Robust novelty detection in the framework of a contamination neighborhood

Lev V. Utkin*

Department of Control, Automation and System Analysis,
St.Petersburg State Forest Technical University,
Institutsky per. 5, 194021 St.Petersburg, Russia
E-mail: lev.utkin@gmail.com
*Corresponding author

Yulia A. Zhuk

Department of Computer Science,
St.Petersburg State Forest Technical University,
Institutsky per. 5, 194021 St.Petersburg, Russia
E-mail: zhuk _ yua@mail.ru

Abstract: A novelty detection robust model is studied in the paper. It is based on contaminated (robust) models which produce a set of probability distributions of data points instead of the empirical distribution. The minimax and minimin strategies are used to construct optimal separating functions. An algorithm for computing the optimal parameters of the novelty detection model is reduced to a finite number of standard SVM tasks with weighted data points. Experimental results with synthetic and some real data illustrate the proposed novelty detection robust model.

Keywords: machine learning; novelty detection; classification; minimax strategy; support vector machine; quadratic programming.

Reference to this paper should be made as follows: Utkin, L.V. and Zhuk, Y.A. (xxxx) 'Robust novelty detection in the framework of a contamination neighborhood', *Int. J. of Intelligent Information and Database Systems*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Lev V. Utkin is currently the Vice-rector for Research and a Professor at the Department of Control, Automation and System Analysis, Saint-Petersburg State Forest Technical University. He holds a PhD in Information Processing and Control Systems (1989) from Saint-Petersburg Electrotechnical University and a DSc in Mathematical Modelling (2001) from Saint-Petersburg State Institute of Technology, Russia. His research interests are focused on imprecise probability theory, reliability analysis, decision making, risk analysis and learning theory.

Yulia A. Zhuk is currently an associate professor at the Department of Computer Science, Saint-Petersburg State Forest Technical University.

She holds a PhD in Pedagogy and Psychology (2010) from Saint-Petersburg State University, Russia. Her research interests are focused on statistics, learning theory, multimedia technologies, decision making.

1 Introduction

An important problem of the statistical machine learning is the classification problem which can be regarded as a task of classifying some objects into classes (group) in accordance with their properties or features. However, for many real-world problems, the task is not to classify but to detect novel or abnormal instances (Campbell (2002); Campbell and Bennett (2001); Cherkassky and Mulier (2007); Scholkopf et al. (2000, 2001)). Novelty detection is the identification of new or unknown data that a machine learning system is not aware of during training. In particular, it aims to detect anomalous observations (Chandola et al. (2007, 2009); Steinwart et al. (2005)). It should be noted that a typical feature of novelty detection models is that only unlabeled samples are available. We make some assumptions on anomalies in order to distinguish between normal and anomalous future observations. One of the most common ways to define anomalies is by saying that anomalies are not concentrated Scholkopf and Smola (2002). The problem of statistical outlier detection is also closely related to that of novelty detection. Detailed reviews of the novelty detection models can be found in works of Bartkowiak (2011); Hodge and Austin (2004); Khan and Madden (2010); Markou and Singh (2003).

The first way to solve the novelty detection problem is to estimate the real-valued density of the data and then threshold it at some value. It is pointed out by many authors (see, for instance, Cherkassky and Mulier (2007)) that this way is likely to fail for sparse high-dimensional data. A better way is to model the support of the (unknown) data distribution directly from data, that is, to estimate a binary-valued function f that is positive in a region where the density is high, and negative elsewhere. This leads to a single-class learning formulation. The function allows us to specify the region in the input space where the data are explained by the model. Sample points outside this region can be regarded as anomalous observations.

Accepting novelty detection as the one-class classification, many novelty detection models using kernel-based methods in the framework of the support vector machine (SVM) have been proposed. These models are called one-class classification support vector machines. There are two main foundation approaches for constructing the one-class SVM. The first approach is proposed by Tax and Duin (Tax and Duin (1999, 2004)). 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 the one-class SVM consists in determining the smallest hyper-sphere containing the training data. By adapting the kernel function, this approach becomes more flexible than just a sphere in the input space. In a nutshell, the approach considers the trade-off between the number of errors made on the training set (number of target objects rejected) and the size of the sphere (its radius).

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. (Scholkopf et al. (2000, 2001)). Under this approach, a hyperplane is used to separate the training data from the origin with maximal margin, i.e., the objective is to separate off the region containing the data points from the surface region containing no data. Here the authors consider the trade-off between the number of errors made on the training set (number of target objects rejected) and the margin separation between the training points and the origin. This is achieved by constructing a hyperplane which is maximally distant from the origin with all data points lying on the opposite side from the origin.

It should be noted that there are other interesting novelty detection models (see for instance, Bicego and Figueiredo (2009); Campbell and Bennett (2001); Hodge and Austin (2004); Kwok et al. (2007)). However, we study the second approach (Scholkopf et al. (2000, 2001)) and modify or extend it in order to take into account the possible fact that the number of training data might be rather small and the use of the empirical probability distribution might lead to incautious decisions. Our main idea is to construct a robust imprecise model using the framework of an ε-contamination neighborhood or ε-contaminated (robust) models (Huber (1981)).

Many robust classification models assume that each data point or example in the training set can move around within an Euclidean ball. These models stem from perturbations of the training data due to unsatisfactory equipment, corrupted data, etc., which are usually described by an additive noise. At that, these models simultaneously assume that all examples in the training set have equal probabilities or weights. This assumption may be too strong when the number of examples is not large. Therefore, in order to relax the strong condition for probabilities of examples, it is natural to suppose that the probabilities may be contaminated or changed. One of the statistical models taking into account these changes is the imprecise ε-contaminated model Walley (1991) which produces a set of probabilities or probability distributions over the training set. We do not know a precise "true" probability distribution over the training set, but we know that it belongs to the set of probability distributions. So, in contrast to the approach with perturbations of the training data, we assume that each probability assigned to every data point can move around within a set of probabilities under certain restrictions.

