme points is large. Moreover, another computation problem we have is to find all extreme points of $\mathcal{P}$. Nevertheless, the algorithm can be significantly simplified if to apply the well known imprecise statistical models for constructing the set $\mathcal{P}$.

## 4. Imprecise statistical models and the classification problem

### 4.1. *The linear-vacuous mixture*

First of all, we consider the popular robust model called the linear-vacuous mixture or the imprecise $\varepsilon$-contaminated model[28]. We have mentioned in the previous sections that it produces the set of probabilities $\mathcal{P}(\varepsilon, p)$. It is simple to show that the set of extreme points of $\mathcal{P}(\varepsilon, p)$ for $p = (n^{-1}, \ldots, n^{-1})$ consists of $n$ vectors such that the $k$-th vector has the $k$-th element $(1-\varepsilon)n^{-1} + \varepsilon$ and the other $n-1$ elements are equal to $(1-\varepsilon)n^{-1}$.

First, we consider possible vectors $h^0$. They can be of two types. Vectors of the first type contain an element $(1-\varepsilon)n^{-1}+\varepsilon$, and vectors of the second type contain only elements $(1-\varepsilon)n^{-1}$. Let us prove that the largest value of $Q(h^0, \mathbf{w})$ is achieved when $h^0$ is a vector of the first type. Indeed, there holds for the first type

$$Q_1(h^0, \mathbf{w}) = \sum_{i=1,\ i\neq k}^{n_0} l(\mathbf{w}, \mathbf{x}_i) \cdot \frac{1-\varepsilon}{n} + l(\mathbf{w}, \mathbf{x}_k)\left(\frac{1-\varepsilon}{n} + \varepsilon\right).$$

The same can be written for the second type

$$Q_2(h^0, \mathbf{w}) = \sum_{i=1}^{n_0} l(\mathbf{w}, \mathbf{x}_i) \cdot \frac{1-\varepsilon}{n}.$$

It is obvious that

$$Q_1(h^0, \mathbf{w}) - Q_2(h^0, \mathbf{w}) = l(\mathbf{w}, \mathbf{x}_k) \cdot \varepsilon \geq 0,$$

as was to be proved.

Hence, we can write that $h_i^1 = (1-\varepsilon)n^{-1}$ for all $i = n_0+1, ..., n$. It is important to note that the vector $h^1$ does not depend on $h^0$. As a result, we get the following linear programming problem:

$$\left(G + \frac{1-\varepsilon}{n} \sum_{i=n_0+1}^{n} l(\mathbf{w}, \mathbf{x}_i)\right) \to \min_{\mathbf{w}} \tag{5}$$

subject to

$$G \geq \frac{1-\varepsilon}{n}\sum_{i=1}^{n_0} l(\mathbf{w},\mathbf{x}_i) + \varepsilon \cdot l(\mathbf{w},\mathbf{x}_k),\ k = 1,...,n_0. \tag{6}$$

We get a simple optimization problem whose solution does not meet any difficulties when the loss function $l(\mathbf{w},\mathbf{x})$ is linear.

It should be also noted that we do not need to consider the strategy taking into account the minimization of $S(h^1,\mathbf{w})$ over set of vectors $h^1$. We deal only with maximization of the function $Q(h^0,\mathbf{w})$ over the set of $h^0$ or over the set $\mathcal{P}$.

Let us consider two special cases when $\varepsilon = 0$ (the precise case of probabilities of examples) and $\varepsilon = 1$ (the total ignorance about probabilities of examples). If $\varepsilon = 0$, then the above optimization problem is reduced to the problem

$$\frac{1}{n}\sum_{i=1}^{n} l(\mathbf{w},\mathbf{x}_i) \to \min_{\mathbf{w}}.$$

One can see from the above that the objective function is the standard empirical expected loss function. If $\varepsilon = 1$, then there holds

$$\max_{k=1,...,n_0} l(\mathbf{w},\mathbf{x}_k) \to \min_{\mathbf{w}}.$$

This implies that only examples from the class $y = -1$ are used for constructing the classifier. Moreover, the decision is not unique because if we take an arbitrary separating function such that it is negative for all examples from the class $y = -1$, then all values of the loss function are 0. Let us imagine that there exist many examples which are linearly separable except for a single example from the negative class. Then the separating line in case of the linear classification passes across the point corresponding to this "abnormal" example.

It can be concluded from the consideration of two "extreme" cases that we should not take too small and too large values of the parameter $\varepsilon$ for constructing the robust classifier.

### 4.2. *The constant odds-ratio model*

Another imprecise neighborhood model we consider is the so-called constant odds-ratio $(\pi,\varepsilon)$ model[28]. For this model, the set of probability distributions defined as the neighborhood of a given distribution $p$ is

$$\mathcal{P}_C(\varepsilon,p) = \{\pi \in S(1,n) : \frac{\pi_i}{\pi_j} \geq (1-\varepsilon)\frac{p_i}{p_j}\ \forall\, i,j \in \{1,\ldots,n\}\}.$$

Here, $\varepsilon \in [0,1)$. Here, for all $i \in \{1,\ldots,n\}$, the probabilities $\pi_i > 0$ because otherwise $\mathcal{P}_C(\varepsilon,p)$ is not well-defined. Under the constant odds-ratio model, the lower and upper weighting probabilities of the $i$-th point are given by

$$\underline{\pi}_i = \frac{(1-\varepsilon)p_i}{1-\varepsilon p_i} \quad \text{and} \quad \overline{\pi}_i = \frac{p_i}{1-\varepsilon\,(1-p_i)}.$$

According to Walley's work[28], this model is convenient for statistical applications because its form is not changed by conditioning or statistical updating.

When we take the empirical distribution of the training data $p = (n^{-1}, \ldots, n^{-1})$ as the probability distribution of interest, the corresponding set $\mathcal{P}_C(\varepsilon, p)$ has $2n$ extreme points. The first $n$ points are such that the element $(n(1-\varepsilon)+\varepsilon)^{-1}$ is located at the $k$-th position, while the other $n-1$ elements have value $(1-\varepsilon)(n(1-\varepsilon)+\varepsilon)^{-1}$. The other $n$ extreme points have at the $k$-th position the element $(1-\varepsilon)(n-\varepsilon)^{-1}$ and $(n-\varepsilon)^{-1}$ at the remaining positions.

