Research Article

Fuzzy One-Class Classification Model Using Contamination Neighborhoods

Lev V. Utkin

Department of Control, Automation and System Analysis, St. Petersburg State Forest Technical University, Institutski per. 5, St. Petersburg 194021, Russia

Correspondence should be addressed to Lev V. Utkin, lev.utkin@gmail.com

Received 8 April 2012; Accepted 16 August 2012

Academic Editor: M. Onder Efe

Copyright © 2012 Lev V. Utkin. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A fuzzy classification model is studied in the paper. It is based on the contaminated (robust) model which produces fuzzy expected risk measures characterizing classification errors. Optimal classification parameters of the models are derived by minimizing the fuzzy expected risk. It is shown that an algorithm for computing the classification parameters is reduced to a set of standard support vector machine tasks with weighted data points. Experimental results with synthetic data illustrate the proposed fuzzy model.

1. Introduction

A main goal of the statistical machine learning is to predict an unobserved output value y based on an observed input vector x . A special very important problem of the statistical machine learning is the classification problem which can be regarded as a task of classifying some objects into classes in accordance with their properties or features. A lot of models have been constructed for solving the machine learning problems in last decades. However, a large part of models is based on restrictive assumptions, for instance, the large amount of training data, the known type of the noise probability distribution, point-valued observations, and so forth. At the same time, real applications cannot satisfy all these assumptions or their part due to several reasons. In order to relax some of the restrictive assumptions, many methods have been proposed. One of the directions for developing the corresponding methods is the fuzzy classification which applies the main ideas of fuzzy set theory to various classification problems.

Most fuzzy classification models can conditionally be divided into three groups. Two groups suppose that there are precise observations, but fuzzy sets are used to take into account different contributions of the observations or in order to take into account imprecision of results,

for instance, imprecision of separating surfaces. According to models of the first group with precise observations, a membership value or a membership function is assigned to every point from the training set in accordance with some rules [1–5]. As a result, different input points can make different contributions to the learning of decision surface. Then the initial classification problem is solved as a weighted classification problem by means of known methods, for example, by means of the support vector machine (SVM) approach proposed by Vapnik [6]. According to models of the second group, a fuzzy separating surface, for instance, a fuzzy hyperplane in the feature space, is constructed [7, 8]. In other words, a result of the classification problem solution is a set of hyperplanes with a membership function derived from the model. The third group of the models supposes that learning observations are interval-valued or fuzzy-valued themselves [9–11] due to imperfection of measurement tools or imprecision of expert information if used as data. It should be noted that methods based on fuzzy sets can be used for generalizing the two-class classification into multiclass classification. In particular, Wilk and Wozniak [12] proposed the corresponding method by exploiting a fuzzy inference system.

The main difficulty of most fuzzy models is how to determine the membership functions for points from the training set. Therefore, a fuzzy classification model is proposed in the

given paper which constructs the membership function on the basis of available statistical data by using an extension of the well-known ε -contamination neighborhood or ε -contaminated (robust) models [13]. Main ideas underlying the proposed model are the following.

The empirical probability distribution accepted for deriving the empirical expected risk (a measure of the classification error) and used in the standard SVM is replaced by a set of probability distributions, which is produced by applying the ε -contaminated model. At that, it is assumed that the contaminating probability distribution can be arbitrary. As a result, we could construct an imprecise classification model for a predefined value ε . However, the main difficulty of using the ε -contaminated model is that the value ε is unknown, and there are no rules or methods for its determination. Therefore, the next idea is to use all values ε and to construct a model which covers all values ε .

Note that the produced set of probability distribution on the basis of the ε -contaminated model for a fixed ε is convex. This implies that the expected risk as an expectation of a loss function lies in an interval with some lower and upper bounds. Moreover, we get a set of nested intervals for different values of ε ranging from 0 to 1, which can be regarded as a fuzzy set. This implies that we obtain the fuzzy expected risk parametrized by classification parameters or by parameters of a separating surface. The next task is to find the classification parameters which minimize the fuzzy expected risk. This task is solved in the framework of SVM by taking a special ranking index for ranging the fuzzy expected risk measures. The fuzziness of the expected risk measure strongly depends on the number of points in the training set. It is assumed that the fuzziness increases with decrease of the number of data points.

It can be seen from the previous idea that the proposed model encompasses peculiarities of the first and the second group of models.

