Available online at www.sciencedirect.com

ScienceDirect

Neurocomputing 71 (2008) 1220–1229

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Support vector machines for interval discriminant analysis ☆

C. Angulo[a,*], D. Anguita[b], L. Gonzalez-Abril[c], J.A. Ortega[d]

[a] *GREC-Knowledge Engineering Research Group, Technical University of Catalonia, 08800 Vilanova i la Geltrú, Spain*

[b] *DIBE-Department of Biophysical and Electronic Engineering, University of Genoa, 16145 Genoa, Italy*

[c] *COSDE-Department of Applied Economics I, University of Seville, 41018 Seville, Spain*

[d] *Computer Engineering Department, University of Seville, 41018 Seville, Spain*

Available online 30 January 2008

Abstract

The use of data represented by intervals can be caused by imprecision in the input information, incompleteness in patterns, discretization procedures, prior knowledge insertion or speed-up learning. All the existing support vector machine (SVM) approaches working on interval data use local kernels based on a certain distance between intervals, either by combining the interval distance with a kernel or by explicitly defining an interval kernel. This article introduces a new procedure for the linearly separable case, derived from convex optimization theory, inserting information directly into the standard SVM in the form of intervals, without taking any particular distance into consideration.

Keywords: Interval analysis; Kernel machines; Convex optimization; Classification

1. Introduction

Support vector machines—SVMs—are learning machines which implement the structural risk minimization inductive principle to obtain good generalization on a limited number of learning patterns. This theory was originally developed by V. Vapnik on the basis of a separable binary classification problem with signed outputs ± 1 [27].

Our aim in this work is to enable SVMs to deal with information in the form of intervals. These intervals can be incorporated for several reasons. For example, an SVM is a powerful paradigm which is not especially well addressed for dealing with large databases due to the quadratic program being solved. The training task can be scaled up by using the interval data concept [1]: large databases are aggregated into smaller data sizes in the form of intervals by using the k-means algorithm, the neural gas algorithm [16] or by another clustering procedure. Incomplete information patterns must sometimes be dealt with by machine learning algorithms before they can be completed, in the case of prognosis for instance. In this case, intervals can represent the incomplete information. In fact, in many practical applications, information is imprecise and incomplete since it is derived from uncertain measurements or linguistic assessments, which can be represented by interval data. Discretization is yet another procedure transforming a continuous attribute into a discrete attribute that can be handled in the form of intervals. Many real-life world applications can be found in the literature, from reliable computing [14] to embedded systems [22] and economics [21].

Interval arithmetic was first developed in [17] to compute interval data as a way of handling uncertainties and many developments have followed [9,20,13]. The main drawback concerning the space of the intervals is that it is not a Euclidian space, and hence a norm cannot be defined. Nevertheless, it is still possible to define a distance and this fact has been used in the first approaches using SVMs on interval data. Thus, it has been proved that SVMs can

☆ This work was partly supported by the ADA Grant (DPI2006-15630-C02-01) from the Spanish Ministry of Education and Science. The corresponding author was granted with an UPC mobility grant during this research at DIBE, University of Genoa, Italy and a Network Action of the EU-funded IST CogSys Coordinated Action, euCognition.

*Corresponding author. Tel.: +34 938967798; fax: +34 938967700.

E-mail address: cecilio.angulo@upc.edu (C. Angulo).

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.12.025

handle interval regression analysis [11,12], whose objective is to design a regression model for the data where coefficients are intervals, of which possibilistic regression analysis [25] is the simplest version. For classification problems, the most used approach [3,28] is to work on interval data by combining an RBF kernel function with a distance defined on the space of intervals, for example the Hausdorff distance. Monotony of the Gaussian function allows the relative distance between intervals to be maintained. This approach has also been considered on other non-Euclidian qualitative spaces such as the orders of magnitude space [23]. Another possibility is to define a particular nonlinear function which enables us with intervals on a pre-defined feature space [6,24].

All the approaches in the precedent cases use local kernels, i.e. they work primarily in a neighbourhood, based on a distance between intervals, sometimes combining the interval distance with a kernel, sometimes by explicitly defining the interval kernel. Nevertheless, it is possible to use another two solutions: (i) to move the problem towards a probabilistic approach [19], in an attempt to deal with interval discriminant analysis, as does interval regression analysis, or (ii) to directly incorporate an interval approach into the SVM. Both new approaches would allow us to deal with the uncertainty associated with the imprecision of data, and the interval nature of the problem. We are concerned with the latter approach in this article.