First, we analyze the first type of extreme points. The set of these extreme points can be divided into two subsets. The first subset contains such extreme points that the element $(n(1-\varepsilon)+\varepsilon)^{-1}$ is located at positions with indices from 1 to $n_0$. The second subset consists of extreme points with the element $(n(1-\varepsilon)+\varepsilon)^{-1}$ located at positions with indices from $n_0+1$ to $n$. Since the element $(n(1-\varepsilon)+\varepsilon)^{-1}$ is larger than other elements, then this case is similar to the linear-vacuous mixture model, i.e., there holds

$$\left( G + \frac{1-\varepsilon}{n(1-\varepsilon)+\varepsilon} \sum_{i=n_0+1}^{n} l(\mathbf{w}, \mathbf{x}_i) \right) \to \min_{\mathbf{w}} \tag{7}$$

subject to

$$G \geq \frac{1-\varepsilon}{n(1-\varepsilon)+\varepsilon} \sum_{i=1}^{n_0} l(\mathbf{w}, \mathbf{x}_i) + \varepsilon \cdot l(\mathbf{w}, \mathbf{x}_k),\ k = 1, ..., n_0. \tag{8}$$

We see from (7)-(8) that the obtained problem differs from the similar problem (5)-(6) only by the multiplier before sums of loss functions. However, we have another type of extreme points. We again divide the set of extreme points of the second type into two subsets. The first subset contains such extreme points that the element $(1-\varepsilon)/(n-\varepsilon)$ is located at positions with indices from 1 to $n_0$. The second subset consists of extreme points with the element $(1-\varepsilon)/(n-\varepsilon)$ located at positions with indices from $n_0+1$ to $n$. It is simple to prove that the largest value of $Q(h^0, \mathbf{w})$ is achieved when $h^0$ is a vector of the second type because $(1-\varepsilon)/(n-\varepsilon)$ is the smallest element among elements of the extreme points. It follows from (3)-(4) that we have to solve $n - n_0$ optimization problems

$$R(h^1, \mathbf{w}_{\rm opt}(h^1)) = \left( \frac{1}{n-\varepsilon} \sum_{i=1}^{n_0} l(\mathbf{w}, \mathbf{x}_i) + \sum_{i=n_0+1}^{n} l(\mathbf{w}, \mathbf{x}_i) \cdot h_i^1 \right) \to \min_{\mathbf{w}}$$

and to find the smallest value of $R(h^1, \mathbf{w}_{\rm opt}(h^1))$ by different vectors $h^1$. We remove constraints (4) because they do not depend on extreme points now. By substituting the vectors $h^1$ into the above problem, we get for $k = n_0 + 1, ..., n$,

$$\left( \frac{1}{n-\varepsilon} \sum_{i=1}^{n} l(\mathbf{w}, \mathbf{x}_i) - \varepsilon \cdot l(\mathbf{w}, \mathbf{x}_k) \right) \to \min_{\mathbf{w}}. \tag{9}$$

In sum, we have $n_0$ optimization problems (7)-(8) such that the optimal solution maximizing the objective function over $k = 1, ..., n_0$ is selected, and $n - n_0$ optimization problems (9) such that the optimal solution minimizing the objective function over $k = n_0 + 1, ..., n$ is selected.

## 5. SVM for the linear-vacuous mixture

Now we apply the above ideas of the robust classifiers to the SVM with the hinge loss function under condition $f(\mathbf{x}, \mathbf{w}) = \langle w, \phi(\mathbf{x})\rangle + w_0$. Let us add the standard Tikhonov regularization term $\frac{1}{2}\langle w, w\rangle$ (this is the most popular penalty or smoothness term) [18] to the objective function (5) and the constant "cost" parameter $C$. 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 [6,7,17,29] for a detailed analysis of regularization methods and SVMs). Moreover, we introduce the following optimization variables:

$$H_i = l(\mathbf{w}, \mathbf{x}_i) = \max(0, f(\mathbf{x}_i, \mathbf{w})),\ i = 1, ..., n.$$

This leads to the quadratic programming problem

$$R(\mathbf{w}_{\rm opt}) = \min\left(\frac{1}{2}\langle w, w\rangle + C\cdot G + C\cdot\frac{1-\varepsilon}{n}\sum_{i\in J_1} H_i\right),$$

subject to

$$H_i \geq 1 - y_i f(\mathbf{x}_i, \mathbf{w}),\ H_i \geq 0,\ i = 1, ..., n,$$

$$G \geq \frac{1-\varepsilon}{n}\sum_{i\in J_0} H_i + \varepsilon\cdot H_k,\ k \in J_0.$$

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

$$L = \frac{1}{2}\langle w, w\rangle + C\cdot G + C\cdot\frac{1-\varepsilon}{n}\sum_{i\in J_1} H_i - \sum_{i=1}^{n}\lambda_i H_i$$

$$-\sum_{i=1}^{n}\varphi_i\left(H_i - 1 + y_i\langle w, \phi(\mathbf{x}_i)\rangle + y_i w_0\right)$$

$$-\sum_{k\in J_0}\eta_k\left(G - \frac{1-\varepsilon}{n}\sum_{i\in J_0} H_i - \varepsilon\cdot H_k\right) \to \max.$$

