A machine learning algorithm for classification under extremely scarce information

Lev V. Utkin*

Department of Industrial Control and Automation,
St.Petersburg State Forest Technical Academy,
Institutsky per. 5, 194021 St.Petersburg, Russia
Fax: +7 812 6709358 E-mail: lev.utkin@gmail.com
*Corresponding author

Yulia A. Zhuk

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

Abstract: When it is difficult to get learning data during the training time, we have to classify objects by having extremely small information about their feature. It is assumed in the paper that only some average or mean value of every feature and the lower and upper bounds of a set of its values are known. The main idea for constructing new classification models taking into account this information is to form a set of probability distributions bounded by some lower and upper probability distribution functions (a p-box). A discriminant function is derived in order to maximize the risk measure over the set of distributions and to minimize it over a set of classification parameters. The algorithm for classification is reduced to a parametric linear programming problem.

Keywords: imprecise probabilities; lower and upper probability distributions; learning problem; risk; Bayesian inference; regression.

Reference to this paper should be made as follows: Utkin, L.V. and Zhuk, Y.A. (xxxx) 'A machine learning algorithm for classification under extremely scarce information', *Int. J. Data Analysis Techniques and Strategies*, 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 Industrial Control and Automation, Saint-Petersburg State Forest Technical Academy. 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. He is on the editorial board of the International Journal of Approximate Reasoning.

Yulia A. Zhuk is currently an assistant at the Department of Computer Science, Saint-Petersburg State Forest Technical Academy. 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, image coding.

1 Introduction

A main goal of the statistical machine learning is to predict an unobserved output value y based on an observed input vector $\mathbf{x}$. This requires us to estimate a predictor f from training data or a set of example pairs of $(\mathbf{x}, y)$. A special very important problem of the statistical machine learning is the binary classification problem which can be regarded as a task of classifying some objects into two classes (group) in accordance with their properties or features. In other words, we have to classify each pattern $\mathbf{x}$ into one of the classes by means of a discriminant function f. A huge number of methods have been proposed for solving the machine learning problems last decades. However, many of them are based on restrictive assumptions, for instance, the large amount of training data, the known type of the noise probability distribution, point-valued observations, etc.

At the same time, real applications can not satisfy all these assumptions or their part due to several reasons. Therefore, a huge number of methods have been developed in order to relax at least a part of the restrictive assumptions. In particular, classification problems with interval-valued observations were studied by many authors (Angulo et al., 2008; Carrizosa et al., 2007; Ishibuchi et al., 1990; Nivlet et al., 2001). A detailed analysis of different models handling missing data in classification is also proposed by (Pelckmans et al., 2005). The classification problem by missing data in the framework of imprecise probability theory was studied also by (de Cooman and Zaffalon, 2004). Similar methods were proposed also by (Corani and Zaffalon, 2008). A series of works devoted to combining partially supervised information in classification tasks were proposed by several authors (Come et al., 2009; Denoeux and Smets, 2006; Masson and Denoeux, 2004). The above is a very small part of the papers devoted to imprecision and partially observed data in classification.

Sometimes, we have to classify objects by having extremely small information about their feature. This might take place in various cases. One of the most frequently occurring case is when we do not observe the objects, but then we have to classify them on the basis of some limited available information. Suppose that we manufactory reinforced concrete beams whose quality and strength depend on a number of parameters such as the weight of reinforcement bars, concrete materials, etc. If we have not observed or measured the parameters before, it is difficult to reject new beams or to classify them into two classes: defective (rejected) or of high quality, because we do not have the learning set of beams with the measured parameters. However, if we know, for instance, how much steel have been used up by manufacturing N beams, then we are able to evaluate the average wight of steel in a beam.

It should be noted that expert judgments are a very important part of information, which also should be exploited. Very often, it is easy for experts to provide judgments about some average value of a feature for every class of objects because this information is the most simple and understandable (Lad and Kulkarni, 2010). However, it is extremely difficult to give, for instance, the variance and other statistical measures of random variables of interest. Therefore, we restrict ourselves by this information and try to solve the classification problem under this restriction. At the same time, it is easy to prove that only mean values can not allow us to construct a classification model if features are unbounded. Therefore, the "smallest" assumption which has to be added for getting non-trivial estimates is finite boundary values of every feature.

In the paper, it is assumed that we know some average or mean value of every feature. Moreover, we assume that the values of every feature are restricted by some lower and upper bounds. In spite of the scarcity of the available information, it can be used for constructing classifiers.

It is interesting to point out that the simplest classification algorithm considered by (Scholkopf and Smola, 2002) in detail uses vectors of means υ and ω of the two classes in feature space, respectively. The algorithm is based on analyzing the distances between a vector $\mathbf{x}$ and two vectors of mean values of features. The smallest distance determines the class of $\mathbf{x}$. It has been noted by (Scholkopf and Smola, 2002) that the proposed decision is the best we can do if we have no prior information about the probabilities of the two classes.

However, the above algorithm does not use the additional information about bounds of the features, which is often available in real applications. For instance, by returning to the above considered example, we can assert that the wight of steel in a beam is positive (larger than 0) and it is less than the wight of the beam. This implies that we can use the additional information in order to improve the classification accuracy. (Huhn and Hullermeier, 2008) illustrated how, for example, the additional information in the form of ordinal structures can improve the classification performances. The main idea for constructing new models taking into account this information is the following. The mean values of features and their boundary values produce a set of probability distributions bounded by some lower and upper cumulative distribution functions (CDFs). This way leads to constructing the so-called p-boxes (Destercke et al., 2008) from data. At that the bounds for the set of probability distributions depend on unknown parameters of a discriminant function which has to be found in order to solve the classification problem. It should be noted that the considered set of distributions is not the set of parametric distributions having the same parametric form as the bounding distributions, but it is the set of all possible distributions restricted by the lower and upper bounds. This is an important feature of the proposed approach in this paper. Two probability distributions are selected from the p-box in order to make a pessimistic decision, which maximize the risk function as a measure of the classification error. In other words, the well-known minimax strategy is applied for solving the classification problem, which appears as an insurance against the worst case (Robert, 1994). The similar idea applied to regression models has been considered by (Utkin, 2010), by (Utkin and Coolen, 2010).

The second idea is the following. By knowing the mean values of feature and assuming that the discriminant function is linear, we can easily find the mean value of the function f and its bounds as functions of the unknown parameters w. Therefore, the second idea is that we construct the p-boxes not for every feature $\mathbf{x}^{(i)}$, but for the function f. As a result the bounds of the p-boxes and the bounds of values of f become functions of the classification parameters.