Starting with a brief introduction of SVM learning in Section 2, a first direct formulation for interval SVM, called I-SVM, is derived in Section 3 by using the approach in [5] on how to introduce prior knowledge in the form of multiple polyhedral sets into a reformulation of a linear SVM classifier, for the case of intervals. The size of the obtained QP problem, i.e. the number of parameters and constraints for both primal and dual formulation, is very large. For this reason, a new approach is developed in Section 4 which drastically reduces the I-SVM size. Furthermore, it is demonstrated that this new procedure can be interpreted as based on interval arithmetic. An illustrative example is provided. The complexity of the approach is analysed and compared with other similar approaches in Section 5. Experimentation with the proposed machine on two real examples from [5] is detailed. Finally, Section 6 presents conclusions and potential lines of research.

2. SVM Learning

The SVM is an implementation of a more general regularization principle known as the large margin principle [2]. Let

$$\mathscr{Z} = \{(x_1, y_1), \ldots, (x_p, y_p)\} = \{z_1, \ldots, z_p\} \in (\mathscr{X} \times \mathscr{Y})^p \tag{1}$$

be a training set, where $\mathscr{X}$ is the input space and $\mathscr{Y} = \{\theta_1, \theta_2\} = \{+1, -1\}$ the output space. Let $\phi : \mathscr{X} \to \mathscr{F} \subseteq \mathbb{R}^d$ be a feature mapping, with $\phi = (\phi_1, \ldots, \phi_d)$, for the usual 'kernel trick'. $\mathscr{F}$ is named *feature space* with a dot product denoted by $\langle \cdot, \cdot \rangle_{\mathscr{F}}$. Let $\mathrm{x} \triangleq \phi(x) \in \mathscr{F}$ be the *representation* of $x \in \mathscr{X}$. A binary linear classifier,

$$f_{\mathrm{w}}(x) = \langle \phi(x), \mathrm{w} \rangle_{\mathscr{F}} + b = \langle \mathrm{x}, \mathrm{w} \rangle_{\mathscr{F}} + b \tag{2}$$

is sought in the space $\mathscr{F}$, with $f_{\mathrm{w}} : \mathscr{X} \to \mathscr{F} \to \mathbb{R}$, $b \in \mathbb{R}$, and where outputs are obtained by thresholding the classifier, $h_{\mathrm{w}}(x) = \mathrm{sign}(f_{\mathrm{w}}(x))$. Henceforth $\langle \cdot, \cdot \rangle_{\mathscr{F}}$ is denoted as $\langle \cdot, \cdot \rangle$.

Let us first consider the linearly separable case, which is the use of a linear kernel from a kernel machine perspective. Let β and α be the minimum and the maximum values for classes labelled as $\{+1, -1\}$, respectively, i.e.

$$\beta = \min_{z_i \in \mathscr{Z}_+} \langle \mathrm{x_i}, \mathrm{w} \rangle, \quad \alpha = \max_{z_i \in \mathscr{Z}_-} \langle \mathrm{x_i}, \mathrm{w} \rangle,$$

where $\mathscr{Z}_+$ and $\mathscr{Z}_-$ are respectively the patterns belonging to the classes labelled as $\{\theta_1, \theta_2\} = \{+1, -1\}$. Hence, according to [10] and by taking into account that the margin is $(\beta - \alpha)/\|\mathrm{w}\|$, the classifier w with the largest geometrical margin on a given training sample $\mathscr{Z}$ can be written as

$$\mathrm{w_{SVM}} \triangleq \arg \max_{\mathrm{w} \in \mathscr{F}} \frac{1}{\|\mathrm{w}\|} \cdot \min_{z_i \in \mathscr{Z}} y_i \langle \mathrm{x_i}, \mathrm{w} \rangle. \tag{3}$$

There are several practical methods of dealing with this problem [7], one of which is to minimize the norm $\|\mathrm{w}\|$ in (3) while $\beta - \alpha = 2$. This canonical form of the optimal separation introduces the bias term by defining $\beta = b + 1$ and $\alpha = b - 1$ which determine the location of the separating hyperplane relative to the origin. Hence, the problem is translated into the optimization problem

$$\begin{aligned} \min_{\mathrm{w} \in \mathscr{F}} \quad & \tfrac{1}{2}\|\mathrm{w}\|^2 \\ \text{s.t.} \quad & y_i(\langle \mathrm{x_i}, \mathrm{w} \rangle - b) \geqslant 1, \; z_i \in \mathscr{Z}. \end{aligned} \tag{4}$$

The solution can be expressed in the form

$$\mathrm{w_{SVM}} = \sum_i \alpha_i y_i \mathrm{x_i}, \quad f_{\mathrm{w_{SVM}}}(x) = \sum_i \alpha_i \, y_i \, k(x_i, x) + b, \tag{5}$$

where $k(x, x') = \langle \phi(x), \phi(x') \rangle = \langle \mathrm{x}, \mathrm{x}' \rangle$ is the kernel function, and where only those few α_i associated with the so-called *support vectors* are not zero. The bias b is usually calculated as the halfway point between hyperplanes [8].