We can not solve the classification problem for all probability distributions from the set of distributions. Therefore, in order to construct a robust classifier, we have to select a probability distribution in accordance with a certain decision rule. First, we select the "worst" distribution providing the largest value of the expected risk. It corresponds to the minimax (pessimistic) strategy in decision making and can be interpreted as an insurance against the worst case (Robert (1994)). This is the most popular strategy in the robust classification. The second distribution minimizes the expected risk and corresponds to the minimin (optimistic) strategy which can not be called robust. However, it is interesting as another extreme strategy. By using these probability distributions and the above assumptions, we construct the robust novelty detection model.

The paper is organized as follows. Section 2 presents the standard novelty detection problem proposed by (Scholkopf et al. (2000, 2001)). An ε-contaminated robust model and its peculiarities are considered in Section 3. A set of probability distributions produced by the model and its extreme points which are very important for constructing the novelty detection model are studied in the same section. The minimax strategy and its realization in the SVM approach is given in Section 4. The minimin strategy is 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 Novelty detection

Suppose we have unlabeled training data $\mathbf{x}_1,...,\mathbf{x}_n \subset \mathcal{X}$, where n is the number of observations, $\mathcal{X}$ is some set, for instance, it is a compact subset of $\mathbb{R}^m$. According to Scholkopf et al. (2000, 2001), a well-known novelty detection 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 the kernel and by separating them from the origin with maximum margin.

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

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

σ is the kernel parameter determining the geometrical structure of the mapped samples in the kernel space. It is pointed out by Wang et al. (2009) that the problem of selecting a proper parameter σ is very important in classification. When a very small σ is used ($\sigma \to 0$), $K(\mathbf{x},\mathbf{y}) \to 0$ for all $\mathbf{x} \neq \mathbf{y}$ and all mapped samples tend to be orthogonal to each other, despite their class labels. In this case, both between-class and within-class variations are very large. On the other hand, when a very large σ is chosen ($\sigma^2 \to \infty$), $K(\mathbf{x},\mathbf{y}) \to 1$ and all mapped samples converge to a single point. This obviously is not desired in a classification task. Therefore, a too large or too small σ will not result in more separable samples in G.

It is shown by Campbell (2002) that the data points lie on the surface of a hypersphere in feature space since $\phi(x)\cdot\phi(x) = K(x,x) = 1$. Now we have to find a hyperplane $f(\mathbf{x},\mathbf{w}) = \langle\mathbf{w},\phi(\mathbf{x}_i)\rangle - \rho = 0$ that separates the data from the origin with maximal margin, i.e., we want ρ to be as large as possible so that the volume of the halfspace $\langle\mathbf{w},\phi(\mathbf{x}_i)\rangle \geq \rho$ is minimized. Let us introduce the parameter $\nu \in [0;1]$ which is analogous to ν used for the ν-SVM (Scholkopf and Smola (2002)). Roughly speaking, it denotes the fraction of input data for which $\langle\mathbf{w},\phi(\mathbf{x}_i)\rangle \leq \rho$.

To separate the data set from the origin, we solve the following quadratic program:

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

subject to

$$\langle\mathbf{w},\phi(\mathbf{x}_i)\rangle \geq \rho - \xi_i, \tag{2}$$

$$\xi_i \geq 0,\ i = 1, ..., n. \tag{3}$$

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

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

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

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

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

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

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

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

subject to

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

The value of ρ can be obtained as

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

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

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

3 A robust model and sets of probability distributions

Robust models have been exploited in classification problems due to the opportunity to avoid some strong assumptions underlying the standard classification models. There are several different definitions of robustness in literature. An exhaustive laconic review of robust models in machine learning was given by Xu et al. (2009). As pointed out by Xu et al. (2009), 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, Bouveyron and Girard (2009); Cerioli et al. (2006); Ghaoui et al. (2003); Lanckriet et al. (2002, 2003); Provost and Fawcett (2001); Trafalis and Gilbert (2007); Xu et al. (2006)) in which box-type uncertainty sets are considered.

One of the interpretations of robustness stems from perturbations of the training data. In some cases, the training data and the testing data are sampled from different processes, they may be corrupted, they may be obtained by means of unsatisfactory equipment, etc. In order to take into account these factors, many popular robust classification models are 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, the i-th "true" data point is only known to belong to the interior of an Euclidian ball of radius δ_i centered at the "nominal" data point $\mathbf{x}_i$. This model has a very clear intuitive geometric interpretation (Ben-Tal et al. (2009)). 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. The choice of the maximally robust classifier is explained by the pessimistic strategy in decision making. Applying the above ideas to the SVM with the hinge loss function provides the following optimization problem (see, for instance, a similar problem for binary classification problem Bi and Zhang (2004)):

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

subject to

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

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

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

Another interpretation of robustness stems from the literature on robust statistics (Huber (1981)), which studies how an estimator or algorithm behaves under a small perturbation of the statistics model. There are many reasons for the perturbation of the statistical model. One of the main reasons is a small amount of learning data. Therefore, another class of robust models is based on relaxing strong assumptions about a probability distribution of data points (see, for instance, Lanckriet et al. (2002)). We propose a model which can be partially regarded as a special case of these models and is based on using the framework of *ε-contaminated (robust) models* (Huber (1981)). They are constructed by eliciting a Bayesian prior distribution $p = (p_1, ..., p_n)$ as an estimate of the true prior distribution. The ε-contaminated model is a class of probabilities which for fixed $\varepsilon \in (0, 1)$ and p_i is the set $\mathcal{M}(\varepsilon) = \{(1 - \varepsilon)p_i + \varepsilon q_i\}$, where q_i is arbitrary and $q_1 + ... + q_n = 1$. The rate ε reflect the amount of uncertainty in p (Berger (1985)). In other words, we take an arbitrary probability distribution $q = (q_1, ..., q_n)$ from the unit simplex denoted by $S(1, n)$. Of course, the assumption that q is restricted by the unit simplex $S(1, n)$ is one of possible types of ε-contaminated models. Generally, there are a lot of different assumptions which produce specific robust models.