In order to simplify the description of the proposed classification model, we consider the one-class classification model or the well-known novelty detection model [14–17]. Moreover, we restrict ourselves by studying a model proposed by Schölkopf et al. [16, 18]. Nevertheless, this model can easily be extended in the case of binary classification models. It is important to show main principles for constructing the fuzzy classification model in the paper.

The paper is organized as follows. Section 2 presents the standard one-class classification problem proposed by [16, 18]. The ε -contaminated robust model and its peculiarities are considered in Section 3. A set of probability distributions produced by the model and the fuzzy expected risk measure are studied in the same section. The SVM approach for computing the optimal classification parameters is provided in Section 4. In this section, one of the possible ways for decomposing the quadratic programming problem is considered. Numerical experiments with synthetic and some real data illustrating accuracy of the proposed model are provided in Section 5. In Section 6, concluding remarks are given.

2. One-Class Classification

Suppose we have unlabeled training data $\mathbf{x}_1, \dots, \mathbf{x}_n \subset \mathcal{X}$, 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 papers [16, 18], a well-known one-class classification (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 $\mathcal{X} \rightarrow G$ 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:

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma^2}\right). \quad (1)$$

σ is the kernel parameter determining the geometrical structure of the mapped samples in the kernel space.

The aim of classification is 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. We use the parameter $\nu \in [0; 1]$ which is analogous to ν used for the ν -SVM [19]. It denotes the fraction of input data for which $\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle \leq \rho$.

To find the optimal parameters $\mathbf{w}$ and ρ , the following quadratic program has to be solved:

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

subject to

$$\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, n. \quad (3)$$

Slack variables ξ_i are used to allow points to violate margin constraints. Using multipliers $\alpha_i, \beta_i \geq 0$, we introduce a Lagrangian:

$$\begin{aligned} L(\mathbf{w}, \xi, \rho, \alpha, \beta) = & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{\nu n} \sum_{i=1}^n \xi_i - \rho \\ & - \sum_{i=1}^n \alpha_i (\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle - \rho + \xi_i) - \sum_{i=1}^n \beta_i \xi_i. \end{aligned} \quad (4)$$

It is shown in [19] that the dual problem is of the form

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j), \quad (5)$$

subject to

$$0 \leq \alpha_i \leq \frac{1}{\nu n}, \quad \sum_{i=1}^n \alpha_i = 1. \quad (6)$$

The value of ρ can be obtained as

$$\rho = \left(\langle \mathbf{w}, \phi(\mathbf{x}_j) \rangle \right) = \sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \mathbf{x}_j). \quad (7)$$

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

$$f(\mathbf{x}, \mathbf{w}) = \text{sgn} \left(\sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \mathbf{x}) - \rho \right). \quad (8)$$

3. Robust Models and Fuzzy Expected Losses

3.1. Robust Models and the Expected Risk. Robust models have been exploited in classification problems due to the opportunity to avoid some strong assumptions underlying the standard classification models. As pointed out by Xu et al. [20], 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, e.g., [21–24]) 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, and every data point is only known to belong to the interior of an Euclidian ball. Another class of robust models is based on relaxing strong assumptions about a probability distribution of data points (see, e.g., [25]).

We consider 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 ([13]). They are constructed by eliciting a Bayesian prior distribution $p = (p_1, \dots, p_n)$ 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 $q_1 + \dots + q_n = 1$. The rate ε reflects the amount of uncertainty in p [26]. In other words, we take an arbitrary probability distribution $q = (q_1, \dots, q_n)$ from the unit simplex denoted by $S(1, n)$. According to these models, for $0 < \varepsilon < 1$, $\mathcal{M}(\varepsilon)$ 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 $S(1, n)$ is one of possible types of ε -contaminated models. Generally, there are a lot of different assumptions which produce specific robust models.

Let us rewrite the problem in a general form of minimizing the expected risk [6]

$$R(\mathbf{w}, \rho) = \int_{\mathbb{R}^m} L(\mathbf{x}) dF_0(\mathbf{x}). \quad (9)$$

Here the loss function $L(\mathbf{x})$ is the hinge loss function which is represented as

$$L(\mathbf{x}) = \max\{0, \rho - \langle \mathbf{w}, \phi(\mathbf{x}) \rangle\} - \rho\nu. \quad (10)$$

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

$$R_{\text{emp}}(\mathbf{w}, \rho) = \frac{1}{n} \sum_{i=1}^n L(\mathbf{w}, \phi(\mathbf{x}_i)) - \rho\nu. \quad (11)$$

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 ε -contaminated model. According to the model, we replace the probability distribution $p = (1/n, \dots, 1/n)$ 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 $\mathcal{M}(\varepsilon)$.