3. Interval-SVM: a convex optimization approach

We are attempting to directly incorporate an interval approach into the SVMs by inserting information into the standard SVM in the form of intervals, so we call Interval-SVM as I-SVM. Henceforth, the problem statement will be constrained to $\mathbb{R}^m$. The training set $\mathscr{Z}$ composed of only exact points (1), is added with the following set of intervals in $\mathbb{R}^m$:

$$\mathscr{T} = \{(I_{p+1}, y_{p+1}), \ldots, (I_n, y_n)\} \in (\mathbb{I}^m \times \mathscr{Y})^r, \tag{6}$$

where $I_i = (I_{i1}, \ldots, I_{im})$ with $I_{ij} = [x_{ij}^{\mathrm{L}}, x_{ij}^{\mathrm{U}}]$ and $x_{ij}^{\mathrm{L}} \leqslant x_{ij}^{\mathrm{U}}$, is an element in the space of the m-dimensional closed intervals $\mathbb{I}^m$ for $i = p+1, \ldots, n$, $n = p+r$ and $j = 1, \ldots, m$. The lower bound of any interval I is given by $x^{\mathrm{L}} \triangleq (x_j^{\mathrm{L}})_j \triangleq (x_1^{\mathrm{L}}, \ldots, x_m^{\mathrm{L}})$, and its upper bound, $x^{\mathrm{U}} \triangleq (x_j^{\mathrm{U}})_j \triangleq (x_1^{\mathrm{U}}, \ldots, x_m^{\mathrm{U}})$. Hence $x^{\mathrm{L}} \leqslant x^{\mathrm{U}}$ and for the sake of simplicity, it is denoted as $I = [x^{\mathrm{L}}, x^{\mathrm{U}}]$. Furthermore, inputs in the form of points can be considered as intervals where the lower and upper bounds are the same, $\{x_i\} = [x_i, x_i]$, thereby obtaining as the general training set,

$$\begin{aligned}\mathscr{X}\mathscr{T} &\triangleq \mathscr{X} \cup \mathscr{T} \\ &= \{(I_1, y_1), \ldots, (I_p, y_p), (I_{p+1}, y_{p+1}), \ldots, (I_n, y_n)\} \\ &= \{z_1, \ldots, z_p, z_{p+1}, \ldots, z_n\} \in (\mathbb{I}^m \times \mathscr{Y})^n. \end{aligned} \tag{7}$$

The interval discriminant learning problem can now be defined as the search for a binary linear classifier in the form of (2) with $f_{\mathrm{w}} : \mathbb{I}^m \to \mathscr{F} \to \mathbb{R}$, $b \in \mathbb{R}$, and where outputs are obtained by thresholding the classifier, $h_{\mathrm{w}}(x) = \mathrm{sign}(f_{\mathrm{w}}(x))$.

3.1. I-SVM: primal formulation by convex optimization

In [5] it is shown how prior knowledge in the form of multiple polyhedral sets can be introduced into a reformulation of a linear SVM classifier. All the points lying in a polyhedral set can be determined by a general set of linear inequalities,

$$\mathscr{B} = \{x | Bx \leqslant d\} \subset \mathbb{R}^m \tag{8}$$

with $B \in \mathbb{R}^{s \times m}$ and $d \in \mathbb{R}^s$.

By using the Farkas theorem [15], it can be derived from Proposition 2.1 in [5] that, given a weight vector w, a nonempty polyhedron $\mathscr{B}$ lies in the half-space $\{x | w' \cdot x \geqslant 1\}$ (alternatively, $\{x | -w' \cdot x \leqslant -1\}$) if and only if there exists a vector u such that $B' \cdot u + w = 0$ (alternatively, $B' \cdot u - w = 0$), $d' \cdot u + 1 \leqslant 0$ and $u \geqslant 0$, where B' indicates the transpose of B.

For the case of an interval $I = \left[x^{\mathrm{L}}, x^{\mathrm{U}}\right] \subset \mathbb{R}^m$, it can be described in the form of (8), $I = \{x | Bx \leqslant d\} \subset \mathbb{R}^m$, which gives

$$B = \begin{pmatrix} 1 & 0 \cdots 0 & 0 \\ -1 & 0 \cdots 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 \cdots 0 & 1 \\ 0 & 0 \cdots 0 & -1 \end{pmatrix} \in \mathbb{R}^{2m \times m},$$

$$d = \begin{pmatrix} x_1^{\mathrm{U}} \\ -x_1^{\mathrm{L}} \\ \vdots \\ x_m^{\mathrm{U}} \\ -x_m^{\mathrm{L}} \end{pmatrix} \in \mathbb{R}^{2m}. \tag{9}$$