The parameters w of the discriminant function are finally computed by minimizing the upper risk measure through solving a linear programming problem.

The paper is organized as follows. A formal statement of the classification problem is given in Section 2. In Section 3, it is considered how to extend the standard classification problem under condition of a set of probability distributions. The minimax strategy for decision making about parameters of classification is studied in the same section. The question of constructing p-boxes from the information about mean value of features and their bounds is considered in Section 4. A method for computing optimal classification parameters by solving a parametric linear programming problem is provided in Section 5. Walley's imprecise Dirichlet model (Walley, 1996) is proposed for computing the prior probabilities of classes in Section 6. A numerical example is considered in Section 7.

2 The standard classification problem

2.1 A classification problem statement

The binary-classification problem can be formulated as follows. There are predictor-response data with a binary response y representing the observation of classes $y = -1$ and $y = 1$. The binary-classification problem is to estimate a region in predictor space in which class 1 is observed with the greatest possible majority. Suppose we are given empirical data

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n) \in \mathbb{R}^n \times \{-1, +1\}.$$

Here $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ is some nonempty set of the patterns or examples; $y_1, ..., y_n$ are labels or outputs. Below we divide the set of training data into two subsets corresponding to $y_i = -1$ with n_{-1} elements and $y_i = 1$ with n_1 elements, respectively, $n_{-1} + n_1 = n$.

Classification problem is usually characterized by an unknown CDF $F_0(\mathbf{x}, y)$ on $\mathbb{R}^n \times \{-1, +1\}$ defined by the training set or examples $\mathbf{x}_i$ and their corresponding class labels y_i.

One of the possible approaches for solving the problem is the discriminant function approach which uses a real valued function $f(\mathbf{x}, w)$ called the discriminant function whose sign denoted $\mathrm{sgn}(f(\mathbf{x}, w))$ or briefly $\mathrm{sgn}(f)$ determines the class label prediction. The discriminant function $f(\mathbf{x})$ may be parametrized with some parameters $w = (w_0, w_1, ..., w_m)$, $w \in \Lambda$, that are determined from the training examples by means of a learning algorithm. So, we will write the discriminant function as $f(\mathbf{x}, w)$. Moreover, we assume that the form of the discriminant function is known and it is linear, i.e.,

$$f(\mathbf{x}, w) = w_0 + \sum_{i=1}^{m} w_i x_i, \ x_i \in \mathbf{x}.$$

Introduce also the notation $x_i^{(k)}$ for the i-th element of the vector $\mathbf{x}_k$.

2.2 The classification problem and the mean values of features

We briefly consider the simplest classification algorithm mentioned in Section 1, which uses only mean values of features for classifying. According to this algorithm, the vectors of means $\upsilon = (\upsilon_1, ..., \upsilon_m)$ and $\omega = (\omega_1, ..., \omega_m)$ corresponding to two classes are computed as

$$\upsilon = \frac{1}{n_{-1}} \sum_{i:y_i=-1} \mathbf{x}_i, \ \omega = \frac{1}{n_1} \sum_{i:y_i=+1} \mathbf{x}_i.$$

A new point $\mathbf{x}$ is assigned to the class whose mean is closest. If the half way between υ and ω lies the point $\mathbf{c} = (\upsilon + \omega)/2$, then the class of $\mathbf{x}$ is determined as follows:

$$\begin{aligned} y &= \operatorname{sgn} \langle (\mathbf{x} - \mathbf{c}), w \rangle \\ &= \operatorname{sgn} \left(\langle \mathbf{x}, \omega \rangle - \langle \mathbf{x}, \upsilon \rangle + w_0 \right). \end{aligned}$$

Here the canonical dot product notation $\langle \cdot, \cdot \rangle$ is used for short. The parameter w_0 is computed as

$$w_0 = \frac{1}{2} \left(\langle \upsilon, \upsilon \rangle - \langle \omega, \omega \rangle \right).$$

If the class means have the same distance to the origin, then w_0 will vanish.

So, by having only the mean values of features, we can construct the discriminant function.

3 The classification problem under a set of probability distributions

First, we have to note that the loss function depends on the discriminant function f. By replacing the CDF $F_0(\mathbf{x}, y)$ by the CDF $F(f, y)$, we can rewrite the risk measure as follows:

$$R(w) = \int_{\mathbb{R} \times \{-1,1\}} \mathbf{1}\{\operatorname{sgn}(f) \neq y\} \mathrm{d}F(f, y) \mathrm{d}y.$$

Moreover, we represent the joint probability as $F(f, y) = F(f \mid y) \cdot P(y)$. Here $P(y)$ is the prior probability that an example $\mathbf{x}$ belongs to the class y.

Let us rewrite the risk measure taking into account two values of y

$$R(w) = P(-1)R_{-1}(w) + P(1)R_{+1}(w).$$

Here

$$R_{-1}(w) = \int_{\mathbb{R}} \mathbf{1}\{\operatorname{sgn}(f) \neq -1\} \mathrm{d}F(f \mid -1),$$

$$R_{+1}(w) = \int_{\mathbb{R}} \mathbf{1}\{\text{sgn}(f) \neq +1\} \mathrm{d}F(f \mid +1).$$

Suppose that the distributions F are unknown. However, we assume that some lower and upper bounds for a set $\mathcal{F}(y)$ of the CDFs $F(f \mid y)$ are known accurate within w, i.e., we know the CDFs as functions of w, but w is unknown. The corresponding lower and upper CDFs are $\underline{F}(f \mid y)$ and $\overline{F}(f \mid y)$, respectively. We can write

$$\mathcal{F}(y) = \{F(f \mid y) \mid \underline{F}(f \mid y) \leq F(f \mid y) \leq \overline{F}(f \mid y), \ \forall f \in \mathbb{R}\}.$$

In other words, there is an unknown precise "true" CDF $F(f \mid y) \in \mathcal{F}(y)$ for every $y \in \{-1, +1\}$, but we do not know it and we only know that it belongs to the set $\mathcal{F}(y)$. It has been mentioned that the set $\mathcal{F}(y)$ is not the set of parametric distributions having the same parametric form as the bounding distributions, but it is the set of all possible distributions restricted by the lower and upper bounds.

The bounds of $\mathcal{F}(y)$ depend on the vector w. Consequently, the set $\mathcal{F}(y)$ is also depends on w. However, we do not indicate this fact explicitly for short.