3.2. The Fuzzy Expected Risk. Let $h = (h_1, \dots, h_n)$ be a probability distribution which belongs to the set $\mathcal{M}(\varepsilon)$. Since the set $\mathcal{M}(\varepsilon)$ is convex, then the values of the expected risk for every value ε are restricted by some lower and upper bounds such that every point of the interval of expected risk corresponds to a probability distribution from $\mathcal{M}(\varepsilon)$. Note that $\mathcal{M}(\varepsilon_1) \subseteq \mathcal{M}(\varepsilon_2)$ by $\varepsilon_1 \leq \varepsilon_2$. This implies that there holds

$$[R_L(\cdot, \varepsilon_1); R_U(\cdot, \varepsilon_1)] \subseteq [R_L(\cdot, \varepsilon_2); R_U(\cdot, \varepsilon_2)]. \quad (12)$$

If we consider all possible values of the parameter ε , then we get the set of nested intervals $[R_L(\mathbf{w}, \rho, \varepsilon); R_U(\mathbf{w}, \rho, \varepsilon)]$ with the parameter ε . Moreover, the expected risk interval is reduced to a point by $\varepsilon = 0$, and the largest interval takes place by $\varepsilon = 1$. This set of intervals can be viewed as a fuzzy set $\tilde{R}(\mathbf{w}, \rho)$ of the expected utility with the membership function μ . The similar idea to obtain fuzzy sets by means of the contamination robust model has been mentioned by Utkin and Zhuk [27].

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

$$\begin{aligned} R_U(\mathbf{w}, \rho, \varepsilon) &= \max_{h \in \mathcal{M}(\varepsilon)} \sum_{i=1}^n h_i L(\mathbf{w}, \phi(\mathbf{x}_i)) - \rho\nu \\ &= \max_{h \in \mathcal{M}(\varepsilon)} \sum_{i=1}^n ((1 - \varepsilon)n^{-1} + \varepsilon q_i) L(\mathbf{w}, \phi(\mathbf{x}_i)) - \rho\nu \\ &= (1 - \varepsilon)R_{\text{emp}}(\mathbf{w}, \rho) + \varepsilon \cdot \max_q \sum_{i=1}^n q_i L(\mathbf{w}, \phi(\mathbf{x}_i)) \\ &\quad - \rho\nu, \end{aligned} \quad (13)$$

subject to

$$0 \leq q_i \leq 1, \quad q_1 + \dots + q_n = 1. \quad (14)$$

The obtained optimization problem is linear with optimization variables $q_1, \dots, q_n$, but the objective function depends on $\mathbf{w}$. Therefore, it cannot 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 previously mentioned 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

$$(1, 0, \dots, 0), (0, 1, \dots, 0), \dots, (0, 0, \dots, 1). \quad (15)$$

This implies that there holds

$$R_U(\mathbf{w}, \rho, \varepsilon) = (1 - \varepsilon)n^{-1} \sum_{i=1}^n L(\mathbf{w}, \phi(\mathbf{x}_i)) + \varepsilon \cdot \max_{i=1, \dots, n} L(\mathbf{w}, \phi(\mathbf{x}_i)) - \rho\nu. \quad (16)$$

The same can be written for the lower bound:

$$R_L(\mathbf{w}, \rho, \varepsilon) = (1 - \varepsilon)n^{-1} \sum_{i=1}^n L(\mathbf{w}, \phi(\mathbf{x}_i)) + \varepsilon \cdot \min_{i=1, \dots, n} L(\mathbf{w}, \phi(\mathbf{x}_i)) - \rho\nu. \quad (17)$$

3.3. The Fuzzy Decision Problem and a Way for Its Solving. Now we can write a new criterion of decision making about optimal parameters $\mathbf{w}, \rho$. The parameters $\mathbf{w}_0, \rho_0$ are optimal iff for all $\mathbf{w}, \rho$, there holds $\tilde{R}(\mathbf{w}_0, \rho_0) \leq \tilde{R}(\mathbf{w}, \rho)$. The next question is how to compare the fuzzy sets. This is one of the most controversial questions in fuzzy literature. Most ranking methods are based on transforming a fuzzy set into a real number called by the ranking index. Here we use the index proposed by [28], which can be written in terms of the considered decision problem as

$$R(\mathbf{w}, \rho) = \int_0^1 (\eta \cdot R_L(\mathbf{w}, \rho, \varepsilon) + (1 - \eta) R_U(\mathbf{w}, \rho, \varepsilon)) d\mu. \quad (18)$$