Without any loss of generality, let us assume that the bias b is initially not considered in the optimization problems. Hence, by applying the result in [5], the primal QP problem can be formulated as (note that the B matrix is the same for any interval),[1]

$$\begin{aligned} &\min_{w \in \mathbb{R}^m, \{u^i\}_{i=1}^n \in \mathbb{R}^{2m}} \quad \tfrac{1}{2}\|w\|^2 \\ &\text{s.t.} \quad \begin{aligned} &B' \cdot u^i + y_i \cdot w = \mathbf{0}, \quad i = 1, \ldots, n, \\ &(d^i)' \cdot u^i + 1 \leqslant 0, \quad i = 1, \ldots, n. \\ &u^i \geqslant \mathbf{0}, \quad i = 1, \ldots, n. \end{aligned} \end{aligned} \tag{10}$$

By applying expressions for B and d^i in (9) for the case of intervals and by naming $u^i = (u_{11}^i, u_{12}^i, \ldots, u_{m1}^i, u_{m2}^i)' \in \mathbb{R}^{2m}$, it can be demonstrated by linear algebra that the QP problem becomes

$$\begin{aligned} &\min_{w \in \mathbb{R}^m, \{u^i\}_{i=1}^n \in \mathbb{R}^{2m}} \quad \tfrac{1}{2}\|w\|^2 \\ &\text{s.t.} \quad \begin{aligned} &y_i \cdot w = (u_{j2}^i)_j - (u_{j1}^i)_j, \quad i = 1, \ldots, n, \\ &y_i \cdot w' \cdot x_i^{\mathrm{L}} \geqslant 1 + (u_{j1}^i)_j' \cdot (\Delta x_i), \quad i = 1, \ldots, n, \\ &u^i \geqslant \mathbf{0}, \quad i = 1, \ldots, n, \end{aligned} \end{aligned} \tag{11}$$

where $\Delta x_i = x_i^{\mathrm{U}} - x_i^{\mathrm{L}} = \left(x_{i1}^{\mathrm{U}} - x_{i1}^{\mathrm{L}}, \ldots, x_{im}^{\mathrm{U}} - x_{im}^{\mathrm{L}}\right) \in \mathbb{R}^m$.

By taking into consideration the $nm + n + 2nm = n(3m+1)$ restrictions to be accomplished, it can be seen that the obtained QP problem is over-parameterized with $(2n+1)m$ parameters being optimized. Hence, by assuming that a solution exist for the problem, weight vector w can be obtained in n forms, by using any of the n first sets of m constraints in (11), $w = y_i \cdot ((u_{j2}^i)_j - (u_{j1}^i)_j)$. It should be borne in mind that all the constraints considered in this first set are equality constraints, and that no constraint is stronger than another.

By replacing $y_i \cdot w$ in the second constraint in (11) for the value in the first constraint, and defining $w = 1/n \sum_i y_i (u_2^i - u_1^i)$ for robustness, with $u_s^i = (u_{js}^i)_j$, for $s = 1, 2$, a similar primal QP problem[2] is obtained,

$$\begin{aligned} &\min_{\{u_1^i, u_2^i\}_{i=1}^n \in \mathbb{R}^m} \quad \tfrac{1}{2n^2} \left\| \sum_i y_i (u_2^i - u_1^i) \right\|^2 \\ &\text{s.t.} \quad \begin{aligned} &(u_2^i)' \cdot x_i^{\mathrm{L}} - (u_1^i)' \cdot x_i^{\mathrm{U}} \geqslant 1, \quad i = 1, \ldots, n, \\ &u_1^i, u_2^i \geqslant \mathbf{0}, \quad i = 1, \ldots, n. \end{aligned} \end{aligned} \tag{12}$$

In this form, $2nm$ parameters must be optimized on $n(2m+1)$ constraints, i.e. the number of parameters and constraints in (12) are much smaller than those in (11).

[1] A note about the notation. When QP problems are settled, we distinguish a set of vectorial constraints or a vector of constraints from a set of one-dimensional constraints by boldfacing the vectors of zeros, $\mathbf{0}$ or ones, $\mathbf{1}$. We also use boldfaced letters for matrices and sub-matrix components involved in QP problems. In particular, we denote the n-dimensional identity matrix by $\mathbf{I_n}$.

[2] Minor computing errors are unavoidable when numerical solutions for the QP problem are calculated, therefore it is more robust to deal with a constraint which is the mean of all the equality constraints, as is usually performed, for instance, when the bias is calculated for the standard biclass SVM.