One of the possible strategies to derive an estimator is the minimax (pessimistic) strategy. According to the minimax strategy, we select a probability distribution from the set $\mathcal{F}(-1)$ and a probability distribution from the set $\mathcal{F}(+1)$ such that the risk measures $R_{-1}(w)$ and $R_{+1}(w)$ achieve their maximum for every fixed w. It should be noted that the "optimal" probability distributions may be different for different values of parameters w. This implies that the corresponding "optimal" probability distributions depend on w. The minimax strategy can be explained in a simple way. We do not know the precise probability distribution F and every distribution from $\mathcal{F}(y)$ can be selected. Therefore, we should take the "worst" distribution providing the largest value of the risk measure. The minimax criterion appears as an insurance against the worst case because it aims at minimizing the expected loss in the least favorable case (Robert, 1994).

Since the sets $\mathcal{F}(-1)$ and $\mathcal{F}(+1)$ are obtained independently for $y = -1$ and $y = 1$, respectively, then

$$\overline{R}(w) = \max_{F(f \mid y) \in \mathcal{F}(-1) \times \mathcal{F}(+1)} R(w)$$
$$= P(-1) \max_{F(f \mid -1) \in \mathcal{F}(-1)} R_{-1}(w) + P(1) \max_{F(f \mid +1) \in \mathcal{F}(+1)} R_{+1}(w).$$

Here we assume that the prior probabilities are precise. However, it will be shown below how to relax this strong assumption and to use the imprecise probabilities.

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

$$\overline{R}(w_{\text{opt}}) = \min_{w \in \Lambda} \overline{R}(w).$$

Let us consider in detail the first problem $\max_{F(f \mid -1) \in \mathcal{F}(-1)} R_{-1}(w)$ and the loss function $L(f, -1) = \mathbf{1}\{\text{sgn}(f) \neq -1\}$. It is easy to prove that the loss function $L(f, -1)$ is increasing with f. Indeed, if $f < 0$, then $\text{sgn}(f) = -1$ and $L(f, -1) =$

0. If $f \geq 0$, then $\mathrm{sgn}(f) = 1$ and $L(f, -1) = 1$. According to (Walley, 1991), this implies that the upper bound for $R_{-1}(w)$, i.e., the maximum of $R_{-1}(w)$ over all distributions from $\mathcal{F}(-1)$ is achieved at the distribution $\underline{F}(f \mid -1)$. Hence, there holds

$$\overline{R}_{-1}(w) = \int_{\mathbb{R}} \mathbf{1}\{\mathrm{sgn}(f) \neq -1\} \mathrm{d}\underline{F}(f \mid -1) = 1 - \underline{F}(0 \mid -1).$$

In the same way, we can consider the second problem $\max_{F(f \mid +1) \in \mathcal{F}(+1)} R_{+1}(w)$. The function $L(f, +1)$ in this case is decreasing. Therefore, the upper bound for $R_{+1}(w)$ is achieved at the distribution $\overline{F}(f \mid 1)$. This implies that

$$\overline{R}_{+1}(w) = \int_{\mathbb{R}} \mathbf{1}\{\mathrm{sgn}(f) \neq +1\} \mathrm{d}\overline{F}(f \mid +1) = \overline{F}(0 \mid 1).$$

Finally, we get the upper bound for the risk measure $R(w)$, which is of the form:

$$\overline{R}(w) = P(1)\overline{F}(0 \mid 1) + P(-1) - P(-1)\underline{F}(0 \mid -1). \tag{1}$$

The optimization problem for computing the optimal values of parameters w can be written as

$$\begin{aligned}\overline{R}(w_{\mathrm{opt}}) &= \min_{w \in \Lambda} \left(P(1)\overline{F}(f \mid 1) + P(-1) - P(-1)\underline{F}(f \mid -1) \right) \\ &= \max_{w \in \Lambda} \left(P(-1)\underline{F}(f \mid -1) - P(1)\overline{F}(f \mid 1) - P(-1) \right) = \max_{w \in \Lambda} Q(w),\end{aligned}$$

subject to $f(\mathbf{x}, w) = 0$.

Here we introduce the measure $Q(w) = -\overline{R}(w)$ for simplicity.

Now we have two tasks. First, we have to define CDFs $\overline{F}(f = 0 \mid y = 1)$ and $\underline{F}(f = 0 \mid y = -1)$ from the available information. Second, we have to define the prior probabilities of classes $P(-1)$ and $P(1)$.

4 P-boxes constructing from the mean and bounds of a random variable

Assume that we are only given information about bounds $X \in [a, b]$, and the mean M of a random variable X such that $a < M < b$. (Ferson et al., 2001) considered the "best" bounds for the CDF of X. To determine the upper bound $\overline{F}(x)$ for the CDF we analyze the case $x < M$. The maximal CDF at x, that is the maximal probability measure on $\{X \leq x\}$ must satisfy the condition on the average $x \cdot \overline{F}(x) + b\left(1 - \overline{F}(x)\right) = M$ to have at least the mean M. We get

$$\overline{F}(x) = \begin{cases} 0, & x < a, \\ \min\left(1, \frac{b-M}{b-x}\right), & a \leq x < b, \\ 1, & x \geq b. \end{cases}$$

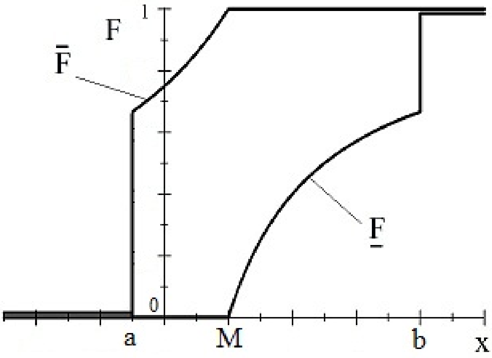

Figure 1 The lower and upper probability distributions

It should be noted that the above expression for the upper bound $\overline{F}(x)$ can be obtained by the natural extension (Kuznetsov, 1991; Walley, 1991) which can be represented as the following linear programming problem:

$$\overline{F}(x) = \min_{c,d} \left(c + d \cdot M \right),$$

subject to $c, d \in \mathbb{R}$, $c + d \cdot z \geq \mathbf{1}\{z \leq x\}$, $\forall z \in [a, b]$.

The lower bound for the CDFs can be obtained in the same way by solving the following programming problem:

$$\underline{F}(x) = \max_{c,d} \left(c + d \cdot M \right),$$

subject to $c, d \in \mathbb{R}$, $c + d \cdot z \leq \mathbf{1}\{z \leq x\}$, $\forall z \in [a, b]$.