Here $\eta_i$, $\varphi_i$, $\lambda_i$ are positive Lagrange multipliers. The saddle point can be found by setting the derivatives equal to zero

$$\partial l/\partial w_0 = \sum_{i=1}^{n} \varphi_i y_i = 0,$$

$$\partial l/\partial w_j = w_j - \sum_{i=1}^{n} \varphi_i y_i x_i^{(j)} = 0,\ j = 1, ..., m,$$

$$\partial l/\partial H_i = C \cdot \frac{1-\varepsilon}{n} - \lambda_i - \varphi_i = 0,\ i \in J_1,$$

$$\partial l/\partial H_i = -\lambda_i - \varphi_i + \frac{1-\varepsilon}{n} \sum_{k \in J_0} \eta_k + \varepsilon \cdot H_i = 0,\ i \in J_0,$$

$$\partial l/\partial G = C - \sum_{k \in J_0} \eta_k = 0.$$

By substituting the above conditions into the Lagrangian, we get the dual optimization problem:

$$\text{maximize}\ \ L = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \varphi_i \varphi_j K(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{n} \varphi_i,$$

subject to

$$\sum_{i=1}^{n} \varphi_i y_i = 0,\ \sum_{k \in J_0} \eta_k = C,$$

$$0 \le \varphi_i \le C \cdot \frac{1-\varepsilon}{n},\ \ i \in J_1,$$

$$0 \le \varphi_i \le C \cdot \frac{1-\varepsilon}{n} + \varepsilon \cdot \eta_i,\ \ i \in J_0.$$

Here $K(\mathbf{x}, \mathbf{y}) = (\phi(\mathbf{x}), \phi(\mathbf{y}))$ is some kernel satisfying the Mercer's condition[26].

One can see that the objective function does not differ from the objective function obtained in the standard SVM with the empirical probability distribution of the examples, i.e. by exploiting the empirical expected loss measure. However, the main difference is in the constraints. Let us replace the variable $\eta_k$ by $\psi_k = \eta_k/C$. Then we can rewrite the constraints with $\eta_k$ as follows:

$$\sum_{k \in J_0} \psi_k = 1,\ 0 \le \varphi_i \le C \cdot \left( \frac{1-\varepsilon}{n} + \varepsilon \cdot \psi_i \right),\ \ i \in J_0.$$

The variable $\psi_i$ can be interpreted as the conditional probability of the $i$-th example under condition that it belongs to the class $y = -1$. This implies that the term in round brackets is the contaminated conditional probability.

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

$$f(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{n} \varphi_i K(\mathbf{x}_i, \mathbf{x}) + w_0.$$

It is simple to see that the problem in the "precise" case $\varepsilon = 0$ is the standard SVM dual problem.

Let us write some of the Karush-Kuhn-Tucker complementarity conditions

$$\eta_k \left( G - \frac{1-\varepsilon}{n} \sum_{i \in J_0} H_i - \varepsilon \cdot H_k \right) = 0.$$

It follows from the definition of $G$ that $G = (1-\varepsilon) n^{-1} \sum_{i \in J_0} H_i + \varepsilon \cdot H_k$ for a single value of $k \in J_0$, say $k = s$, providing its largest value. It is assumed that values $(1-\varepsilon) n^{-1} \sum_{i \in J_0} H_i + \varepsilon \cdot H_k$ do not coincide for different $k$. It should be noted that it is simple to get the same results when the corresponding values may coincide. This implies that $\eta_k = 0$ for all $k \neq s$. In other words, we have optimal vectors of variables $\eta_k$

$$(\eta_1, 0, ..., 0), \ (0, \eta_2, ..., 0), ..., (0, ..., 0, \eta_{n_0}).$$

It follows from the saddle point that $\eta_s = C$, and we get new optimal vectors of variables $\eta_k$

$$(C, 0, ..., 0), \ (0, C, ..., 0), ..., (0, ..., 0, C).$$

Hence, we can replace the obtained dual optimization problem by $n_0$ problems substituting one of the above optimal vectors of variables $\eta_k$, i.e., we get for $s = 1, ..., n_0$

$$\text{maximize} \ \ L_s = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \varphi_i \varphi_j K(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{n} \varphi_i, \tag{10}$$

subject to

$$\sum_{i=1}^{n} \varphi_i y_i = 0, \tag{11}$$

$$0 \leq \varphi_i \leq C \cdot \frac{1-\varepsilon}{n}, \quad i = 1, ..., n, \ i \neq s, \tag{12}$$

$$0 \leq \varphi_s \leq C \cdot \left( \frac{1-\varepsilon}{n} + \varepsilon \right). \tag{13}$$

Finally, we get $n_0$ problems corresponding to standard weighted SVM with the identical weights $1-\varepsilon$ for all examples and with the weight $1+(n-1)\varepsilon$ assigned to the $s$-th example from the class $y = -1$. It is interesting to note that the increased weight is assigned to an example only from the negative class. We select a solution $\mathbf{w}_{\text{opt}}$ such that $\mathbf{w}_{\text{opt}} = \arg\max_{s \in J_0}(L_s(\mathbf{w}))$.

## 6. SVM for the constant odds-ratio model

By using (7)-(8) and (9) and taking into account the derivation of the SVM for the linear-vacuous mixture model, we can similarly write a set of quadratic programming problems for the constant odds-ratio model. The objective function and one of the constraints coincide with (10) and (11), respectively, but other constraints take into account the weights or extreme points of the constant odds-ratio model.

If $s = 1, ..., n_0$, then we have the following constraints:

$$0 \leq \varphi_i \leq C\cdot \ , \quad i = 1, ..., n, \ i \neq s,$$

$$0 \leq \varphi_s \leq C \cdot \frac{1}{n(1-\varepsilon)+\varepsilon}.$$

If $s = n_0 + 1, ..., n$, the constraints are of the form:

$$0 \leq \varphi_i \leq C \cdot \frac{1}{n-\varepsilon}, \quad i = 1, ..., n, \ i \neq s,$$