The most methods for solving classification problems are based on minimizing the expected risk (Vapnik (1998)) and they differ mainly by loss functions. The expected risk can be written as follows:

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

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

$$\begin{aligned} L(\mathbf{w}, \phi(\mathbf{x})) &= \max\{0, \rho - \langle \mathbf{w}, \phi(\mathbf{x})\rangle\} - \rho\nu \\ &= L^*(\mathbf{w}, \phi(\mathbf{x})) - \rho\nu. \end{aligned}$$

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

$$R_{\text{emp}}(\mathbf{w},\rho) = \frac{1}{n}\sum_{i=1}^{n} L^*(\mathbf{w}, \phi(\mathbf{x}_i)) - \rho\nu. \tag{4}$$

The assumption of the empirical probability distribution means that every point $\mathbf{x}_i$ has the probability $p_i = 1/n$. This is a too strong assumption when the number of points is not large. Its validity might give rise to doubt in this case. Therefore, in order to relax the strong condition for probabilities of points, we apply the aforementioned ε-contaminated model. According to the model, we replace the probability distribution $p = (1/n, ..., 1/n)$ accepted in the empirical expected risk by the set of probability distributions $\mathcal{M}(\varepsilon) = \{(1 - \varepsilon)n^{-1} + \varepsilon q_i\}$. In other words, there is an unknown precise “true” probability distribution in $\mathcal{M}(\varepsilon)$, but we do not know it and only know that it belongs to the set $\mathcal{M}(\varepsilon)$. If we have assumed in some robust models (Ben-Tal et al. (2009)) that each point can move around within an Euclidean ball, then the proposed robust model assumes that the probability $1/n$ of each point (but not a data point itself) can move around within a unit simplex under some restrictions. This is the main idea for constructing the robust novel detection models below.

One of the possible ways for dealing with the set $\mathcal{M}(\varepsilon)$ of probability distributions produced by the above constraints is to use the minimax (pessimistic) strategy. According to the minimax strategy, we select a probability distribution from the set $\mathcal{M}(\varepsilon)$ such that the expected risk $R(\mathbf{w},\rho)$ achieves its maximum for every fixed $\mathbf{w}$. It should be noted that the "optimal" probability distributions may be different for different values of parameters $\mathbf{w}$. The minimax strategy can be explained in a simple way. We do not know a precise probability distribution and every distribution from $\mathcal{M}(\varepsilon)$ can be selected. Therefore, we should take the "worst" distribution providing the largest value of the expected risk. The minimax criterion can be interpreted as an insurance against the worst case because it aims at minimizing the expected loss in the least favorable case (Robert (1994)). This criterion of decision making can be regarded as the well-known Γ-minimax[1] (Berger (1985); Gilboa and Schmeidler (1989); Troffaes (2007)).

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

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

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

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

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

$$\begin{aligned}\overline{R}(\mathbf{w},\rho) &= \max_{h\in\mathcal{M}(\varepsilon)} \sum_{i=1}^{n} h_i L^*(\mathbf{w}, \phi(\mathbf{x}_i)) - \rho\nu \\ &= \max_{h\in\mathcal{M}(\varepsilon)} \sum_{i=1}^{n} \left((1-\varepsilon)n^{-1} + \varepsilon q_i\right) L^*(\mathbf{w}, \phi(\mathbf{x}_i)) - \rho\nu \\ &= (1-\varepsilon)R_{\text{emp}}(\mathbf{w},\rho) + \varepsilon \cdot \max_{q} \sum_{i=1}^{n} q_i L^*(\mathbf{w}, \phi(\mathbf{x}_i)) - \rho\nu,\end{aligned}$$

subject to

$$0 \leq q_i \leq 1,\ q_1 + ... + q_n = 1. \tag{5}$$

The obtained optimization problem is linear with optimization variables $q_1, ..., q_n$, but the objective function depends on $\mathbf{w}$. Therefore, it can not be directly solved by well-known methods. In order to overcome this difficulty, note, however, that all points q belong to the simplex $S(1,n)$ 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 of the simplex, and the number of its extreme points is n. Extreme points of the simplex $S(1,n)$ are of the form:

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

This implies that there holds

$$\overline{R}(\mathbf{w},\rho) = (1-\varepsilon)n^{-1} \sum_{i=1}^{n} L^*(\mathbf{w}, \phi(\mathbf{x}_i)) + \varepsilon \cdot \max_{i=1,...,n} L^*(\mathbf{w}, \phi(\mathbf{x}_i)) - \rho\nu. \tag{6}$$

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

4 The minimax strategy and the SVM

The detailed descriptions of approaches for deriving the SVM can be found in works (Cherkassky and Mulier (2007); Scholkopf and Smola (2002); Smola and Scholkopf (2004); Vapnik (1998)). We consider only some features of general approaches, which are specific for the robust model.

First, we add the standard Tikhonov regularization term $\frac{1}{2}\langle \mathbf{w}, \mathbf{w}\rangle$ (this is the most popular penalty or smoothness term) (Tikhonov and Arsenin (1977)) to the objective function (6). 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 Evgeniou et al. (2002) for a detailed analysis of regularization methods). Moreover, we introduce the following optimization variables:

$$\xi_i = \max\left\{0, \rho - \langle \mathbf{w}, \phi(\mathbf{x})\rangle\right\}, \ G = \max_{i=1,...,n} \xi_i.$$

This leads to the quadratic programming problem

$$\overline{R}(\mathbf{w}_{\text{opt}},\rho_{\text{opt}}) = \min_{\mathbf{w},\rho,G,\xi_i} \left(\frac{1}{2}\langle \mathbf{w}, \mathbf{w}\rangle + (1-\varepsilon)n^{-1}\sum_{i=1}^{n}\xi_i + \varepsilon G - \rho\nu \right), \tag{7}$$

subject to

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

$$G \geq \xi_i, \ i = 1, ..., n.$$

One can see from the above optimization problem that it differs from the standard SVM (1)-(3). The objective function is added by the term εG, and there are additional constraints $G \geq \xi_i$ which restrict the set of values of the optimization variable G. These marginal changes, nevertheless, lead to quite different results.