Hence, there holds

$$\underline{F}(x) = \begin{cases} 0, & x < a, \\ \max\left(0, \frac{x-M}{x-a}\right), & a \leq x < b, \\ 1, & x \geq b. \end{cases}$$

The lower and upper CDFs are shown in Fig. 1, where $M = 2$, $a = -1$, $b = 8$. The resulting bounds are optimal in the sense that they could not be any tighter without excluding at least some portion of a CDF from a distribution satisfying the specified constraints. However, this does not mean that any distribution whose CDF is inscribed within this bounded probability region would satisfy the constraints (Ferson et al., 2001). The obtained set is more rich and produces the p-box. This leads to a more conservative and cautious solution of the classification problem.

5 The risk measure and p-boxes

5.1 The upper risk measure as a function of classification parameters

Our goal is to find the parameters $w_0, ..., w_m$ of the discriminant function minimizing the upper bound for the risk measure $\overline{R}(w)$. In order to solve this problem, we need to represent the upper risk measure as a function of parameters w, and then we have to minimize the function.

We suppose that the vector of input variables (features) $\mathbf{x} = (x_1, ..., x_m)$ is bounded, i.e., its every element fulfils the condition

$$\underline{x}_k \leq x_k \leq \overline{x}_k,\ k = 1, ..., m. \tag{2}$$

Here $\underline{x}_k$ and $\overline{x}_k$ are lower and upper bounds for the k-th elements of the vector $\mathbf{x}$.

We have denoted in Section 2 the vectors of mean values of m features by $\upsilon = (\upsilon_1, ..., \upsilon_m)$ and $\omega = (\omega_1, ..., \omega_m)$ corresponding to the classes with $y = -1$ and $y = 1$, respectively. In particular, if we knew the values of features, then the mean values of the k-th feature could be written as

$$\upsilon_k = \frac{1}{n_-} \sum_{i:y_i=-1} x_k^{(i)},\quad \omega_k = \frac{1}{n_+} \sum_{i:y_i=1} x_k^{(i)}.$$

We also denote the mean values of the function f corresponding to every class y by M_y. It is assumed that $\upsilon_0 = 1$ and $\omega_0 = 1$ for writing below constraints for the first moments M_{-1} and M_1 in a more short form.

First, we represent the upper risk measure as a function of parameters w. By using the results provided in Sections 3 and 4 and assuming that features of every class are characterized by their mean values and by bounds of their values, we can substitute the corresponding lower and upper CDFs for each class into (1). In other words, we substitute the functions

$$\underline{F}(f \mid -1) = \begin{cases} 0, & f < a_{-1}, \\ \max\left(0, \frac{f - M_{-1}}{f - a_{-1}}\right), & a_{-1} \leq f < b_{-1}, \\ 1, & f \geq b_{-1}, \end{cases} \tag{3}$$

$$\overline{F}(f \mid 1) = \begin{cases} 0, & f < a_1, \\ \min\left(1, \frac{b_1 - M_1}{b_1 - f}\right), & a_1 \leq f < b_1, \\ 1, & f \geq b_1, \end{cases} \tag{4}$$

into the expression for $\overline{R}(w)$.

It follows from the condition (1) and the functions (3)-(4) that the upper bound for the risk measure is

$$\overline{R}(w) = -P(-1) \cdot \max\left(0, \frac{f - M_{-1}}{f - a_{-1}}\right) + P(1) \cdot \min\left(1, \frac{b_1 - M_1}{b_1 - f}\right) - P(-1).(5)$$

The next question is to define the parameters of the lower and upper CDFs M_{-1}, M_1, a_{-1}, a_1, b_{-1}, b_1. The fact is that we do not know these parameters and we know only means and the corresponding bounds of the features, but not of the function f. Therefore, it is necessary to express the above parameters through the parameters of the features. This is a crucial point because the above parameters are functions of parameters $w_0, ..., w_m$ of the discriminant function.

Due to the linearity of the discriminant function f, the first moments M_{-1} and M_1 are

$$M_{-1} = w_0 + \sum_{k=1}^{m} w_k \upsilon_k, \ M_1 = w_0 + \sum_{k=1}^{m} w_k \omega_k.$$

It is also simply to prove that $a_{-1} = a_1 = a$ and $b_{-1} = b_1 = b$.

Let us consider different cases of the relationship between parameters if the considered random variables.

Suppose that $a < M_{-1} < M_1 < b$. Then it is easy to show that the upper bound for the risk measure is

$$\overline{R}(w) = \begin{cases} P(1)\frac{b-M_1}{b-f} + P(-1), & a \leq f < M_{-1}, \\ -P(-1)\frac{f-M_{-1}}{f-a} + P(1)\frac{b-M_1}{b-f} + P(-1), & M_{-1} \leq f < M_1, \\ -P(-1)\frac{f-M_{-1}}{f-a} + P(1) + P(-1), & M_1 \leq f < b. \end{cases}$$

By taking into account the necessary condition for the discriminant function $f = 0$ (see Section 3), we write

$$\overline{R}(w) = \begin{cases} 1 - P(1)M_1/b, & a \leq 0 < M_{-1}, \\ 1 - P(-1)M_{-1}/a - P(1)M_1/b, & M_{-1} \leq 0 < M_1, \\ 1 - P(-1)M_{-1}/a, & M_1 \leq 0 < b. \end{cases}$$

We will use the measure $Q(w)$ in place of $\overline{R}(w)$ below, which has to be maximized over w.

5.2 Optimization problems for computing the optimal classification parameters

Let us consider three cases of the relationship between M_{-1}, M_1, 0. Every relationship produces a set of parameters w.

Case 1. $M_{-1} \leq 0 < M_1$. In order to find the optimal values of parameters w, we define bounding conditions, i.e., we define bounds a and b. Note that the last expression for the risk measure deals with the bounds a and b of f, but not input variables. Therefore, we can not write in an explicit form these bounds. At the same time, we can write

$$a_{-} = \min_{x_k} \left\{ w_0 + \sum_{k=1}^{m} w_k x_k \right\}$$

subject to (2).