$$0 \leq \varphi_s \leq C \cdot \frac{1-\varepsilon}{n-\varepsilon}.$$

So, we get $n_0$ optimization problems corresponding to standard weighted SVM with the identical weights $n(1-\varepsilon)/(n(1-\varepsilon)+\varepsilon)$ for all examples and with the weight $n/(n(1-\varepsilon)+\varepsilon)$ assigned to the $s$-th example from the class $y = -1$. In addition, we get $n - n_0$ optimization problems corresponding to standard weighted SVM with the identical weights $n/(1-\varepsilon)$ for all examples and with the weight $n(1-\varepsilon)/(n-\varepsilon)$ assigned to the $s$-th example from the class $y = 1$. The optimal solution is selected by maximizing the Lagrangian $L_s$ over $s = 1, ..., n_0$ and by minimizing $L_s$ over $s = n_0 + 1, ..., n$. Suppose that we get solutions $\mathbf{w}^*$ and $\mathbf{w}^{**}$ corresponding to the first and the second subsets of extreme points, respectively. Finally, we select a solution $\mathbf{w}_{\rm opt}$ such that

$$\mathbf{w}_{\rm opt} = \arg\max(L_s(\mathbf{w}^*), L_s(\mathbf{w}^{**})).$$

## 7. Numerical experiments

We illustrate the model proposed in this paper via several examples, all computations have been performed using the statistical software R [15]. We investigate the performance of the proposed model and compare it with the standard SVM by considering two error measures ($E$ and $E_{10}$), which are the proportion of misclassified examples on a sample of data and the proportion of misclassified examples from the negative class with label $y = -1$. The first measure is often used to quantify the predictive performance of classification models. However, we are interested in the second measure because it is important for us to minimize errors in one of the classes, namely, in the negative class. The measures can formally be written as

$$E = N_T/N, \quad E_{10} = N_{01}/N.$$

Here $N_T$ ($N_{01}$) is the number of all (negative) test examples for which the predicted class for an example does not coincide with its true class, $N$ is the total number of test data. It should be noted that the measure $E_{10}$ is called very often as the false positive rate.

All experiments use a standard Gaussian radial basis function (RBF) kernel with the kernel parameter $\gamma$

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

Different values for the parameter $\gamma$ have been tested, choosing those leading to the best results. Moreover, we use the linear-vacuous mixture with the parameter $\varepsilon$. The number of instances for training will be denoted as $n$. The weighted version of the well known support vector machine (SVM) method is used in all experiments.

First, we consider the performance of the proposed model 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 $n/2$ examples (the first subset) are generated with mean values $\mathbf{m}_1 = (m_1, m_1)$, and $n/2$ examples (the second subset) have mean values $\mathbf{m}_2 = (m_2, m_2)$. Here we take identical mean values of features. The standard deviation is $\sigma$ for both subsets and both features.

We investigate how the accuracy measures depend on the number of examples in the training set. We accept $m_1 = 5$, $m_2 = 8$, $\sigma = 3$, $\varepsilon = 0.2$, $\gamma = 6$. The "cost" parameter is $C = 100$. Three pictures illustrating the proposed model are shown in Figs. 1-3. They correspond to cases of $n = 20, 40, 80$, respectively, and depict contours $f(\mathbf{x}) = 0$ and generated data points with the above parameters. The solid curves correspond to the proposed model, the dashed curves correspond to the standard SVM with identical weights of training examples. Small triangles and circles correspond to training data from the negative and positive classes, respectively. One can see from the pictures that the difference between models is reduced with increase of $n$. Moreover, it can be seen from the first picture that the contour obtained from the proposed model covers all triangles. In other words, we prefer to extend the region of the negative examples in order to be sure that the test examples will be in this region. It should be noted that the synthetic training set is generated by using the precise model with normally distributed examples. However, the imprecision here is a result of a small amount of training data, for instance, $n = 20$. Moreover, a quite small number of training data ($n = 10$) will be studied in numerical experiments with real data. This imprecision can be taken into account by means of the proposed imprecise models.

In order to study how the dispersion of data impacts on the classification performance, we compute the measures $E_{10}$ and $E$ for $\sigma = 3$ and $\sigma = 4$ by the same other parameters. The results are shown in Fig. 4 for different values of the parameter $\varepsilon$. It should be noted that the case $\varepsilon = 0$ corresponds to the standard model and it is given for comparison of the classification performance between two models. The first picture illustrates how the classification error $E_{10}$ is changed by different values of the parameter $\varepsilon$. The measure $E$ is shown in the second picture. The solid curves

Table 1. A brief introduction about data sets

| | $m$ | $n_0$ | $n - n_0$ |
|---|---|---|---|
| PID | 8 | 268 | 500 |
| Musk | 166 | 499 | 269 |
| BCWD | 30 | 212 | 357 |
| HS | 3 | 225 | 81 |
| Parkinsons | 22 | 147 | 48 |
| BT | 9 | 36 | 70 |
| ILP | 10 | 416 | 167 |
| Seeds | 7 | 70 | 140 |
| Ionosphere | 34 | 225 | 126 |
| Ecoli | 7 | 143 | 193 |
| VC2C | 6 | 210 | 100 |
| Yeast | 8 | 463 | 1021 |

correspond to the case $\sigma = 3$, the dashed curves correspond to the case $\sigma = 4$. One can see from the pictures that the proposed model allows us to decrease the classification error by taking non-zero values of $\varepsilon$. Of course, this improvement can be clearly observed only for the measure $E_{10}$. The measure $E$ does not show any visible changes.

The proposed algorithm has been evaluated and investigated by the following publicly available data sets: Pima Indian Diabetes (PID), Musk, Breast Cancer Wisconsin Diagnostic (BCWD), Haberman's Survival (HS), Parkinsons, Breast Tissue (BT), Indian Liver Patient (ILP), Seeds, Ionosphere, Ecoli, Vertebral Column 2C (VC2C), Yeast. All data sets are from the UCI Machine Learning Repository[8]. A brief introduction about these data sets are given in Table 1, while more detailed information can be found from, respectively, the data resources. It should be noted that the first two classes (carcinoma, fibro-adenoma) in the Breast Tissue data set are united and regarded as the negative class. Two classes (disk hernia, spondylolisthesis) in the Vertebral Column 2C data set are united and regarded as the negative class. The class CYT (cytosolic or cytoskeletal) in Yeast data set is regarded as negative. Other classes are united as the positive class. The first class in Seeds data set is regarded as negative.