Here $\eta \in [0, 1]$ is a parameter of pessimism. It can be regarded as a caution parameter proposed by [29]. The caution parameter reflects the degree of ambiguity aversion. The more ambiguity averse the decision maker is, the higher is the influence of the lower interval limit of generalized expected utility. $\eta = 1$ corresponds to strict ambiguity aversion; $\eta = 0$ expresses maximal ambiguity seeking attitudes. It is assumed here that the variable μ is a function of ε , for example, $\mu = 1 - \varepsilon$.

Then we write a new criterion of decision making. Parameters $\mathbf{w}_0, \rho_0$ are optimal iff for all $\mathbf{w}, \rho$, there holds

$$R(\mathbf{w}_0, \rho_0) \leq R(\mathbf{w}, \rho). \quad (19)$$

It follows from the above that the optimal parameters $\mathbf{w}, \rho$ can be obtained by solving the following optimization problem:

$$R(\mathbf{w}_0, \rho_0) = \min R(\mathbf{w}, \rho). \quad (20)$$

Denote $A = \int_0^1 \varepsilon d\mu$. The optimal expected risk is now of the form

$$R(\mathbf{w}_{\text{opt}}, \rho_{\text{opt}}) = \min_{\mathbf{w}, \rho} R(\mathbf{w}, \rho), \quad (21)$$

where

$$R(\mathbf{w}, \rho) = \frac{1 - A}{n} \sum_{i=1}^n L(\mathbf{w}, \phi(\mathbf{x}_i)) - \rho\nu + A\eta \cdot \min_{i=1, \dots, n} L(\mathbf{w}, \phi(\mathbf{x}_i)) + A(1 - \eta) \max_{i=1, \dots, n} L(\mathbf{w}, \phi(\mathbf{x}_i)). \quad (22)$$

We can see that the previous objective function $R(\mathbf{w}, \rho)$ consists of two main parts. The first part is the modified empirical expected risk. The second part can be regarded as Hurwicz criterion with optimism parameter $1 - \eta$, which is exploited in decision problems when we do not have information about states of nature. If A takes values from 0 to 1, then the objective function is a convex combination of the Hurwicz criterion and the expected utility under condition that probabilities of states of nature are identical $1/n$. Moreover, $R(\mathbf{w}, \rho)$ can be also regarded as the objective function with respect to the Hodges-Lehmann criteria [30] by $\eta = 0$. This means that we are not sure that the probability of every point is $1/n$, and this disbelief is compensated by a more guaranteed Hurwicz approach with some coefficient depending on the function $\mu(\varepsilon)$. It is interesting that the fuzzy classification problem has transformed to the standard decision problem comprising several decision criteria.

Before solving the optimization problem for computing parameters $\mathbf{w}, \rho$, we have to define the function $\mu(\varepsilon)$ which determines the membership function of fuzzy expected losses. As we have pointed out, the simplest way is to assume that $\mu = 1 - \varepsilon$. Then we get

$$A = 1 - \int_0^1 \mu d\mu = 0.5. \quad (23)$$

However, this way does not take into account the possible dependence of the fuzzy set (its fuzziness) on the number of points in the training set. We assume that the fuzziness has to increase with decrease of n .

Let us consider the meaning of A in the objective function (22). If $A = 1$, then the optimal parameters are defined only by two single points maximizing and minimizing $L(\mathbf{w}, \phi(\mathbf{x}_i))$. This is an extreme case when we suppose that the empirical probability distribution is totally wrong. This takes place when we have a small number of points in the training set. Another extreme case is $A = 0$. This case corresponds to the standard approach in classification based on the empirical expected risk. It was shown by Vapnik [6] that the second case can be applied by large values of observations. Hence, we can state that $A = 0$ by $n \rightarrow \infty$ and $A \rightarrow 1$ by small values of n .