3.2. I-SVM QP dual formulation by convex optimization

The dual of QP primal formulations can be obtained, as usual, by using Lagrangian formulation based on multipliers and the KKT (Karush–Kuhn–Tucker) conditions. By using the simplest expression obtained (12) when convex optimization approach was employed as the primal QP problem, the Lagrangian function to be maximized as a dual formulation becomes

$$\begin{aligned} L(\{u_1^i, u_2^i\}_{i=1}^n) = \frac{1}{2n^2}\left\|\sum_{i=1}^n y_i(u_2^i - u_1^i)\right\|^2 & \\ - \sum_{i=1}^n \alpha_i \cdot ((u_2^i)' \cdot x_i^{\mathrm{L}} - (u_1^i)' \cdot x_i^{\mathrm{U}} - 1) & \\ - \sum_{i=1}^n \mu_i' \cdot u_1^i & \\ - \sum_{i=1}^n v_i' \cdot u_2^i & \qquad (13) \end{aligned}$$

with $\alpha_i \geqslant 0$ and $\mu_i, v_i \in \mathbb{R}^m$ such that $\mu_i, v_i \geqslant \mathbf{0}$.

The KKT conditions, $\partial L/\partial u_1^i = 0$ and $\partial L/\partial u_2^i = 0$, for $i = 1, \ldots, n$, result in the following expressions:

$$\frac{1}{n^2} y_i \sum_{j=1}^n y_j(u_2^j - u_1^j) = \alpha_i \cdot x_i^{\mathrm{U}} - \mu_i,$$

$$\frac{1}{n^2} y_i \sum_{j=1}^n y_j(u_2^j - u_1^j) = \alpha_i \cdot x_i^{\mathrm{L}} + v_i. \qquad (14)$$

The Lagrangian function in (13) can be expressed as

$$\begin{aligned} L(\{u_1^i, u_2^i\}_{i=1}^n) = \frac{1}{2n^2}\left\|\sum_{i=1}^n y_i(u_2^i - u_1^i)\right\|^2 & \\ - \sum_{i=1}^n u_2^i \cdot (\alpha_i \cdot x^{\mathrm{L}} + v_i) & \\ + \sum_{i=1}^n u_1^i \cdot (\alpha_i \cdot x^{\mathrm{U}} - \mu_i) & \\ + \sum_{i=1}^n \alpha_i & \qquad (15) \end{aligned}$$

and substituting KKT conditions into this expression, enables the Lagrangian function in (13) to be expressed as

$$L(\{u_1^i, u_2^i\}_{i=1}^n) = -\frac{1}{2n^2}\left\|\sum_{i=1}^n y_i(u_2^i - u_1^i)\right\|^2 + \sum_{i=1}^n \alpha_i. \qquad (16)$$

Finally, two sets of expressions derived from KKT conditions can be employed to calculate the weight vector $w = 1/N \sum_i y_i(u_2^i - u_1^i)$. Without any loss of generality, let us use the second set in (14). The dual QP problem becomes,[3]

$$\begin{aligned} \min_{\alpha\in\mathbb{R}^n, v\in\mathbb{R}^{nm}} \quad & \tfrac{1}{2}(\alpha', v') \cdot \mathbf{Q_L} \cdot \begin{pmatrix} \alpha \\ v \end{pmatrix} - (\mathbf{1}', \mathbf{0}') \cdot \begin{pmatrix} \alpha \\ v \end{pmatrix} \\ \text{s.t.} \quad & \alpha_i \geqslant 0, \quad i = 1, \ldots, n, \\ & v_i \geqslant \mathbf{0}, \quad i = 1, \ldots, n, \end{aligned} \qquad (17)$$

where

$$\mathbf{Q_L} = \left(\begin{array}{c|c} \mathbf{q_L'q_L} & \mathbf{q_L'} \cdot (\mathbf{I_m}, \ldots, \mathbf{I_m}) \\ \hline (\mathbf{I_m}, \ldots, \mathbf{I_m})' \cdot \mathbf{q_L} & \mathbf{I_{nm}} \end{array}\right) \qquad (18)$$

provided that $\mathbf{q_L'} = (y_1 \cdot x_1^{\mathrm{L}}, \ldots, y_n \cdot x_n^{\mathrm{L}})$, $\alpha' = (\alpha_1, \ldots, \alpha_n)$ and $v' = (v_1', \ldots, v_n')$.

The weight vector w solution can be written as

$$w = \mathbf{q_L} \cdot \boldsymbol{\alpha} + \sum_{i=1}^n y_i \cdot v_i \qquad (19)$$

and bias b is obtained a posteriori as usual.

4. I-SVM: towards an interval analysis approach

The objective in this article is to reduce the oversized QP problem obtained from the result in [5] for the case of intervals to one more amenable QP problem leading to a direct interval-like SVM formulation. Let us now introduce the formulation for the interval discriminant learning problem.