This is the linear programming problem having m variables and $2m$ constraints. This implies that m constraints are equalities, i.e., we should search the problem solution at a combination of boundary points $\underline{x}_k$ and $\overline{x}_k$. Therefore, the bound a satisfies the following set of constraints:

$$a \leq w_0 + \sum_{k=1}^{m} w_k x_k^*, \quad \forall x_k^* \in \{\underline{x}_k, \overline{x}_k\}.$$

In other words, we have 2^m inequalities such that one of them is the equality. Note that the bound a is negative. Consequently, in order to make $1 - \overline{R}(w)$ as large as possible, the values of $|a|$ should be taken as small as possible. Hence, the value of a should be taken as large as possible, but its "growth" is restricted by the above linear constraints.

The same can be derived for the bound b which is also unknown, but it can be found from the linear programming problem

$$b = \max_{x_k} \left\{ w_0 + \sum_{k=1}^{m} w_k x_k \right\}$$

subject to (2).

Hence, the bound b fulfils the following set of constraints:

$$b \geq w_0 + \sum_{k=1}^{m} w_k x_k^*, \quad \forall x_k^* \in \{\underline{x}_k, \overline{x}_k\}.$$

Note that the bound b is positive. Therefore, in order to make $1 - \overline{R}(w)$ as large as possible, the value of b should be taken as small as possible, but its decrease is restricted by the above linear constraints.

Now the optimization problem for computing the optimal vector of parameters w can be written as follows:

$$Q_1(w) = \max_{w} \left(P(-1) \frac{\sum_{k=0}^{m} w_k \upsilon_k}{a} + P(1) \frac{\sum_{k=0}^{m} w_k \omega_k}{b} \right)$$

subject to

$$a \leq w_0 + \sum_{k=1}^{m} w_k x_k^*, \quad b \geq w_0 + \sum_{k=1}^{m} w_k x_k^*, \quad \forall x_k^* \in \{\underline{x}_k, \overline{x}_k\},$$

$$\sum_{k=0}^{m} w_k \upsilon_k \leq 0, \quad \sum_{k=0}^{m} w_k \omega_k \geq 0.$$

Here two final inequalities correspond to conditions $M_{-1} \leq 0$ and $M_1 \geq 0$.

Let us introduce the variables $q_k = w_k/a$, $k = 0, ..., m$, $q = (q_1, ..., q_m)$. Moreover, we introduce a parameter r such that $r = a/b$. Then we can rewrite the above optimization problem as follows:

$$\begin{aligned} Q_1(q, r) &= \max_{q} \left(P(-1) \sum_{k=0}^{m} q_k \upsilon_k + P(1) r \sum_{k=0}^{m} q_k \omega_k \right) \\ &= \max_{q} \sum_{k=0}^{m} q_k \left(P(-1) \upsilon_k + r P(1) \omega_k \right) \end{aligned}$$

subject to

$$1 \geq q_0 + \sum_{k=1}^{m} q_k x_k^*,\ 1 \geq r q_0 + r \sum_{k=1}^{m} q_k x_k^*,\ \forall x_k^* \in \{\underline{x}_k, \overline{x}_k\}, \tag{6}$$

$$\sum_{k=0}^{m} q_k \upsilon_k \geq 0,\ \ \sum_{k=0}^{m} q_k \omega_k \leq 0. \tag{7}$$

Since r is negative, then we can rewrite constraints (6) as follows:

$$\frac{1}{r} \leq q_0 + \sum_{k=1}^{m} q_k x_k^* \leq 1, \ \ \forall x_k^* \in \{\underline{x}_k, \overline{x}_k\}. \tag{8}$$

The above problem can be regarded as the linear programming problem with the parameter $r \leq 0$ which is the ratio of lower and upper bounds for f. By changing the parameter r and by solving the linear programming problem for every r, we can find the optimal parameters $q_0, ..., q_m$ corresponding to the largest value of $Q_1(q, r)$ for all r.

The objective function can be written also as

$$Q_1(q, r) = P(-1) M_{-1}/a + r \cdot P(1) M_1/a.$$

Since $M_1/a \leq 0$, then $Q_1(q, r)$ increases as r decreases. This implies that we have to decrease r. However, by decreasing r, we increase $1/r$ and the feasible region is reduced. Therefore, there exists a compromised point r providing the maximum of the objective function.

Note that it is impossible to find the optimal values $w_0, ..., w_m$ from the optimal values $q_0, ..., q_m$ because the variable a can not be found. However, we do not need to search for the single optimal vector w. The classification problem requires to find a linear function $f(\mathbf{x}, w)$ such that its sign determines the class label prediction. But in case of the linear discriminant function, its sign does not change when we multiply the function by some positive number. This implies that all linear functions $f(\mathbf{x}, w)$ with parameters $w_k = q_k a$, $k = 0, ..., m$, $a > 0$, can be regarded as the discriminant functions. If we assume, for instance, $w_0 = 1$, then there holds $a = 1/q_0$, and we can compute every parameter $w_k = q_k/q_0$.

Case 2. $a \leq 0 < M_{-1}$. Denote the corresponding measure as $Q_2(w) = P(1) M_1/b$. Let us introduce the variables $q_k = w_k/b$, $k = 0, ..., m$. Then we get the linear programming problem

$$Q_2(q_{\text{opt}}) = P(1) \cdot \max_q \sum_{k=0}^{m} q_k \omega_k$$

subject to

$$1 \geq q_0 + \sum_{k=1}^{m} q_k x_k^*, \ \ x_k^* \in \{\underline{x}_k, \overline{x}_k\},\ \sum_{k=0}^{m} q_k \upsilon_k \leq 0.$$

Case 3. $M_1 \leq 0 < b$. Denote the corresponding measure as $Q_3(q) = P(-1)M_{-1}/a$. Let us introduce the variables $q_k = w_k/a$, $k = 0, ..., m$. Then we get the linear programming problem

$$Q_3(q_{\text{opt}}) = P(-1) \cdot \max_q \sum_{k=0}^{m} q_k \upsilon_k$$

subject to

$$1 \geq q_0 + \sum_{k=1}^{m} q_k x_k^*, \quad x_k^* \in \{\underline{x}_k, \overline{x}_k\}, \ \sum_{k=0}^{m} q_k \omega_k \leq 0.$$

Now the optimal vector q_{opt} is defined from the condition

$$q_{\text{opt}} = \arg\max \left\{ \max_q \max_r Q_1(q, r), \max_q Q_2(q), \max_q Q_3(q) \right\}.$$

6 Prior probabilities

One of the simplest ways for estimating the probabilities $P(y)$, $y = -1, 1$, is their direct computing by using the following expressions:

$$P(-1) = n_{-1}/n, \quad P(1) = n_{+1}/n = 1 - P(-1).$$

In particular, if we do not have information about n_{-1} and n_{+1}, then the prior probabilities of classes can be taken 0.5.