One of the possible functions satisfying the above conditions for A is $\mu = 1 - (n(1 - \varepsilon))^n$ or $\varepsilon = 1 - (1/n)\mu^{1/n}$. Hence

$$A = \frac{1}{n} \int_0^1 \mu^{1/n} d\mu = \frac{1}{n+1}. \quad (24)$$

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

4. The SVM Approach

In order to use the SVM approach to the fuzzy classification problem, we use the hinge loss function (10). Let us add the standard Tikhonov regularization term $(1/2)\langle \mathbf{w}, \mathbf{w} \rangle$ (this is the most popular penalty or smoothness term) [31] to the objective function (22). 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 [32] for a detailed analysis of regularization methods). Moreover, we introduce the following optimization variables:

$$\begin{aligned}\xi_i &= \max\{0, \rho - \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle\}, \\ G &= \max_{i=1, \dots, n} \xi_i.\end{aligned}\quad (25)$$

This leads to the quadratic programming problem

$$\begin{aligned}R(\mathbf{w}_{\text{opt}}, \rho_{\text{opt}}) &= \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle + \frac{1-A}{n} \sum_{i=1}^n \xi_i \\ &+ A\eta \min_{i=1, \dots, n} \xi_i \\ &+ A(1-\eta)G - \rho\nu \longrightarrow \min_{\mathbf{w}, \rho, G, \xi_i},\end{aligned}\quad (26)$$

subject to

$$\begin{aligned}\xi_i &\geq \rho - \langle \mathbf{w}, \phi(\mathbf{x}) \rangle, \quad \xi_i \geq 0, \quad i = 1, \dots, n, \\ G &\geq \xi_i, \quad i = 1, \dots, n.\end{aligned}\quad (27)$$

If we suppose that $\min_{i=1, \dots, n} \xi_i = \xi_k$, then the objective function can be decomposed to n objective functions of the form

$$\begin{aligned}R_k(\mathbf{w}_{\text{opt}}^{(k)}, \rho_{\text{opt}}^{(k)}) &= \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle \\ &+ \frac{1-A}{n} \sum_{i=1}^n \xi_i \\ &+ A\eta \xi_k + A(1-\eta)G - \rho\nu.\end{aligned}\quad (28)$$

The optimal parameters $\mathbf{w}_{\text{opt}}$ and ρ_{opt} correspond to the smallest value of R_k , $k = 1, \dots, n$.

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. Moreover, if we minimize the primary objective function, the dual objective function denoting L_k has to be maximized. The Lagrangian is

$$\begin{aligned}L_k &= \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle + \frac{1-A}{n} \sum_{i=1}^n \xi_i \\ &+ A\eta \xi_k + A(1-\eta)G - \rho\nu \\ &- \sum_{i=1}^n \zeta_i \xi_i - \sum_{i=1}^n \varphi_i (\xi_i - \rho + \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle) \\ &- \sum_{i=1}^n \psi_i (G - \xi_i).\end{aligned}\quad (29)$$

Here $\zeta_i, \psi_i, \varphi_i$, $i = 1, \dots, n$ are Lagrange multipliers. Hence, the dual variables have to satisfy positivity constraints $\zeta_i \geq 0$,

$\varphi_i \geq 0, \psi_i \geq 0$ for all $i = 1, \dots, n$. The saddle point can be found by setting the derivatives equal to zero:

$$\frac{\partial L}{\partial \rho} = -\nu + \sum_{i=1}^n \varphi_i = 0, \quad (30)$$

$$\frac{\partial L}{\partial G} = A(1-\eta) - \sum_{i=1}^n \psi_i = 0, \quad (31)$$

$$\frac{\partial L}{\partial \xi_i} = \frac{1-A}{n} + \mathbf{1}_k(i)A\eta - \zeta_i - \varphi_i + \psi_i = 0, \quad (32)$$

$$\frac{\partial L}{\partial \mathbf{w}_j} = \mathbf{w}_j - \sum_{i=1}^n \varphi_i \phi(\mathbf{x}_j^{(i)}) = 0, \quad j = 1, \dots, m. \quad (33)$$

Here $\mathbf{1}_k(i)$ is the indicator function taking the value 1 if $i = k$. Using (30)–(33), we get the following dual optimization problem:

$$L_k = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j K(\mathbf{x}_i, \mathbf{x}_j), \quad (34)$$

subject to

$$0 \leq \varphi_i \leq \frac{1-A}{n} + \mathbf{1}_k(i)A\eta + \psi_i, \quad i = 1, \dots, n, \quad (35)$$

$$\begin{aligned}\sum_{i=1}^n \varphi_i &= \nu, \quad \sum_{i=1}^n \psi_i = A(1-\eta), \\ \psi_i &\geq 0, \quad i = 1, \dots, n.\end{aligned}\quad (36)$$