4.1. Interval SVM problem

Let $I^1 = [x_1^{\mathrm{L}}, x_1^{\mathrm{U}}] = (I_1^1, \ldots, I_m^1) \in \mathbb{I}^m$ and $I^2 = [x_2^{\mathrm{L}}, x_2^{\mathrm{U}}] = (I_1^2, \ldots, I_m^2) \in \mathbb{I}^m$ be two intervals, where $I_i^j = [x_{ji}^{\mathrm{L}}, x_{ji}^{\mathrm{U}}]$ for $j = 1, 2$ and $i = 1, \ldots, m$. The following operations are defined:

- Sum: $I^1 + I^2 = [x_1^{\mathrm{L}} + x_2^{\mathrm{L}}, x_1^{\mathrm{U}} + x_2^{\mathrm{U}}] \in \mathbb{I}^m$;
- Difference: $I^1 - I^2 = [x_1^{\mathrm{L}} - x_2^{\mathrm{U}}, x_1^{\mathrm{U}} - x_2^{\mathrm{L}}] \in \mathbb{I}^m$;
- Product by a scalar: $\lambda I^1 = [\min\{\lambda x_1^{\mathrm{L}}, \lambda x_1^{\mathrm{U}}\}, \max\{\lambda x_1^{\mathrm{L}}, \lambda x_1^{\mathrm{U}}\}] \in \mathbb{I}^m$;
- Product: $I^1 \cdot I^2 = [\min A, \max A] \in \mathbb{I}^1$ where
$$A = \left\{\sum_{i=1}^m x_{1i}^{\mathrm{L}} x_{2i}^{\mathrm{L}}, \sum_{i=1}^m x_{1i}^{\mathrm{L}} x_{2i}^{\mathrm{U}}, \sum_{i=1}^m x_{1i}^{\mathrm{U}} x_{2i}^{\mathrm{L}}, \sum_{i=1}^m x_{1i}^{\mathrm{U}} x_{2i}^{\mathrm{U}}\right\}; \qquad (20)$$
- Minimum: $\min I^1 = x^{\mathrm{L}} \in \mathbb{R}^m$;
- Inequality: $I^1 \succeq I^2 \Leftrightarrow x_1^{\mathrm{L}} \geqslant x_2^{\mathrm{U}}$, and $I^1 \succeq q \Leftrightarrow x_1^{\mathrm{L}} \geqslant q$ where $q \in \mathbb{R}^m$.

The QP problem to be dealt with from an interval arithmetic perspective is,

$$\begin{aligned} \min_{w\in\mathbb{R}^m} \quad & \tfrac{1}{2}\|w\|^2 \\ & y_i \cdot w' \cdot I_i \succeq 1, \quad z_i \in \mathscr{Z}\mathscr{T}, \end{aligned} \qquad (21)$$

[3] Using the first set of conditions, equivalent results are obtained.

which can be considered as a generalization of the standard definition in (4). A solution function from $\mathbb{I}^m$ to $\mathbb{I}^1$ is considered instead of a real function,

$$f_w(I) \triangleq [f_w^{\mathrm{L}}(I), f_w^{\mathrm{U}}(I)] = w' \cdot I \tag{22}$$

provided that $I = [x_i^{\mathrm{L}}, x_i^{\mathrm{U}}]$ and w is a vector, with

$$\min_{x \in I} w' \cdot x = f_w^{\mathrm{L}}(I).$$

$$\max_{x \in I} w' \cdot x = f_w^{\mathrm{U}}(I), \tag{23}$$

which leads to a new QP formulation,

$$\begin{aligned} \min_{w \in \mathbb{R}^m} \quad & \tfrac{1}{2}\|w\|^2 \\ & y_i \cdot [f_i^{\mathrm{L}}, f_i^{\mathrm{U}}] \succeq 1, \quad z_i \in \mathscr{Z}\mathscr{T} \end{aligned} \tag{24}$$

with $w' \cdot I_i = \{w' \cdot x \mid x \in I_i\} = [f_i^{\mathrm{L}}, f_i^{\mathrm{U}}]$.

In fact, the standard primal QP problem for SVM can be recovered from the I-SVM extension if all the information is exact, i.e. no interval exists. In this case $x_i^{\mathrm{L}} = x_i^{\mathrm{U}} = x_i$, and hence $f_i^{\mathrm{L}} = f_i^{\mathrm{U}} = \min_{x_i \in I_i} w' \cdot I_i = w' \cdot x_i$.

Let β and α be the minimum and the maximum absolute values for each class, i.e.

$$\beta = \min_{z_i \in \mathscr{Z}_+} f_i^{\mathrm{L}} = \min_{z_i \in \mathscr{Z}_+} y_i f_i^{\mathrm{L}},$$

$$\alpha = \max_{z_i \in \mathscr{Z}_-} f_i^{\mathrm{U}} = \min_{z_i \in \mathscr{Z}_-} y_i f_i^{\mathrm{U}}.$$

The QP problem (21) or (24) corresponds to finding the classifier w with the largest geometrical margin $(\beta - \alpha)/\|w\|$ on a given training sample $\mathscr{Z}\mathscr{T}$ of intervals,

$$w_{\mathrm{I-SVM}} \triangleq \arg\max_{w \in \mathbb{R}^m} \frac{1}{\|w\|} \cdot \min_{z_i \in \mathscr{Z}\mathscr{T}} \left(\min_{x \in I_i} y_i \cdot w' \cdot x \right). \tag{25}$$

A bias can be added as the value halfway between values α and β [10,7] to solution (22).