Figs. 5-10 illustrate how the classification error $E_{10}$ depends on $\varepsilon$ for the above data sets. The solid curve with triangle markers, the dashed curve with circle markers and the thin curve with cross markers correspond to cases $n = 10$, $n = 20$ and $n = 40$, respectively. We take $n/2$ examples from every class for training, which are randomly selected from the classes. The remaining instances in every data set are used for validation. One can see from the pictures that the most data sets demonstrate the apparent improvement of the classification performance in some intervals of the parameter $\varepsilon$. Moreover, we can see that this improvement depends on the

Table 2. The classification performance by different $\varepsilon$ for $n = 10$

| $\varepsilon$ | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 |
|---|---|---|---|---|---|---|---|---|---|
| PID | .250 | .247 | .244 | .243 | .243 | .243 | .240 | .232 | **.224** |
| Musk | .221 | .168 | .150 | .136 | .133 | .130 | .125 | **.123** | .127 |
| BCWD | .055 | .055 | .056 | .056 | .054 | **.053** | .056 | .057 | .061 |
| HS | .206 | .214 | .208 | .207 | .195 | .204 | .198 | **.190** | .191 |
| Parkinsons | .268 | .224 | .209 | .207 | .204 | .207 | .206 | .205 | **.195** |
| BT | .057 | .058 | .059 | .056 | .052 | .053 | .047 | **.044** | .060 |
| ILP | .330 | .310 | .305 | .302 | .304 | .302 | .292 | .285 | **.274** |
| Seeds | .053 | .052 | .052 | .051 | .051 | .051 | .052 | .051 | **.050** |
| Ionosphere | .188 | .169 | .159 | .152 | .147 | .144 | .143 | .141 | **.139** |
| Ecoli | .031 | .030 | .030 | .029 | .029 | .029 | .028 | **.027** | .028 |
| VC2C | .154 | .123 | .117 | .112 | .109 | .108 | .107 | .105 | **.104** |
| Yeast | .102 | .098 | .104 | .104 | .116 | .101 | **.099** | .106 | .136 |

Table 3. The classification performance by different $\varepsilon$ for $n = 20$

| $\varepsilon$ | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 |
|---|---|---|---|---|---|---|---|---|---|
| PID | .221 | .220 | .219 | .217 | .216 | .214 | .207 | .194 | **.176** |
| Musk | .163 | .151 | .150 | .148 | .148 | .150 | .151 | .148 | **.079** |
| BCWD | **.039** | .040 | .041 | .041 | .041 | .042 | .044 | .052 | .068 |
| HS | .174 | .173 | .178 | .171 | .180 | .161 | .160 | **.146** | .147 |
| Parkinsons | .199 | .191 | .190 | **.185** | .186 | .191 | .187 | .194 | .280 |
| BT | .030 | .027 | .030 | .025 | .022 | .020 | .018 | **.016** | .022 |
| ILP | .316 | **.315** | .320 | .322 | .318 | .319 | .332 | .328 | .350 |
| Seeds | .041 | .041 | .040 | .040 | .039 | .040 | .038 | .038 | **.035** |
| Ionosphere | .131 | .130 | .129 | .129 | .129 | .128 | .127 | .123 | **.063** |
| Ecoli | .022 | .022 | .021 | .021 | .021 | .020 | .020 | .019 | **.015** |
| VC2C | .124 | .116 | .115 | .118 | .115 | .114 | **.113** | .116 | .205 |
| Yeast | .102 | .107 | .113 | .111 | .106 | **.097** | .100 | .108 | .176 |

number of training examples. The better results can be observed when the number of training examples is rather small, for instance, when $n = 10$. At the same time, we see that some data sets, for example, Breast Cancer Wisconsin, do not show the improvement of the classification performance. For better comparison and analysis, the results shown in Figs. 5-10 are represented as Tables 2-4 in accordance with the value $n$ of training examples. The smallest values of $E_{10}$ are shown in bold to highlight the top performances. It can be seen from Tables 2-4 that the optimal value of $\varepsilon$ providing the smallest value of $E_{10}$ decreases for many data sets with $n$. The mean values of the optimal values of $\varepsilon$ for $n = 10$, 20 and 40 by taking into

Table 4. The classification performance by different $\varepsilon$ for $n = 40$

| $\varepsilon$ | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 |
|---|---|---|---|---|---|---|---|---|---|
| PID | .184 | .179 | .176 | .169 | .165 | .160 | .157 | **.151** | .153 |
| Musk | .157 | .157 | .158 | .157 | .152 | .142 | .096 | **.087** | .164 |
| BCWD | **.036** | .037 | .039 | .040 | .042 | .046 | .050 | .056 | .068 |
| HS | .152 | .141 | .139 | .136 | .131 | .141 | .126 | **.118** | .135 |
| Parkinsons | **.161** | .163 | .168 | .176 | .192 | .217 | .257 | .295 | .318 |
| BT | .013 | .010 | .012 | .008 | .009 | .005 | .004 | **.003** | .020 |
| ILP | .108 | **.107** | .110 | .109 | .113 | .116 | .121 | .133 | .124 |
| Seeds | .029 | .028 | .028 | .027 | .027 | .026 | .024 | **.023** | .026 |
| Ionosphere | .086 | .085 | .085 | .085 | .083 | .071 | .053 | .033 | **.032** |
| Ecoli | .013 | .013 | .012 | .012 | .011 | .010 | .008 | .006 | **.005** |
| VC2C | .127 | .126 | .126 | .127 | .131 | .137 | **.191** | .311 | .338 |
| Yeast | .070 | .074 | .068 | .077 | .071 | **.059** | .067 | .103 | .190 |

account 12 analyzed data sets are 0.725, 0.717 and 0.542. This corresponds to the assumption that the imprecision of the model is reduced as the amount of training data increases.

The linear-vacuous mixture model is used for the experiments because this model is the most simple and is explainable in a simple way as a contaminated model. For some data sets (PID and Musk), we use also the constant odds-ratio model (see Fig. 11).