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

$$f(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^n \varphi_i K(\mathbf{x}_i, \mathbf{x}) - \rho. \quad (37)$$

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

$$\rho = \sum_{i=1}^n \varphi_i K(\mathbf{x}_i, \mathbf{x}_j). \quad (38)$$

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

$$\begin{aligned}\varphi_i (\xi_i - \rho + \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle) &= 0, \\ \psi_i (G - \xi_i) &= 0.\end{aligned}\quad (39)$$

It follows from the second condition that $G = \xi_i$ for a single value of $i = l$ such that $l = \arg \max_{i=1, \dots, n} \xi_i$. Here we assume that $\xi_1, \dots, \xi_n$ do not coincide. Therefore, $\eta_i = 0$ for all $i \neq l$. Returning to the constraints (35)–(36), we get $\psi_l = A(1-\eta)$ and

$$\begin{aligned}0 \leq \varphi_i &\leq \frac{1-A}{n} + A\eta \cdot \mathbf{1}_k(i) + A(1-\eta) \cdot \mathbf{1}_l(i), \\ i &= 1, \dots, n.\end{aligned}\quad (40)$$

FIGURE 1: The contours $f(\mathbf{x}, \mathbf{w}) = 0$ by different values of $\eta = 0, 0.5, 1$ and by $n = 10$.

It follows from the previous constraints that the optimization problem (34)–(36) can be decomposed into n problems:

$$L_{k,l} = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \varphi_i \varphi_j K(\mathbf{x}_i, \mathbf{x}_j), \quad (41)$$

subject to (40) and $\sum_{i=1}^n \varphi_i = \nu$.

The optimal values of φ_i , $i = 1, \dots, n$ correspond to the *smallest* value of objective function $L_{k,l}$, $l = 1, \dots, n$ and to the *largest* value of objective functions $L_{k,l}$, $k = 1, \dots, n$.

So, we have n^2 simple quadratic programming problems whose solution can be obtained by means of well-known methods and tools.

5. 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 measure of the performance of 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

$$\text{ACC} = \frac{1}{N} \sum_{i=1}^N (I(y_i \cdot f(\mathbf{x}_i, \mathbf{w}) \geq 0)), \quad (42)$$

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

FIGURE 2: The contours $f(\mathbf{x}, \mathbf{w}) = 0$ by different values of $\eta = 0, 0.5, 1$ and by $n = 20$.

We will denote the accuracy measure for the proposed model as $\text{ACC}_{\text{fuzzy}}$ and for the standard SVM as ACC_{st} .

All the experiments use a standard Gaussian radial basis function (GRBF) kernel with the kernel parameter $\sigma = 0.04$. Different values for the parameter σ have been tested, choosing those leading to the best results.

We consider the performance of our method with synthetic data having two features x_1 and x_2 . The training set consisting of two subsets is generated in accordance with the normal probability distributions such that $N_1 = (1 - \varepsilon_0)N$ examples (the first subset) are generated with mean values $\mathbf{m}_1 = (4, 4)$ and $N_2 = \varepsilon_0 N$ examples (the second subset) have mean values $\mathbf{m}_2 = (12, 12)$. The standard deviation is $s = 2$ for both subsets and both features. Here ε_0 is a portion of abnormal examples in training set. The parameters ν and ε_0 are 0.3. The parameter A is taken $1/(n + 1)$.

Figure 1 illustrates how the contours $f(\mathbf{x}, \mathbf{w}) = 0$ depend on the parameter η for the fuzzy model (thick curve) and for the standard SVM (dashed curve) by the very small number

of points in the training set ($n = 10$). The parameter η takes values 0, 0.5, 1. One can see from the pictures that the region bounded by the “fuzzy” contour decreases with η . The value $\eta = 0$ provides the cautious strategy of decision making. It can be seen from the first picture that the region bounded by the “fuzzy” contour is larger than the same region corresponding to the standard model. The regions almost coincide for optimistic strategy when $\eta = 1$ (see Figure 1(c)). Figures 2 and 3 illustrate similar dependencies by $n = 20$ and $n = 40$, respectively.

The accuracy measure for different conditions is shown in Table 1. It can be seen from the results given in the table that the accuracy measure of the fuzzy model is larger than the same measure of the standard SVM. Of course, this relation takes place when n is rather small, and the contamination of the normal data has a large (in examples 0.3) ratio.