4.2. I-SVM: reduced primal formulation

The size of the QP problem (12), i.e. the number of parameters and constraints, could be reduced because it is still very large. An approach from interval arithmetics will be developed which drastically reduces this size. Let us consider the QP problem (24) defined in the form,

$$\begin{aligned} \min_{w^+, w^- \in \mathbb{R}^m} \quad & \tfrac{1}{2}\|w^+ - w^-\|^2 \\ & y_i \cdot [f_i^{\mathrm{L}}, f_i^{\mathrm{U}}] \succeq 1, \quad z_i \in \mathscr{Z}\mathscr{T}, \\ & w^+, w^- \geqslant \mathbf{0}, \end{aligned} \tag{26}$$

where $w = w^+ - w^-$ is the solution.

Due to the positiveness of the terms w^+ and w^-, and observing that

$$f_i^{\mathrm{L}} = \sum_{j=1}^m (w_j^+)' \cdot x_{ij}^{\mathrm{L}} - \sum_{j=1}^m (w_j^-)' \cdot x_{ij}^{\mathrm{U}},$$

$$f_i^{\mathrm{U}} = \sum_{j=1}^m (w_j^+)' \cdot x_{ij}^{\mathrm{U}} - \sum_{j=1}^m (w_j^-)' \cdot x_{ij}^{\mathrm{L}}, \tag{27}$$

the QP problem (26) can be translated into

$$\begin{aligned} \min_{w^+, w^- \in \mathbb{R}^m} \quad & \tfrac{1}{2}\|w^+ - w^-\|^2 \\ & y_i \cdot ((w^+)' \cdot x_i^{\mathrm{L}} - (w^-)' \cdot x_i^{\mathrm{U}}) \geqslant 1, \quad y_i = 1, \\ \text{s.t.} \quad & y_i \cdot ((w^+)' \cdot x_i^{\mathrm{U}} - (w^-)' \cdot x_i^{\mathrm{L}}) \geqslant 1, \quad y_i = -1, \\ & w^+, w^- \geqslant \mathbf{0}, \end{aligned} \tag{28}$$

which is representing a reduced form of the QP problem (12) by assuming that every parameter associated with positive- or negative-labelled patterns has the same value:

$$u_1^i = w^- u_2^i = w^+, \quad y_i = 1,$$

$$u_1^i = w^+ u_2^i = w^-, \quad y_i = -1. \tag{29}$$

Due to assumption (29), we can infer that not all possible original solutions for (12) will be maintained in our formulation (28), but the new formulation is built on a good assumption for several reasons.

Firstly, it should be explained that the procedure introduced for the linearly separable case will allow to insert information directly into the standard SVM in the form of intervals, without taking any particular distance into consideration, because both terms defining w (w^+ and w^-) are positive and constant.

Secondly, this formulation is much simpler than those previously obtained in (11) and (12), since $2m$ parameters are optimized on an $n + 2m$ constrained optimization problem. For the dual case, this reduced training task effort is similar to that used for standard SVM on punctual data, so intuitively a good performance in terms of accuracy should be expected.

Finally, experimentation with the proposed machine on the previous two real examples used in [5] will allow us to observe that similar or better accuracy is obtained with the simplified approach than with the original full optimization approach.

4.3. I-SVM QP dual by interval analysis

Starting from the original primal formulation (28), the Lagrangian function to be maximized as dual formulation is

$$\begin{aligned} L(w^+, w^-) = {} & \frac{1}{2}\|w^+ - w^-\|^2 \\ & - \sum_{\mathscr{Z}_+} \gamma_i \cdot ((w^+)' \cdot x_i^{\mathrm{L}} - (w^-)' \cdot x_i^{\mathrm{U}} - 1) \\ & - \sum_{\mathscr{Z}_-} \beta_i \cdot ((w^-)' \cdot x_i^{\mathrm{L}} - (w^+)' \cdot x_i^{\mathrm{U}} - 1) \\ & - \sum_{j=1}^m \mu_j \cdot w_j^+ - \sum_{j=1}^m \nu_j \cdot w_j^- \end{aligned} \tag{30}$$

with $\gamma_i, \beta_i \geqslant 0$, $\mu_j, v_j \geqslant 0$. For notational simplicity, it is assumed that the first training patterns are positive instances, $\mathscr{X}_+ \subset \mathscr{X}$, and the latter are negative instances, $\mathscr{X}_- \subset \mathscr{X}$.