## 8. Conclusion

In this paper we have presented a new approach for robust classification under condition that one of the classes is more important in comparison with another class. Two well-known imprecise statistical models have been used for producing sets of probabilities which can be regarded as weights of examples from the training set. However, the approach can be extended on other imprecise statistical models, for example, pari-mutuel models[28], Kolmogorov-Smirnov bounds[23,24], etc. Moreover, if we can determine all extreme points of a set of probability distributions produced by a statistical model, then the approach can be adapted to this set of distributions in a simple way.

It should be noted that the approach differs from the standard method in the framework of minimax strategies where we have a single objective function which has to be maximized over the set of probability distributions and minimized over the set of parameters $\mathbf{w}$. Here we have two different objective functions. This is the main difficulty we had to overcome.

In order to numerically evaluate the proposed model, we have compared it with the standard classifier (SVM) by using some sets of real data, including well-known Pima Indian Diabetes, Haberman's Survival, Yeast, Ecoli and many other data

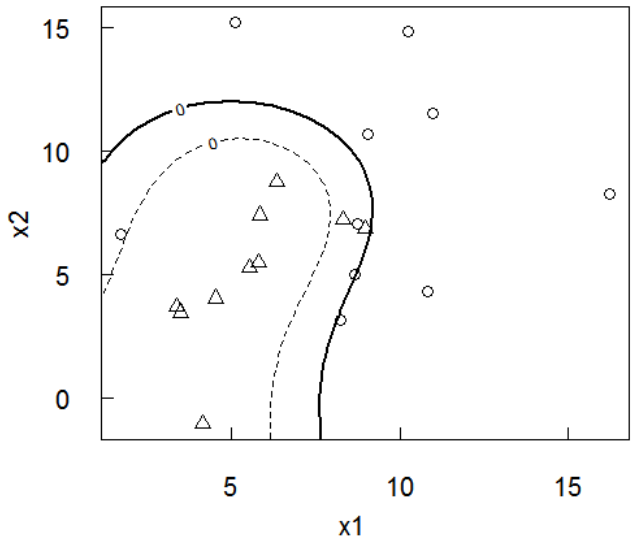


Fig. 1. Contours $f(\mathbf{x}) = 0$ and generated data points for $n = 20$

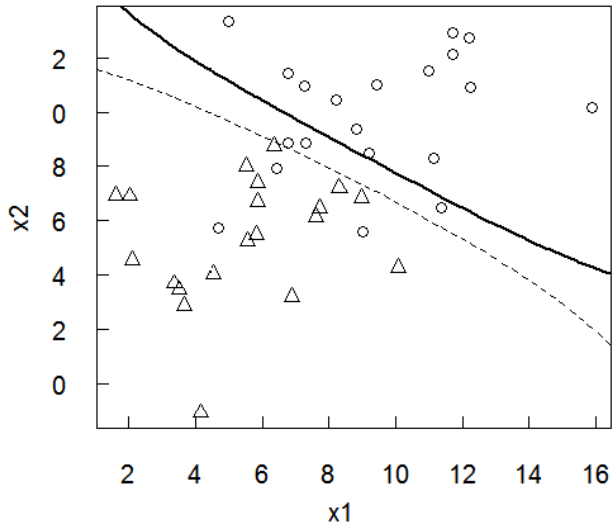


Fig. 2. Contours $f(\mathbf{x}) = 0$ and generated data points for $n = 40$

sets. The standard classifier has been obtained by taking $\varepsilon = 0$. Most numerical experiments have shown that incorporating the imprecise statistical models into

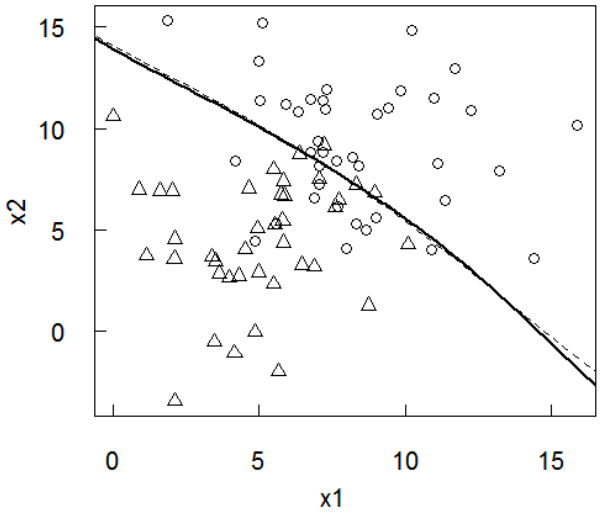


Fig. 3. Contours $f(\mathbf{x}) = 0$ and generated data points for $n = 80$

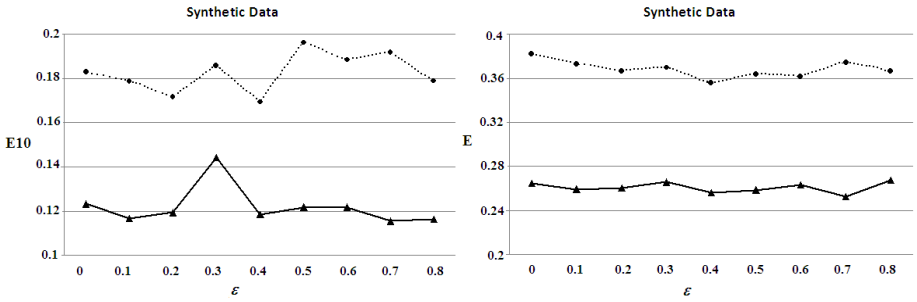


Fig. 4. The classification performances $E_{10}$ and $E$ by different $\varepsilon$ for synthetic data

the classification problems of the special form leads to improving the classification accuracy.

Additional experiments have been made to test the model and to compare it with the standard SVM by using synthetic data and generating examples in accordance with the normal probability distribution.