As a further example we applied all analyzed models to the well-known “Iris” data set from the UCI Machine

FIGURE 3: The contours $f(\mathbf{x}, \mathbf{w}) = 0$ by different values of $\eta = 0, 0.5, 1$ and by $n = 40$.

TABLE 1: Accuracy measures for different n and ν .

n	$\eta = 0$		$\eta = 0.5$		$\eta = 1$	
	$\text{ACC}_{\text{fuzzy}}$	ACC_{st}	$\text{ACC}_{\text{fuzzy}}$	ACC_{st}	$\text{ACC}_{\text{fuzzy}}$	ACC_{st}
10	0.852	0.556	0.824	0.584	0.56	0.56
20	0.768	0.528	0.744	0.514	0.484	0.48
40	0.714	0.666	0.712	0.68	0.724	0.7

Learning Repository [33]. 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 examples from the Iris Setosa class are abnormal. 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 $\nu = 0.333$, $\sigma = 0.001$. It is investigated how

the accuracy measures depend on the amount n of training data. In particular, if we take $n = 50$, then $\text{ACC}_{\text{fuzzy}} = 0.834$, $\text{ACC}_{\text{st}} = 0.773$. If we take $n = 80$, then $\text{ACC}_{\text{fuzzy}} = 0.833$, $\text{ACC}_{\text{st}} = 0.866$. One can see that the fuzzy approach provides better numerical results in comparison with the standard approach when the number of examples n is rather small.

6. Conclusion

In this paper, a fuzzy one-class classification model has been proposed, which is based on applying the ε -contaminated model. The algorithm for computing the optimal parameters of classification is reduced to a finite number of standard SVM tasks with weighted data points, where the weights are assigned in accordance with a predefined rule derived from comparison of fuzzy expected risk measures. It is easy to be implemented with standard functions of the statistical software package R.

Experimental results with synthetic data and with the well-known “Iris” data set from the UCI Machine Learning Repository reported have illustrated that the proposed fuzzy model outperforms the standard approach by the small number of observations. Due to the proposed fuzzy model, we do not need to assign a certain value for the contamination parameter ε . However, we have to define the function $\mu(\varepsilon)$ and the parameter η .

It should be noted that the proposed model can easily be extended on the case of binary or multiclass classification. This is a direction for future work.

We have investigated only one fuzzy ranking index for comparison fuzzy numbers. However, there are many efficient indices whose application to the classification problems in the framework of the proposed model could give better accuracy of classification and better models. This is also another direction for future work.

Acknowledgment

The author would like to express his appreciation to the anonymous referees and the Editor of this Journal whose valuable comments have improved the paper.