Instead of minimizing the primary objective function (7), a dual objective function, the so-called Lagrangian, can be formed of which the saddle point is the optimum. The Lagrangian is

$$\begin{aligned} L(\mathbf{w}, \xi, \rho, \lambda, \eta, \varphi, G) =& \frac{1}{2}\langle \mathbf{w}, \mathbf{w}\rangle + (1-\varepsilon)n^{-1}\sum_{i=1}^{n}\xi_i + \varepsilon G - \rho\nu \\ & - \sum_{i=1}^{n}\lambda_i\xi_i - \sum_{i=1}^{n}\varphi_i\left(\xi_i - \rho + \langle \mathbf{w},\phi(\mathbf{x}_i)\rangle\right) \\ & - \sum_{i=1}^{n}\eta_i\left(G - \xi_i\right). \end{aligned}$$

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

$$\partial L/\partial \rho = -v + \sum_{i=1}^{n} \varphi_i = 0, \tag{8}$$

$$\partial L/\partial G = \varepsilon - \sum_{i=1}^{n} \eta_i = 0, \tag{9}$$

$$\partial L/\partial \xi_i = (1-\varepsilon)n^{-1} - \lambda_i - \varphi_i + \eta_i = 0, \tag{10}$$

$$\partial L/\partial w_j = w_j - \sum_{i=1}^{n} \varphi_i \phi(x_j^{(i)}) = 0,\ j = 1, ..., m. \tag{11}$$

Using (8)-(10), we simplify the objective function as

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

Finally, substituting w_j from (11), we get the following dual optimization problem

$$L(\eta, \varphi) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \varphi_i \varphi_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{12}$$

subject to

$$0 \leq \varphi_i \leq (1-\varepsilon)n^{-1} + \eta_i,\ i = 1, ..., n, \tag{13}$$

$$\sum_{i=1}^{n} \varphi_i = \nu,\ \sum_{i=1}^{n} \eta_i = \varepsilon,\ \eta_i \geq 0,\ i = 1, ..., n. \tag{14}$$

It is very interesting to note that the objective function does not differ from the objective function obtained in the standard SVM for novel detection with the empirical probability distribution, i.e., by exploiting the empirical expected risk. However, the main difference is in the constraints.

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

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

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

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

Let us consider how the above problem can be modified in the “precise” case when we have a single precise nonparametric probability distribution. In this case, we write $\varepsilon = 0$. Hence, $\eta_i = 0$ and the constraints for φ_i become

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

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

In case of complete ignorance, we can write $\varepsilon = 1$. This implies that a single data point $\mathbf{x}_i$ with the largest loss function $L(\mathbf{w}, \phi(\mathbf{x}_i))$ defines the decision function f.

Let us write the Karush-Kuhn-Tucker complementarity conditions

$$\begin{aligned} \varphi_i \left(\xi_i - \rho + \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle \right) &= 0, \\ \eta_i \left(G - \xi_i \right) &= 0. \end{aligned}$$

It follows from the second condition that $G = \xi_i$ for a single value of $i = k$ such that $k = \arg\max_{i=1,...,n} \xi_i$. Here we assume that $\xi_1, ..., \xi_n$ do not coincide. Therefore, $\eta_i = 0$ for all $i \neq k$. Returning to the constraints (13)-(14), we get $\eta_k = \varepsilon$ and

$$0 \leq \varphi_i \leq (1-\varepsilon)n^{-1}, \ i = 1, ..., n, \ i \neq k,$$

$$0 \leq \varphi_i \leq (1-\varepsilon)n^{-1} + \varepsilon, \ i = k.$$

It follows from the above constraints that the optimization problem (12)-(14) can be decomposed into n problems

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

subject to

$$0 \leq \varphi_i \leq (1-\varepsilon)n^{-1} + \varepsilon \cdot \mathbf{1}_k(i), \ i = 1, ..., n, \ \sum_{i=1}^{n} \varphi_i = \nu.$$

Here $\mathbf{1}_k(i)$ is the indicator function taking the value 1 if $i = k$. The optimal values of φ_i, $i = 1, ..., n$, correspond to the *smallest* value of objective function L_k, $k = 1, ..., n$.

It should be noted that the optimization problem (12)-(14) has $2n$ variables and $3n + 2$ constraints. As a result, it may be hard to solve it especially when there are many observations. Therefore, the above decomposition is an important result because we get n simple optimization problems having n variables and $2n + 1$ constraints.

5 The minimin strategy and the SVM

The minimin strategy can be regarded as a direct opposite to the minimax strategy. According to the minimin strategy, the expected risk R is minimized over all probability distributions from the set $\mathcal{M}(\varepsilon)$ as well as over all values of parameters $\mathbf{w}$, ρ. The strategy can be called optimistic because it selects the "best" probability distribution from the set $\mathcal{M}(\varepsilon)$.

Similarly to the minimax strategy, we can write

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

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

$$\underline{R}(\mathbf{w},\rho) = (1-\varepsilon)R_{\mathrm{emp}}(\mathbf{w},\rho) + \varepsilon \cdot \min_{q} \sum_{i=1}^{n} q_i L^*(\mathbf{w},\phi(\mathbf{x}_i)) - \rho\nu,$$

subject to (5).