However, this way might lead to incorrect probabilities when the number of examples n is small. For making cautious decisions, we use Walley's imprecise Dirichlet model (Walley, 1996) which in case of two classes is called also the imprecise beta model. This model has been used in classification in several works, for instance, in (Corani and Zaffalon, 2008; Zaffalon, 2002). According to this model, the lower and upper bounds for the probability $P(y)$ can be written as follows:

$$\underline{P}(y) = \frac{n_y}{n+s}, \quad \overline{P}(y) = \frac{n_y + s}{n+s}, \quad y = -1, 1.$$

Here s is the hyperparameter of the Dirichlet distribution. According to the work (Walley, 1996), the hyperparameter s should be taken 1 or 2. The above expressions are obtained from the following probabilities by maximizing and minimizing the probabilities

$$P(-1) = \frac{\gamma s + n_{-1}}{n+s}, \quad P(1) = \frac{(1-\gamma)s + n_{+1}}{n+s}$$

over the set of $\gamma \in (0, 1)$.

By using the expressions for the prior probabilities, we write the optimization problem for the minimax strategy (Case 1)

$$Q(w_{\text{opt}}) = \max_{w\in\Lambda} \max_{0\leq\gamma\leq 1} \left((\gamma s + n_{-1})\, \underline{F}(0 \mid -1) - ((1-\gamma)s + n_{+1})\, \overline{F}(0 \mid 1) \right),$$

subject to $f(\mathbf{x}, w) = 0$.

Rewrite the objective function as follows:

$$\begin{aligned} Q(w) &= \gamma s \left(\underline{F}(0 \mid -1) + \overline{F}(0 \mid 1) \right) \\ &\quad + n_{-1}\underline{F}(0 \mid -1) - n_{+1}\overline{F}(0 \mid 1) - s\overline{F}(0 \mid 1). \end{aligned}$$

One can see from the above expression that the maximum of $Q(w)$ over the set of γ is achieved at point $\gamma = 1$ and it is

$$Q(w_{\text{opt}}) = \max_{w\in\Lambda} \left((s + n_{-1})\, \underline{F}(0 \mid -1) - n_{+1}\overline{F}(0 \mid 1) \right).$$

Finally, the linear programming problem for computing the optimal parameters w_{opt} has the objective function

$$Q(w, r) = \sum_{k=0}^{m} q_k \left((s + n_{-1})\, v_k + r n_{+1} \omega_k \right)$$

and constraints (7) and (8).

Similar expressions can be obtained for Case 2 and Case 3.

7 Numerical example

Let us consider a toy example in order to illustrate the programming problem statement. Suppose that we have to reject defective wooden beams. We have three features for classifying defective beams – the width x_1, the length x_2 of a crack, the brightness x_3 of the wood. The width and the length of cracks are restricted by the beam sizes. In particular, the width is in the interval from 0 to 0.01 meters ($\underline{x}_1 = 0,\ \overline{x}_1 = 0.01$), the length can be from 0 to 0.5 meters ($\underline{x}_2 = 0,\ \overline{x}_2 = 0.5$). The brightness is measured from 16 to 64 ($\underline{x}_3 = 16,\ \overline{x}_3 = 64$). Assume that prior probabilities of classes are 0.5.

The experts provide judgments concerning the mean values of defective beams ($y = 1$) and high-quality beams ($y = -1$):

$$v_1 = 0.002,\ v_2 = 0.1,\ v_3 = 38,\ \omega_1 = 0.005,\ \omega_2 = 0.3,\ \omega_3 = 54.$$

First, we find the parameters of the discriminant function by using the simplest classification method described in Section 2. According to this method, we can write that

$$\begin{aligned} f_{\text{simp}}(\mathbf{x}, w) &= \sum_{k=1}^{3} (\omega_k - v_k)\, x_k + \frac{1}{2} \sum_{k=1}^{3} \left(v_k^2 - w_k^2 \right) \\ &= 0.003x_1 + 0.2x_2 + 16x_3 - 736.05. \end{aligned}$$

In accordance with the proposed method, the parametric linear programming problem for the case $M_{-1} \leq 0 < M_1$ is

$$\begin{aligned} Q_1(q_{\text{opt}}, r) = \max_q \{ & q_0(1+r) + q_1\,(0.002 + r \cdot 0.005) \\ & + q_2\,(0.1 + r \cdot 0.3) + q_3\,(38 + r \cdot 54)\} \end{aligned}$$

subject to

$$\begin{aligned} 1/r &\leq q_0 + 0q_1 + 0q_2 + 16q_3 \leq 1, \\ 1/r &\leq q_0 + 0.01q_1 + 0q_2 + 16q_3 \leq 1, \\ 1/r &\leq q_0 + 0q_1 + 0.5q_2 + 16q_3 \leq 1, \\ 1/r &\leq q_0 + 0.01q_1 + 0.5q_2 + 16q_3 \leq 1, \\ 1/r &\leq q_0 + 0q_1 + 0q_2 + 64q_3 \leq 1, \\ 1/r &\leq q_0 + 0.01q_1 + 0q_2 + 64q_3 \leq 1, \\ 1/r &\leq q_0 + 0q_1 + 0.5q_2 + 64q_3 \leq 1, \\ 1/r &\leq q_0 + 0.01q_1 + 0.5q_2 + 64q_3 \leq 1, \end{aligned}$$

$$\begin{aligned} q_0 + 0.002q_1 + 0.1q_2 + 38q_3 \geq 0, \\ q_0 + 0.005q_1 + 0.3q_2 + 54q_3 \leq 0. \end{aligned}$$

The largest value 0.834 of the objective function $Q_1(q, r)$ is achieved by $r = -0.7$ at the point $q_0 = 1$, $q_1 = 0$, $q_2 = -4.857$, $q_3 = 0$ ($q_{\text{opt}} = (1, 0, -4.857, 0)$).