We believe that this work opens ways for future work which is mainly directed towards three goals. Firstly, we are working on developing a set of models dealing with imbalanced data. Secondly, we are working on models which linearly combine two parts of the expected loss function in order to take into account the joint importance of the classes. Thirdly, we try to extend the proposed model to multiclass

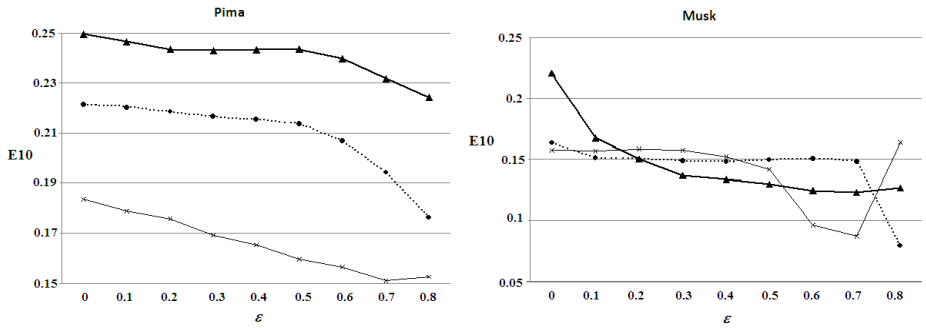


Fig. 5. The classification performance by different $\varepsilon$ and $n$ for PID and Musk data sets

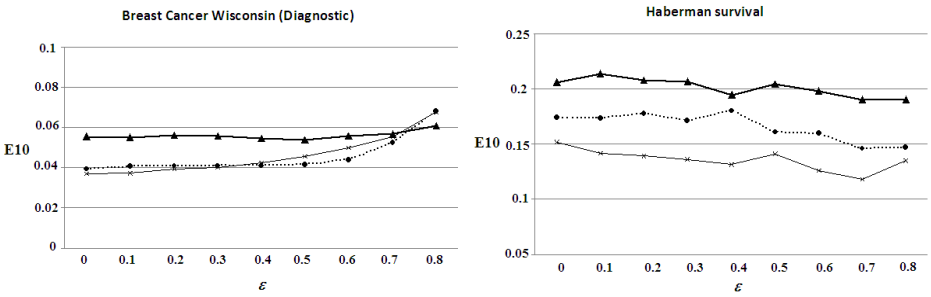


Fig. 6. The classification performance by different $\varepsilon$ and $n$ for BCWD and HS data sets

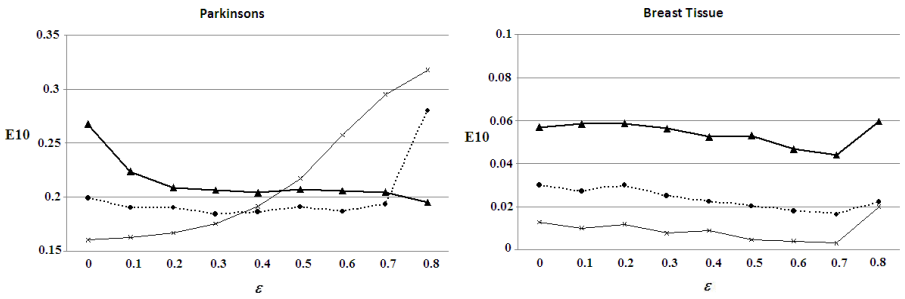


Fig. 7. The classification performance by different $\varepsilon$ and $n$ for Parkinsons and BT data sets

classification.

Finally, it is necessary to note that the choice of the best parameters of the imprecise statistical models is still an open question. Numerical experiments have illustrated that there are some optimal values of the parameters providing the smallest classification error. However, they can be found only by considering all

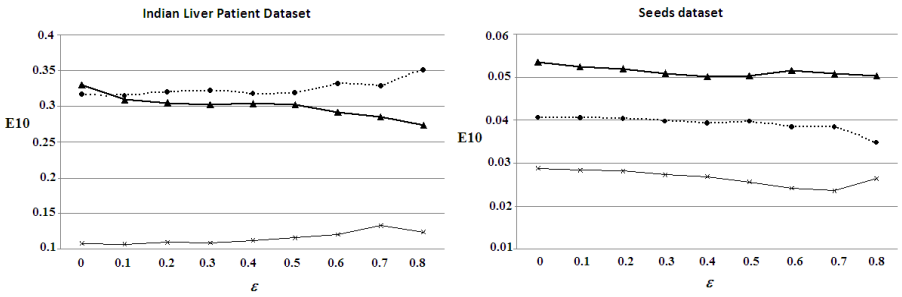


Fig. 8. The classification performance by different $\varepsilon$ and $n$ for ILP and Seeds data sets

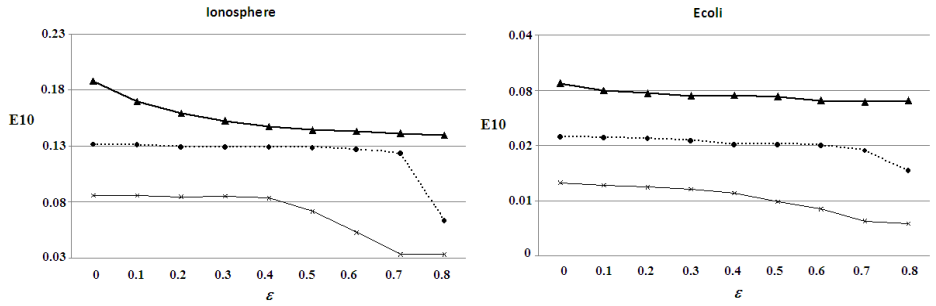


Fig. 9. The classification performance by different $\varepsilon$ and $n$ for Ionosphere and Ecoli data sets

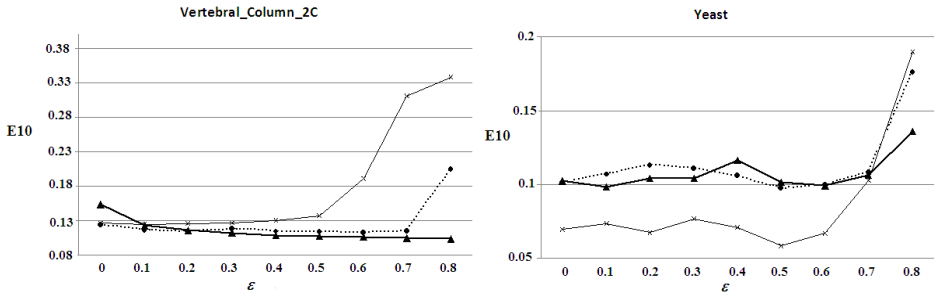


Fig. 10. The classification performance by different $\varepsilon$ and $n$ for VC2C and Yeast data sets

possible values in a predefined grid.