Hence, there holds

$$\underline{R}(\mathbf{w},\rho) = \min_{k=1,\dots,n} \left\{ (1-\varepsilon)n^{-1}\sum_{i=1}^{n} L^*(\mathbf{w},\phi(\mathbf{x}_i)) + \varepsilon \cdot L^*(\mathbf{w},\phi(\mathbf{x}_k)) - \rho\nu \right\}.$$

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

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

subject to

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

The Lagrangian is

$$\begin{aligned} L_k(\mathbf{w},\xi,\rho,\lambda,\varphi) =& \frac{1}{2}\langle \mathbf{w},\mathbf{w}\rangle + (1-\varepsilon)n^{-1}\sum_{i=1}^{n}\xi_i + \varepsilon\xi_k - \rho\nu \\ & - \sum_{i=1}^{n}\lambda_i\xi_i - \sum_{i=1}^{n}\varphi_i\left(\xi_i - \rho + \langle \mathbf{w},\phi(\mathbf{x}_i)\rangle\right). \end{aligned}$$

Here λ_i, φ_i, $i = 1,...,n$, are Lagrange multipliers. The following dual optimization problem can be derived from the Lagrangian

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

subject to

$$0 \leq \varphi_i \leq (1-\varepsilon)n^{-1},\ i = 1, ..., n, ,\ i \neq k,$$

$$0 \leq \varphi_k \leq (1-\varepsilon)n^{-1} + \varepsilon,\ \sum_{i=1}^{n} \varphi_i = \nu.$$

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

6 Experiments

We illustrate the method proposed in this paper via several examples, all computations have been performed using the statistical software R. We investigate the performance of the proposed method and compare it with the standard SVM approach by considering the accuracy (ACC), which is the proportion of correctly classified cases on a sample of data and is often used to quantify the predictive performance of classification methods. ACC is an estimate of a classifier's probability of a correct response, and it is an important statistical measures of the performance a one-class classification test. In novelty detection, ACC is the sum of two accuracy measures: the normal accuracy rate which measures how well the algorithm recognizes new examples of the known examples, and the novelty accuracy rate which does the same for examples of an unknown novel example. ACC can formally be written as

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

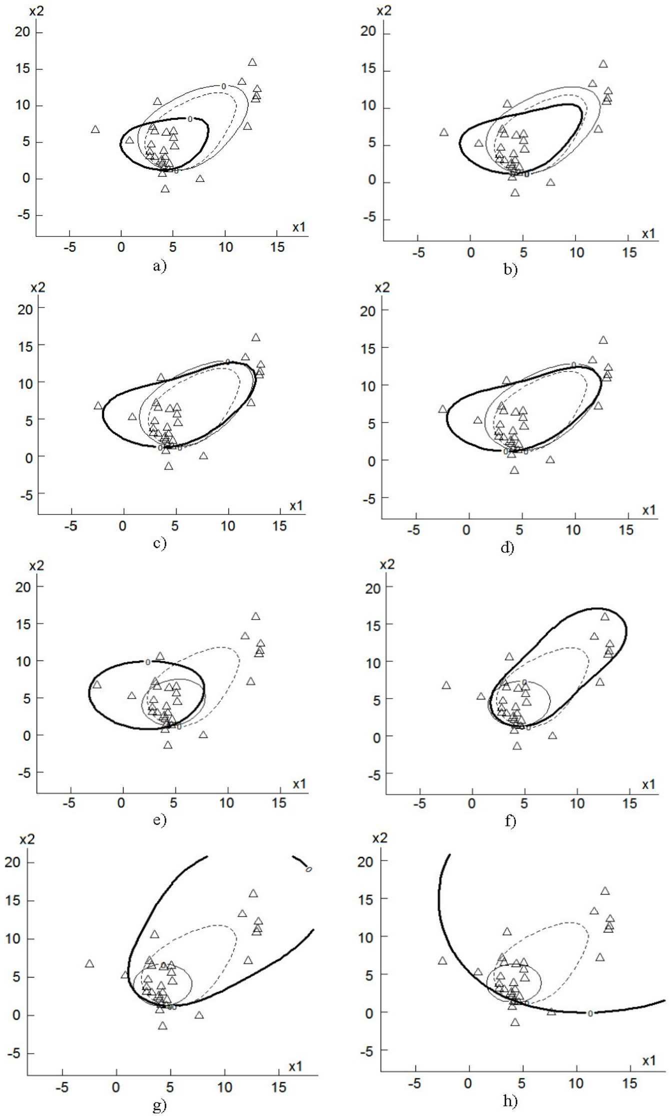
where y_i is the label of the i-th test example $\mathbf{x}_i$.

We will denote the accuracy measure for the proposed minimax strategy as ACC_{mx}, for the proposed minimin strategy as ACC_{mn} and for the standard SVM as ACC_{st}.

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

We first consider the performance of our method for a small example with synthetic data having two features x_1 and x_2. The training set consisting of two subsets is generated in accordance with the normal probability distributions such that $N_1 = (1-\varepsilon_0)N$ examples (the first subset) are generated with mean values $\mathbf{m}_1 = (4, 4)$ and $N_2 = \varepsilon_0 N$ examples (the second subset) have mean values $\mathbf{m}_2 = (12, 12)$. The standard deviation is $s = 2$ for both subsets and both features. Here ε_0 is a portion of abnormal examples in training set. The kernel parameter σ is 0.01. The parameters ν and ε_0 are 0.2.