In the second case $a \leq 0 < M_{-1}$, the problem is

$$Q_2(q_{\text{opt}}) = \max_q \{q_0 + 0.005q_1 + 0.3q_2 + 54q_3\}$$

subject to

$$\begin{aligned} q_0 + 0q_1 + 0q_2 + 16q_3 &\leq 1, \\ q_0 + 0.01q_1 + 0q_2 + 16q_3 &\leq 1, \\ q_0 + 0q_1 + 0.5q_2 + 16q_3 &\leq 1, \\ q_0 + 0.01q_1 + 0.5q_2 + 16q_3 &\leq 1, \\ q_0 + 0q_1 + 0q_2 + 64q_3 &\leq 1, \\ q_0 + 0.01q_1 + 0q_2 + 64q_3 &\leq 1, \\ q_0 + 0q_1 + 0.5q_2 + 64q_3 &\leq 1, \\ q_0 + 0.01q_1 + 0.5q_2 + 64q_3 &\leq 1, \\ q_0 + 0.002q_1 + 0.1q_2 + 38q_3 &\leq 0. \end{aligned}$$

The problem has the optimal solution $q_0 = -19/13$, $q_1 = 0$, $q_2 = 0$, $q_3 = 1/26$. Hence, $Q_2(q_{\text{opt}}) = 0.615$.

In the third case $M_1 \leq 0 < b$, the problem is

$$Q_3(q_{\text{opt}}) = \max_q \{q_0 + 0.002q_1 + 0.1q_2 + 38q_3\}$$

subject to

$$\begin{aligned}
q_0 + 0q_1 + 0q_2 + 16q_3 &\leq 1,\\
q_0 + 0.01q_1 + 0q_2 + 16q_3 &\leq 1,\\
q_0 + 0q_1 + 0.5q_2 + 16q_3 &\leq 1,\\
q_0 + 0.01q_1 + 0.5q_2 + 16q_3 &\leq 1,\\
q_0 + 0q_1 + 0q_2 + 64q_3 &\leq 1,\\
q_0 + 0.01q_1 + 0q_2 + 64q_3 &\leq 1,\\
q_0 + 0q_1 + 0.5q_2 + 64q_3 &\leq 1,\\
q_0 + 0.01q_1 + 0.5q_2 + 64q_3 &\leq 1,\\
q_0 + 0.005q_1 + 0.3q_2 + 54q_3 &\leq 0.
\end{aligned}$$

The problem has the optimal solution $q_0 = 1$, $q_1 = 0$, $q_2 = -10/3$, $q_3 = 0$. Hence, $Q_3(q_{\text{opt}}) = 0.666$.

As a result, we get the minimax discriminant function $f(\mathbf{x}, w) = 1 - 4.857x_2$ because $Q_1(q_{\text{opt}}, 0.7) \geq Q_3(q_{\text{opt}}) \geq Q_2(q_{\text{opt}})$. It follows from the result that the second feature (the length) gives the largest value of the risk measure. If we remove all information concerning the features x_1 and x_3, then we get the same discriminant function. It is very interesting property of the considered algorithm. It determines that the second feature provides the largest error and then it searches for the discriminant function minimizing this error.

In order to evaluate the proposed method, we randomly generate test data in accordance with the uniform probability distributions having the mean values υ_k, ω_k, and the bounds $2\upsilon_k - \overline{x}_k, \overline{x}_k$ or $\underline{x}_k, 2\omega_k - \underline{x}_k$, $k = 1, 2, 3$. Then we exclude all points which do not belong to the interval $[\underline{x}_k, \overline{x}_k]$, $k = 1, 2, 3$, from consideration such that $N = 10000$ points remain inside the intervals $[\underline{x}_k, \overline{x}_k]$, $k = 1, 2, 3$. Predictive performance is quantified by means of the percentage correctly classified (PCC) on test data, also known as the overall classification accuracy, which is an estimate of a classifier's probability of a correct response. In other words, it is the proportion of correctly classified cases on a sample of data. PCC is the most widely used measure of classifier discriminatory power. Formally, it can be written as

$$\text{PCC} = \frac{1}{N} \sum_{i=1}^{N} \delta(y_i, t_i),$$

where y_i is the predicted class for vector $\mathbf{x}_i$, t_i is its true class; δ is 1 if both arguments are equal, 0 otherwise.

The simple comparison of the standard and proposed methods can not be correct because the proposed method uses the minimax or pessimistic strategy. It "copes" with extreme data which locate far from the mean values. Therefore, in order to take into account this fact, we exclude also points which lie inside the intervals $[\upsilon_k - \upsilon_k t, \upsilon_k + \upsilon_k t]$ and $[w_k - w_k t, w_k + w_k t]$. Table 1 (the second and the third rows) shows how the PCCs for the standard method (PCC1) and for the proposed method (PCC2) depend on t by uniformly distributed test data. It can be seen from the table that the standard method gives almost the same classification accuracy when $t = 0$. However, the proposed method is more accurate by dealing

Table 1 The PCC for two classification methods by uniformly and normally distributed random features

	Uniform distribution		Normal distribution	
t	PCC1	PCC2	PCC1	PCC2
0.0	0.50	0.49	0.52	0.28
0.1	0.39	0.54	0.37	0.31
0.2	0.22	0.60	0.23	0.39
0.3	0.22	0.64	0.21	0.41
0.4	0.18	0.70	0.20	0.47
0.5	0.16	0.74	0.17	0.51
0.6	0.00	0.84	0.00	0.62

with the “extreme” points. The small values of PCCs for both methods say that the available “training” information is too partial and imprecise in order to provide a reliable classification. Nevertheless, the additional information about the bounds of features and the minimax strategy can significantly improve the classification accuracy.

The forth and the fifth rows of Table 1 show similar results under condition that test data are generated in accordance with the normal probability distributions having the mean values v_k, ω_k, and the same standard deviations, $k = 1, 2, 3$. It can be seen from the table that the proposed minimax strategy can be compared with the standard method only by larger values of t. In other words, its classification accuracy decreases in this case. The proposed model does better for the uniformed distribution than the normal distribution because test data governed by the normal distribution are concentrated around the mean value of a feature. The uniformly distributed points contain a larger number of outliers. Therefore, this distribution is closer to the worst-case probability distributions.

It can be concluded from the example that the proposed method provides the better classification accuracy when test data are scattered and sparse for each class. From this point of view, the method akin to the robust classification approach where it is taken into account the fact that each point or observation can move around within a Euclidean ball or within a box (Lanckriet et al., 2002; Xu el al., 2009). However, the problem of determining an optimal ball radius or an optimal box size in the robust classification is successfully solved in the proposed method by applying sets of probability distributions consistent with the available information about mean values of features and their bounds.

8 Conclusion

A classification problem by the extremely limited information in the form of conditional expectations of features for every class has been studied in the paper. Its solution is based on the pessimistic (minimax) decision strategy. However, the optimistic (minimin) strategy can be also analyzed by means of the same approach based on representation of the available initial information in the form of p-boxes. Moreover, a cautious strategy as a linear “combination” of pessimistic and optimistic strategies with a predefined caution parameter can also be studied in the same way. This is a direction for further research.