## Acknowledgement

We would like to express our appreciation to the anonymous referees whose valuable comments have improved the paper. The reported study was partially supported

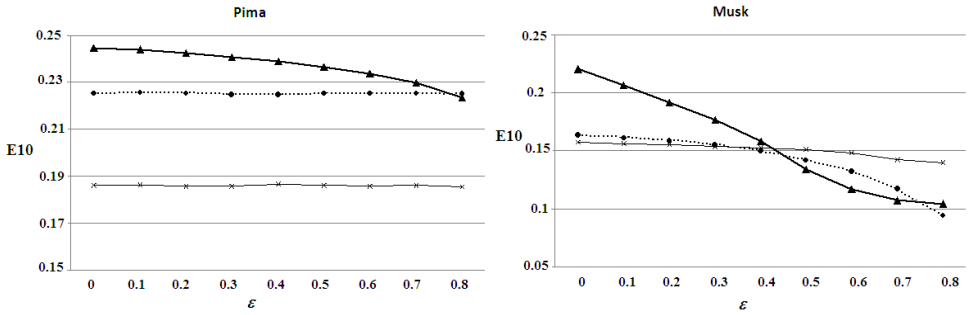


Fig. 11. The classification performance by different $\varepsilon$ and $n$ for PID and Musk data sets by using the constant odds-ratio model

by RFBR, research project No. 14-01-00165-a, and by the Ministry of Education and Science of Russian Federation, research project No. 2014/181-2220.

## References

1. A. Ben-Tal, L.E. Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, Princeton, New Jersey, 2009.
2. J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 1985.
3. J. Bi and T. Zhang. Support vector classification with input data uncertainty. In L.K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17, pages 161–168. MIT Press, Cambridge, MA, 2004.
4. C. Bouveyron and S. Girard. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, 42(11):2649 – 2658, 2009.
5. A. Cerioli, M. Riani, and A.C. Atkinson. Robust classification with categorical variables. In A. Rizzi and M. Vichi, editors, *Compstat 2006 - Proceedings in Computational Statistics*, pages 507–519. Physica-Verlag HD, 2006.
6. V. Cherkassky and F.M. Mulier. *Learning from Data: Concepts, Theory, and Methods*. Wiley-IEEE Press, UK, 2007.
7. T. Evgeniou, T. Poggio, M. Pontil, and A. Verri. Regularization and statistical learning theory for data analysis. *Computational Statistics & Data Analysis*, 38(4):421 – 432, 2002.
8. A. Frank and A. Asuncion. UCI machine learning repository, 2010.
9. L.E. Ghaoui, G.R.G. Lanckriet, and G. Natsoulis. Robust classification with interval data. Technical Report Report No. UCB/CSD-03-1279, University of California, Berkeley, California 94720, 2003.
10. I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.
11. P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.
12. G.R.G. Lanckriet, L.E. Ghaoui, C. Bhattacharyya, and M.I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582, 2002.
13. G.R.G. Lanckriet, L.E. Ghaoui, and M.I. Jordan. Robust novelty detection with

single-class mpm. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, pages 905–912. MIT Press, Cambridge, MA, 2003.

14. F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.
15. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005.
16. C.P. Robert. *The Bayesian Choice*. Springer, New York, 1994.
17. B. Scholkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, Massachusetts, 2002.
18. A.N. Tikhonov and V.Y. Arsenin. *Solution of Ill-Posed Problems*. W.H. Winston, Washington DC, 1977.
19. T.B. Trafalis and R.C. Gilbert. Robust support vector machines for classification and computational issues. *Optimization Methods and Software*, 22(1):187–198, 2007.
20. M.C.M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, 2007.
21. L.V. Utkin. A framework for imprecise robust one-class classification models. *International Journal of Machine Learning and Cybernetics*, 2014. 10.1007/s13042-012-0140-6.
22. L.V. Utkin and Th. Augustin. Efficient algorithms for decision making under partial prior information and general ambiguity attitudes. In T. Seidenfeld F.G. Cozman, R. Nau, editor, *Proc. of the 4th Int. Symposium on Imprecise Probabilities and Their Applications, ISIPTA'05*, pages 349–358, Pittsburgh, USA, July 2005. Carnegie Mellon University, SIPTA.
23. L.V. Utkin and F.P.A. Coolen. On reliability growth models using Kolmogorov-Smirnov bounds. *International Journal of Performability Engineering*, 7(1):5–19, 2011.
24. L.V. Utkin and F.P.A. Coolen. Classification with support vector machines and Kolmogorov-Smirnov bounds. *Journal of Statistical Theory and Practice*, 8(2):297-318, 2014.
25. L.V. Utkin and Y.A. Zhuk. Robust novelty detection in the framework of a contamination neighbourhood. *International Journal of Intelligent Information and Database Systems*, 7(3):205–224, 2013.
26. V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.
27. V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
28. P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.
29. A.R. Webb. *Statistical Pattern Recognition, 2nd Edition*. Wiley, 2002.
30. H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *The Journal of Machine Learning Research*, 10(7):1485–1510, 2009.
31. L. Xu, K. Crammer, and D. Schuurmans. Robust support vector machine training via convex outlier ablation. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06)*, volume 21, pages 536–542, Boston, Massachusetts, 2006. AAAI Press; MIT Press.