Fig. 1 illustrates how the contours $f(\mathbf{x}, \mathbf{w}) = 0$ depend on the robust parameter ε for three models corresponding to the minimax strategy (thick curve), to the minimin strategy (thin curve) and to the standard SVM (dashed curve). The



a) b) c) d) e) f) g) h)

Figure 1 The contours $f(\mathbf{x}, \mathbf{w}) = 0$ by different robust parameters ε

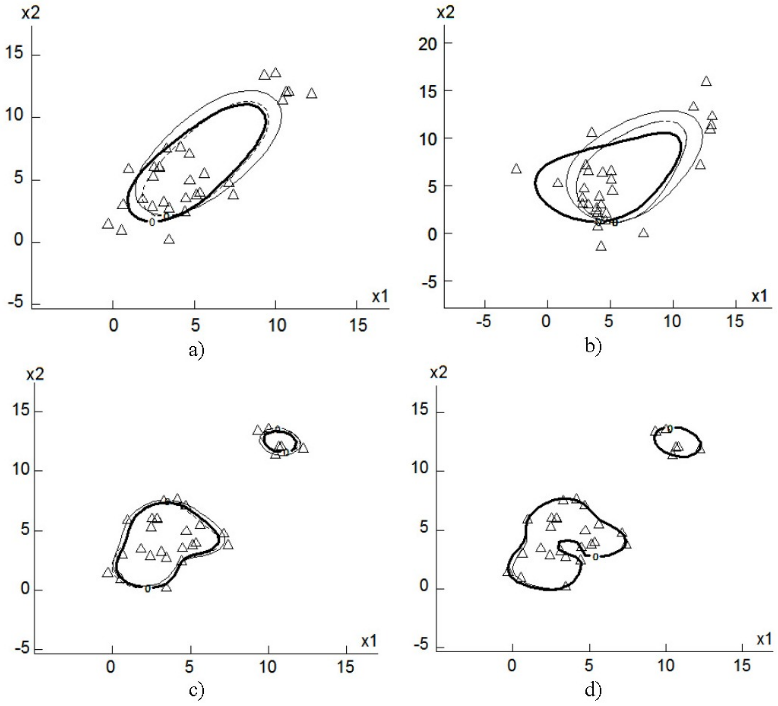


Figure 2 The contours $f(\mathbf{x},\mathbf{w})=0$ by different kernel parameters σ

parameter ε takes values 0.1 (a), 0.2 (b), 0.3 (c), 0.4 (d), 0.7 (e), 0.8 (f), 0.9 (g), 0.99 (h). One can see from the pictures that the region bounded by the minimax contour increases with ε. The reason of this peculiarity is that the minimax strategy assigns the largest weight $(1-\varepsilon)/n+\varepsilon$ to a single "bad" point and identical weights $(1-\varepsilon)/n$ to other $n-1$ points. It is obvious that the "bad" point lies in a large distance from the origin in the feature space G in order to maximize the expected risk. When the parameter ε increases, the weight of the "bad" points increases and weights of other points become to be rather small. Hence, the minimax contour tends to cover this "bad" point from the subset of abnormal examples by large values of ε. On the contrary, the region bounded by the minimin contour decreases with ε because the minimin strategy assigns the largest weight $(1-\varepsilon)/n+\varepsilon$ to a single "good" point which lies very close to the origin. When the parameter ε increases, the minimin contour tends to cover only this point. Since we want ρ to be as large as possible, the the contour converges to the "good" point.

Let us investigate how the models analyzed depend on the kernel parameter σ. The parameter σ takes values 0.008 (a), 0.01 (b), 0.08 (c), 0.1 (d). This dependence can be observed in Fig. 2, where every picture corresponds to a value of σ. It is interesting to note that all novelty detection models transform to the cluster models by $\sigma=0.08$ and $\sigma=0.1$. These parameters lead to "overfitting" when we deal with novelty detection because abnormal examples are recognized as a separate cluster.

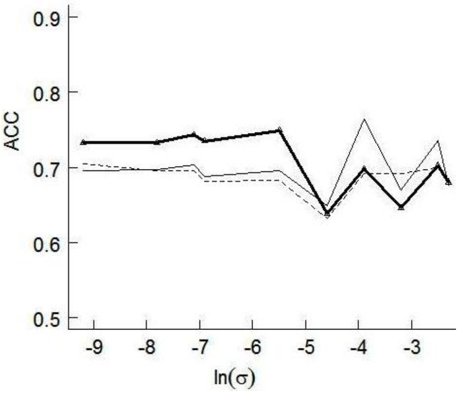


Figure 3 Accuracy measures by different kernel parameters

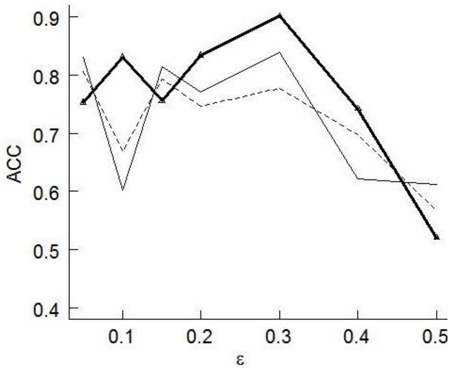


Figure 4 Accuracy measures by different robust parameters

Fig. 3 shows how the accuracy measures for all proposed models depend on the kernel parameter σ. One can see from the picture that there is a region of σ ($\ln \sigma < -5$), where the minimax model provides better accuracy in comparison with other models. At the same time, we observe the advantage of the minimin and standard models by $\ln \sigma > -5$. Other parameters for the experiment are $n = 30$, $\varepsilon = 0.2$, $s = 2$, $\mathbf{m}_1 = (4, 4)$, $\mathbf{m}_2 = (12, 12)$, $\nu = 0.2$, $\varepsilon_0 = 0.2$.

Taking the same parameters and fixing $\sigma = 0.004$, we investigate how accuracy measures depend on the robust parameter ε. The corresponding curves are depicted in Fig. 4.