References

- [1] Y. Hao, Z. Chi, D. Yan, and X. Yue, “An improved fuzzy support vector machine for credit rating,” in *Network and Parallel Computing*, K. Li, C. Jesshope, H. Jin, and J.-L. Gaudiot, Eds., vol. 4672 of *Lecture Notes in Computer Science*, pp. 495–505, Springer, Berlin, Germany, 2007.
- [2] C. F. Lin and S. D. Wang, “Fuzzy support vector machines,” *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 464–471, 2002.
- [3] C. F. Lin and S. D. Wang, “Training algorithms for fuzzy support vector machines with noisy data,” *Pattern Recognition Letters*, vol. 25, no. 14, pp. 1647–1656, 2004.
- [4] C. F. Lin and S. D. Wang, “Fuzzy support vector machines with automatic membership setting,” in *Support Vector Machines: Theory and Applications*, L. Wang, Ed., vol. 177 of *Studies in Fuzziness and Soft Computing*, pp. 629–629, Springer, Berlin, Germany, 2005.
- [5] W. M. Tang, “Fuzzy SVM with a new fuzzy membership function to solve the two-class problems,” *Neural Processing Letters*, vol. 34, pp. 209–219, 2011.
- [6] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.
- [7] P. Y. Hao, “Fuzzy one-class support vector machines,” *Fuzzy Sets and Systems*, vol. 159, no. 18, pp. 2317–2336, 2008.
- [8] X. Jiang, Z. Yi, and J. C. Lv, “Fuzzy SVM with a new fuzzy membership function,” *Neural Computing and Applications*, vol. 15, no. 3–4, pp. 268–276, 2006.
- [9] R. Burduk, “Probability error in global optimal hierarchical classifier with intuitionistic fuzzy observations,” in *Hybrid Artificial Intelligence Systems*, E. Corchado, X. Wu, E. Oja, A. Herrero, and B. Barua, Eds., vol. 5572 of *Lecture Notes in Computer Science*, pp. 533–540, Springer, Berlin, Germany, 2009.
- [10] R. Burduk, “Classification error in Bayes multistage recognition task with fuzzy observations,” *Pattern Analysis and Applications*, vol. 13, no. 1, pp. 85–91, 2010.
- [11] L. V. Utkin and F. P. A. Coolen, “Interval-valued regression and classification models in the framework of machine learning,” in *Proceedings of the 7th International Symposium on Imprecise Probability: Theories and Applications (ISIPTA '11)*, F. Coolen, G. de Cooman, Th. Fetz, and M. Oberguggenberger, Eds., pp. 371–380, Innsbruck, Austria, 2011.
- [12] T. Wilk and M. Wozniak, “Soft computing methods applied to combination of one-class classifiers,” *Neurocomputing*, vol. 75, no. 1, pp. 185–193, 2012.
- [13] P. J. Huber, *Robust Statistics*, John Wiley & Sons, New York, NY, USA, 1981.
- [14] C. Campbell and K. P. Bennett, “A linear programming approach to novelty detection,” in *Advances in Neural Information Processing Systems*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds., vol. 13, pp. 395–401, MIT Press, 2001.
- [15] V. Cherkassky and F. M. Mulier, *Learning From Data: Concepts, Theory, and Methods*, Wiley-IEEE Press, 2007.
- [16] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, “Support vector method for novelty detection,” in *Advances in Neural Information Processing Systems*, pp. 526–532, 2000.
- [17] D. M. J. Tax and R. P. W. Duin, “Support vector domain description,” *Pattern Recognition Letters*, vol. 20, no. 11–13, pp. 1191–1199, 1999.
- [18] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, “Estimating the support of a high-dimensional distribution,” *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.
- [19] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, Mass, USA, 2002.
- [20] H. Xu, C. Caramanis, and S. Mannor, “Robustness and regularization of support vector machines,” *Journal of Machine Learning Research*, vol. 10, pp. 1485–1510, 2009.
- [21] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski, *Robust Optimization*, Princeton University Press, Princeton, NJ, USA, 2009.
- [22] J. Bi and T. Zhang, “Support vector classification with input data uncertainty,” in *Advances in Neural Information Processing Systems*, L. K. Saul, Y. Weiss, and L. Bottou, Eds., vol. 17, pp. 161–168, MIT Press, Cambridge, Mass, USA, 2004.
- [23] F. Provost and T. Fawcett, “Robust classification for imprecise environments,” *Machine Learning*, vol. 42, no. 3, pp. 203–231, 2001.
- [24] 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)*, pp. 536–542, MIT Press, Boston, Mass, USA, July 2006.
- [25] G. R. G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. I. Jordan, “A robust minimax approach to classification,” *Journal of Machine Learning Research*, vol. 3, no. 3, pp. 555–582, 2003.
- [26] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer, New York, NY, USA, 1985.
- [27] L. V. Utkin and Y. A. Zhuk, “Fuzzy decision making using the imprecise Dirichlet model,” *International Journal of Mathematics in Operational Research*, vol. 5, no. 1, pp. 74–90, 2013.
- [28] L. M. Campos and A. Gonzalez, “A subjective approach for ranking fuzzy numbers,” *Fuzzy Sets and Systems*, vol. 29, no. 2, pp. 145–153, 1989.
- [29] J. Schubert, “On ρ in a decision-theoretic apparatus of Dempster-Shafer theory,” *International Journal of Approximate Reasoning*, vol. 13, no. 3, pp. 185–200, 1995.

- [30] J. L. Hodges and E. Lehmann, “The use of previous experience in reaching statistical decisions,” *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 396–407, 1952.
- [31] A. N. Tikhonov and V. Y. Arsenin, *Solution of Ill-Posed Problems*, W.H. Winston, Washington, DC, USA, 1977.
- [32] T. Evgeniou, T. Poggio, M. Pontil, and A. Verri, “Regularization and statistical learning theory for data analysis,” *Computational Statistics and Data Analysis*, vol. 38, no. 4, pp. 421–432, 2002.
- [33] A. Frank and A. Asuncion, *UCI Machine Learning Repository*, 2010.