What are the main advantages of the proposed method? If to compare the method with the simplest classification algorithm considered by (Scholkopf and Smola, 2002), then it uses additional information in the form of bounds for features, which can be very useful and can significantly improve the solution. The proposed method has a strong probabilistic background, and this fact allows us to use it in arbitrary applications where the initial information is scarce. In comparison with many classification methods where additional and often unjustified assumptions are employed, for instance, the known type of a probability distribution of features, the proposed method does not use any assumptions. It is based only on means of features and their bounds. Finally, it exploits the well-known minimax strategy which has a strong explanation, and it can be extended on a case of using different decision strategies. The minimax strategy allows us to classify "extreme" data.

The linear discriminant function f has been used in the paper. However, it is not difficult to show that the strong assumption of linearity of the discriminant function can be relaxed by considering the additive model with the discriminant function of the form:

$$f(\mathbf{x}, w) = w_0 + \sum_{i=1}^{m} w_i \psi_i(x_i).$$

Here ψ_i are some functions.

When we said about the possibility to apply the expert judgments, we did not mention that experts often prefer to provide intervals of mean values instead of their precise values due to several reasons. Of course, the proposed approach is simply extended on this case. The lower and upper CDFs of the corresponding p-box by interval-valued mean values of features is totally defined by bounds of the intervals. In other words, we have to use the bounds for expectations M_{-1} and M_1 obtained through the bounds of feature mean values. The number of constraints in the parametric linear programming problem increases in this case.

Another direction for further research is combined methods when some features are provided by sufficient statistical data and another part of features is described by mean values. These methods presuppose to apply a unified representation of different types of initial information and its incorporation into standard well-known approaches. One of the possible ways for exploiting such the combined information is to develop an extended Support Vector Machine (see for examples (Chandra et. al., 2010), (Doloc-Mihu, 2011), (Xu and Wang, 2011)).

It should be also noted that the knowledge of mean values is really extremely restrictive case. The next step which could give better (more informative) estimates for classification is to use Chebyshev's inequality that additionally draws the variance or the sample variance. This is also a direction for research in future.

Acknowledgement

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

References

Angulo, C., Anguita, D., Gonzalez-Abril, L., and Ortega, J. (2008). Support vector machines for interval discriminant analysis. *Neurocomputing*, 71(7-9):1220-1229.

Carrizosa, E., Gordillo, J., and Plastria, F. (2007). Classification problems with imprecise data through separating hyperplanes. Technical Report MOSI/33, MOSI Department, Vrije Universiteit Brussel.

Chandra, D.K., Ravi, V., and Ravisankar, P. (2010). Support vector machine and wavelet neural network hybrid: application to bankruptcy prediction in banks. *Int. J. of Data Mining, Modelling and Management*, 2(1):1-21.

Come, E., Oukhellou, L., Denoeux, T., and Aknin, P. (2009). Learning from partially supervised data using mixture models and belief functions. *Pattern Recognition*, 42(3):334-348.

Corani, G. and Zaffalon, M. (2008). Learning reliable classifiers from small or incomplete data sets: the naive credal classifier 2. *Journal of Machine Learning Research*, 9:581-621.

de Cooman, G. and Zaffalon, M. (2004). Updating beliefs with incomplete observations. *Artificial Intelligence*, 159(1-2):75–125.

Denoeux, T. and Smets, P. (2006). Classification using belief functions: the relationship between the case-based and model-based approaches. *IEEE Transactions on Systems, Man and Cybernetics B*, 36(6):1395-1406.

Destercke, S., Dubois, D., and Chojnacki, E. (2008). Unifying practical uncertainty representations - i: Generalized p-boxes. *Int. J. of Approximate Reasoning*, 49(3):649-663.

Doloc-Mihu, A. (2011). Kernel method for improving image retrieval performance: a survey. *Int. J. of Data Mining, Modelling and Management*, 3(1):42-74.

Ferson, S., Ginzburg, L., and Akcakaya, R. (2001). Whereof one cannot speak: When input distributions are unknown. Technical report, Applied Biomathematics Report. http://www.ramas.com/whereof.pdf.

Huhn, J. and Hullermeier, E. (2008). Is an ordinal class structure useful in classifier learning? *Int. J. Data Mining, Modelling and Management*, 1(1):45-67.

Ishibuchi, H., Tanaka, H., and Fukuoka, N. (1990). Discriminant analysis of multi-dimensional interval data and its application to chemical sensing. *Int. J. of General Systems*, 16(4):311-329.

Kuznetsov, V. P. (1991). *Interval Statistical Models.* Radio and Communication, Moscow. in Russian.

Lad, B. K., Kulkarni, M. S. (2010). A parameter estimation method for machine tool reliability analysis using expert judgement. *Int. J. of Data Analysis Techniques and Strategies*, 2(2):155-169.

Lanckriet, G.R.G., El Ghaoui, L., Bhattacharyya, C., and Jordan, M.I. (2003). A robust minimax approach to classification. *The Journal of Machine Learning Research*, 3:555-582.

Masson, M.-H. and Denoeux, T. (2004). Clustering interval-valued data using belief functions. *Pattern Recognition Letters*, 25(2):163-171.

Nivlet, P., Fournier, F., and Royer, J.-J. (2001). Interval discriminant analysis: An efficient method to integrate errors in supervised pattern recognition. In *Second International Symposium on Imprecise Probabilities and Their Applications*, pages 284–292, Ithaca, NY, USA.

Pelckmans, K., Brabanter, J. D., Suykens, J., and Moor, B. D. (2005). Handling missing values in support vector machine classifiers. *Neural Networks*, 18(5-6):684-692.

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

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

Utkin, L. (2010). Regression analysis using the imprecise Bayesian normal model. *Int. J. of Data Analysis Techniques and Strategies*, 2:356–372.

Utkin, L. and Coolen, F. (2011) On reliability growth models using Kolmogorov-Smirnov bounds. *Int. J. of Performability Engineering*, 7:5-19.

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

Walley, P. (1996). Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58:3-57.

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 Y., Wang H. (2011). A new feature selection method based on support vector machines for text categorisation. *Int. J. of Data Analysis Techniques and Strategies*, 3(1):1-20.

Zaffalon, M. (2002). Exact credal treatment of missing data. *Journal of Statistical Planning and Inference*, 105(1):105-122.