If we increase the distance between the region where the data mainly lie and the abnormal data by taking $\mathbf{m}_1 = (4, 4)$, $\mathbf{m}_2 = (20, 20)$ with parameters $n = 30$, $\varepsilon = 0.2$, $\sigma = 0.009$, $s = 2$, $\varepsilon_0 = 0.2$, then it is interesting to consider how accuracy measures of models depend on the parameter v. The corresponding curves are shown in Fig. 5. One can see from Fig. 5 that the minimax strategy provides better results by large values of the parameter v. This fact can be explained as follows. The increased parameter v leads to a larger number of examples which can be viewed as abnormal. This implies that some "normal" points are viewed as abnormal and the corresponding contours converge in this case. However, the

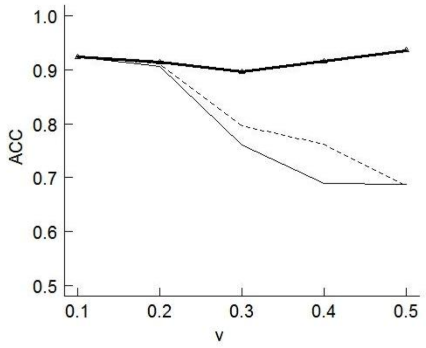


Figure 5 Accuracy measures by different parameters v

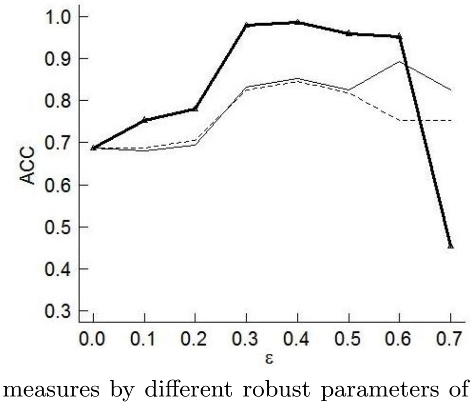


Figure 6 Accuracy measures by different robust parameters of the Iris data set

minimax strategy assigns the largest weight to a point which lies in a large distance from the origin. This feature compensates the influence of v.

As a further example we applied all analyzed models to the well-known “Iris” data set from the UCI Machine Learning Repository (Frank and Asuncion (2010)). The data set contains 3 classes (Iris Setosa, Iris Versicolour, Iris Virginica) of 50 instances each. The number of features is 4 (sepal length in cm, sepal width in cm, petal length in cm, petal width in cm). We suppose that data from the Iris Setosa class are abnormal. The number of training data n is 30. For the experiment, we randomly select n points such that $(1-\varepsilon_0)n$ points are taken from the set of positively labelled examples and $\varepsilon_0 n$ points are from negatively labelled examples. Here $\varepsilon_0 = 50/150 \simeq 0.333$. The parameters for modelling are $v = 0.333$, $\sigma = 0.001$. Fig. 6 illustrates how accuracy measures depend on the robust parameter ε. One can see that the optimal values of the parameter are in the interval from 0.3 to 0.4 where the accuracy measure has the largest value. This says about correspondence or closeness of the optimal values of ε and the parameter ε_0.

We also investigate how the accuracy measures depend on the amount n of training data. In particular, if we take $n = 50$, then $ACC_{\mathrm{mx}} = 0.947$, $ACC_{\mathrm{mn}} =$

0.707, $ACC_{\mathrm{st}} = 0.773$. If we take $n = 80$, then $ACC_{\mathrm{mx}} = 0.833$, $ACC_{\mathrm{mn}} = 0.62$, $ACC_{\mathrm{st}} = 0.713$. It follows from the above experiments that there is some optimal number of training examples providing the largest accuracy measures.

As a further example we consider the "Pima Indian Diabetes" data set, also from the UCI Machine Learning Repository. The Pima data set has eight features ($m = 8$), with 768 training instances (examples) of which 500 are labeled as positive. We view 268 positive examples as abnormal observations. Then it is supposed that $\varepsilon_0 = 268/768 = 0.35$. For the experiment, we randomly select n points such that $(1 - \varepsilon_0)n$ points are taken from the set of positively labelled examples and $\varepsilon_0 n$ points are from negatively labelled examples. If we take $n = 50$, then $ACC_{\mathrm{mx}} = 0.683$, $ACC_{\mathrm{mn}} = 0.656$, $ACC_{\mathrm{st}} = 0.66$. If we take $n = 100$, then $ACC_{\mathrm{mx}} = 0.684$, $ACC_{\mathrm{mn}} = 0.677$, $ACC_{\mathrm{st}} = 0.667$. It can be seen from the results, that the minimax strategy provides better results in comparison with the minimin and the standard models. However, this is an incidental advantage. It is difficult to select abnormal data in the "Pima Indian Diabetes" data set due to its structure.

7 Conclusion

In this paper, a robust novelty detection model has been proposed, which is based on the ε-contaminated model. The simplicity of the model follows from the finite number of extreme points of the set of probability distributions produced by the robust ε-contaminated model and their simple determination. Both the minimax and minimin strategies have a clear explanation and justification in the framework of decision theory. It should be noted that the algorithm for computing the optimal parameters of the novelty detection model is reduced to a finite number of standard SVM tasks with weighted data points (Bicego and Figueiredo (2009); Yang et al. (2007)), where the weights are assigned in accordance with a predefined rule. In contrast to the weighted models, the weights of data points are probabilities of the points assigned in a specific way which is determined by an underlying imprecise probabilistic model. Moreover, the proposed robust model deals with a set of weighted SVM tasks such that every task has different weights defined by the extreme points of the probability set.

Experimental results with synthetic and some real data reported have shown that the proposed robust novelty detection model outperforms the standard approach proposed by Scholkopf et al. (2000, 2001) for certain initial parameters.

One of the important advantages of the minimax strategy is that it allows us to process small training sets. Indeed, when we have a large amount of data for training and normal observations are concentrated, the standard approach (Tax and Duin (1999, 2004)) properly works. However, it is difficult to select abnormal observations when the training set is small because these observations have the same probabilities and are viewed as normal ones. The minimax strategy assigns the largest probabilities (weights) to points which are outlying in order to separate the concentrated and outlying points. Even if we have a small training set, then we artificially separate the data points by assigning different probabilities to these points.

Another advantage of the proposed model is that it reflects the possible violation of strong assumptions about uniformity of the probability mass function accepted in the standard approach during testing. This violation is taken into account by the second distribution q in the ε-contaminated model.

An important direction for future work is to apply the ε-contaminated model to the novelty detection model proposed by Tax and Duin (1999, 2004), to the model proposed by Campbell and Bennett (2001), which uses linear programming techniques. Another interesting direction for further research is to investigate the "double" robust model where we combine two robust approaches: the first approach is when each data point is only known to belong to the interior of an Euclidian ball; the second approach is the ε-contaminated model for empirical probabilities.

It should be noted that two "extreme" strategies (minimax and minimin) have been considered. However, it is interesting to analyze the so-called cautious decision making which is a linear combination of the minimax and minimin strategies and 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 (2005). The use of ideas underlying the method can help to construct new efficient novelty detection models.

Finally, it is also worth noticing that the proposed model 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 whose very valuable comments have improved the paper.

References

Bartkowiak, A. (2011). Anomaly, novelty, one-class classification: A comprehensive introduction. *International Journal of Computer Information Systems and Industrial Management Applications*, 3:61–71.

Ben-Tal, A., Ghaoui, L., and Nemirovski, A. (2009). *Robust optimization*. Princeton University Press, Princeton, New Jersey.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

Bi, J. and Zhang, T. (2004). Support vector classification with input data uncertainty. In Saul, L., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 17, pages 161–168. MIT Press, Cambridge, MA.

Bicego, M. and Figueiredo, M. (2009). Soft clustering using weighted one-class support vector machines. *Pattern Recognition*, 42:27–32.

Bouveyron, C. and Girard, S. (2009). Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, 42(11):2649 – 2658.

Campbell, C. (2002). Kernel methods: a survey of current techniques. *Neurocomputing*, 48(1-4):63–84.

Campbell, C. and Bennett, K. (2001). A linear programming approach to novelty detection. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13, pages 395–401. MIT Press.

Cerioli, A., Riani, M., and Atkinson, A. (2006). Robust classification with categorical variables. In Rizzi, A. and Vichi, M., editors, *Compstat 2006 - Proceedings in Computational Statistics*, pages 507–519. Physica-Verlag HD.

Chandola, V., Banerjee, A., and Kumar, V. (2007). Anomaly detection: A survey. Technical Report TR 07-017, University of Minnesota, Minneapolis, MN, USA.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41:1–58.

Cherkassky, V. and Mulier, F. (2007). *Learning from Data: Concepts, Theory, and Methods*. Wiley-IEEE Press, UK.

Evgeniou, T., Poggio, T., Pontil, M., and Verri, A. (2002). Regularization and statistical learning theory for data analysis. *Computational Statistics & Data Analysis*, 38(4):421 – 432.

Frank, A. and Asuncion, A. (2010). UCI machine learning repository.

Ghaoui, L., Lanckriet, G., and Natsoulis, G. (2003). Robust classification with interval data. Technical Report Report No. UCB/CSD-03-1279, University of California, Berkeley, California 94720.

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153.

Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126.

Huber, P. (1981). *Robust Statistics*. Wiley, New York.

Khan, S. and Madden, M. (2010). A survey of recent trends in one class classification. In Coyle, L. and Freyne, J., editors, *Artificial Intelligence and Cognitive Science*, volume 6206 of *Lecture Notes in Computer Science*, pages 188–197. Springer Berlin / Heidelberg.

Kwok, J., Tsang, I.-H., and Zurada, J. (2007). A class of single-class minimax probability machines for novelty detection. *IEEE Transactions on Neural Networks*, 18(3):778–785.

Lanckriet, G., Ghaoui, L., Bhattacharyya, C., and Jordan, M. (2002). A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582.

Lanckriet, G., Ghaoui, L., and Jordan, M. (2003). Robust novelty detection with single-class mpm. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems*, volume 15, pages 905–912. MIT Press, Cambridge, MA.

Markou, M. and Singh, S. (2003). Novelty detection: a reviewpart 1: statistical approaches. *Signal Processing*, 83(12):2481–2497.

Provost, F. and Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231.

Robert, C. (1994). *The Bayesian Choice*. Springer, New York.

Scholkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., and Williamson, R. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471.

Scholkopf, B. and Smola, A. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, Massachusetts.

Scholkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., and Platt, J. (2000). Support vector method for novelty detection. In *Advances in Neural Information Processing Systems*, pages 526–532.

Smola, A. and Scholkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14:199–222.

Steinwart, I., Hush, D., and Scovel, C. (2005). A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6:211–232.

Tax, D. and Duin, R. (1999). Support vector domain description. *Pattern Recognition Letters*, 20:1191–1199.

Tax, D. and Duin, R. (2004). Support vector data description. *Machine Learning*, 54:45–66.

Tikhonov, A. and Arsenin, V. (1977). *Solution of Ill-Posed Problems.* W.H. Winston, Washington DC.

Trafalis, T. and Gilbert, R. (2007). Robust support vector machines for classification and computational issues. *Optimization Methods and Software*, 22(1):187–198.

Troffaes, M. (2007). Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29.

Utkin, L. and Augustin, T. (2005). Efficient algorithms for decision making under partial prior information and general ambiguity attitudes. In F.G. Cozman, R. Nau, T. S., editor, *Proc. of the 4th Int. Symposium on Imprecise Probabilities and Their Applications, ISIPTA'05*, pages 349–358, Pittsburgh, USA. Carnegie Mellon University, SIPTA.

Vapnik, V. (1998). *Statistical Learning Theory.* Wiley, New York.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London.

Wang, J., Lu, H., Plataniotis, K., and Lu, J. (2009). Gaussian kernel optimization for pattern classification. *Pattern Recognition*, 42(7):1237 – 1247.

Xu, H., Caramanis, C., and Mannor, S. (2009). Robustness and regularization of support vector machines. *The Journal of Machine Learning Research*, 10:1485–1510.

Xu, L., Crammer, K., and Schuurmans, D. (2006). 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. AAAI Press; MIT Press.

Yang, X., Song, Q., and Wang, Y. (2007). A weighted support vector machine for data classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 21(5):961–976.