By imposing KKT conditions, $\partial L/\partial w^+ = 0$ and $\partial L/\partial w^- = 0$, we obtain

$$w^+ - w^- = \sum_{\mathscr{X}_+} \gamma_i x_i^{\mathrm{L}} - \sum_{\mathscr{X}_-} \beta_i x_i^{\mathrm{U}} + \mu \triangleq w_{\inf} + \mu,$$

$$w^+ - w^- = \sum_{\mathscr{X}_+} \gamma_i x_i^{\mathrm{U}} - \sum_{\mathscr{X}_-} \beta_i x_i^{\mathrm{L}} - v \triangleq w_{\sup} - v. \tag{31}$$

The Lagrangian function in (30) can be expressed as

$$\begin{aligned} L(w^+, w^-) = {} & \frac{1}{2}\|w^+ - w^-\|^2 \\ & + (w^+)' \cdot \left(-\sum_{\mathscr{X}_+} \gamma_i x_i^{\mathrm{L}} + \sum_{\mathscr{X}_-} \beta_i x_i^{\mathrm{U}} - \mu \right) \\ & + (w^-)' \cdot \left(-\sum_{\mathscr{X}_+} \gamma_i x_i^{\mathrm{U}} - \sum_{\mathscr{X}_-} \beta_i x_i^{\mathrm{L}} - v \right) \\ & + \sum_{\mathscr{X}_+} \gamma_i + \sum_{\mathscr{X}_-} \beta_i \end{aligned} \tag{32}$$

and the substitution of KKT conditions into this expression, enables the Lagrangian function in (30) to be expressed as

$$L(w^+, w^-) = -\frac{1}{2}\|w^+ - w^-\|^2 + \sum_{\mathscr{X}_+} \gamma_i + \sum_{\mathscr{X}_-} \beta_i. \tag{33}$$

Finally, either of the two expressions derived from the KKT conditions (31) can be employed to calculate the weight vector $w = w^+ - w^-$. Without any loss of generality we can use the former expression. Weights $\alpha_i \triangleq \gamma_i$ for positive labels and $\alpha_i \triangleq \beta_i$ for negative labels are also defined. Hence, the dual QP problem becomes:

$$\begin{aligned} \min_{\alpha_i, \mu_j} \quad & \tfrac{1}{2}(\alpha', \mu') \cdot \mathbf{Q} \cdot \begin{pmatrix} \alpha \\ \mu \end{pmatrix} - (\mathbf{1}', \mathbf{0}') \cdot \begin{pmatrix} \alpha \\ \mu \end{pmatrix} \\ \text{s.t.} \quad & \alpha_i \geqslant 0, \quad i = 1, \ldots, n, \\ & \mu_j \geqslant 0, \quad j = 1, \ldots, m, \end{aligned} \tag{34}$$

where

$$\mathbf{Q} = \left(\begin{array}{cc|c} \mathbf{q}_1'\mathbf{q}_1 & \mathbf{q}_1'\mathbf{q}_2 & \mathbf{q}_1' \\ \mathbf{q}_2'\mathbf{q}_1 & \mathbf{q}_2'\mathbf{q}_2 & -\mathbf{q}_2' \\ \hline \mathbf{q}_1 & -\mathbf{q}_2 & \mathbf{I_m} \end{array} \right) = \left(\begin{array}{cc|c} \left(x_i^{\mathrm{L}} \cdot x_j^{\mathrm{L}}\right)_{\mathscr{X}_+, \mathscr{X}_+} & -\left(x_i^{\mathrm{L}} \cdot x_j^{\mathrm{U}}\right)_{\mathscr{X}_+, \mathscr{X}_-} & (x_i^{\mathrm{L}})'_{\mathscr{X}_+} \\ -\left(x_i^{\mathrm{U}} \cdot x_j^{\mathrm{L}}\right)_{\mathscr{X}_-, \mathscr{X}_+} & \left(x_i^{\mathrm{U}} \cdot x_j^{\mathrm{U}}\right)_{\mathscr{X}_-, \mathscr{X}_-} & (x_i^{\mathrm{U}})'_{\mathscr{X}_-} \\ \hline (x_i^{\mathrm{L}})'_{\mathscr{X}_+} & (x_i^{\mathrm{U}})'_{\mathscr{X}_-} & \mathbf{I_m} \end{array} \right) \tag{35}$$

provided that $\mathbf{q} = (\mathbf{q_1}, \mathbf{q_2}) = ((y_i \cdot x_i^{\mathrm{L}})_{\mathscr{X}_+}, (y_i \cdot x_i^{\mathrm{U}})_{\mathscr{X}_-})$.

Weight vector w solution can be written as

$$w = \mathbf{q} \cdot \alpha + \mu \tag{36}$$

and bias b is obtained as usual.

Remark 1. Weight vector $w = w^+ - w^-$ is calculated from the first KKT condition in (31). When using the second condition, the expressions obtained are similar if we exchange all over x_i^{L} with x_i^{U} and μ with v throughout the function to be minimized.

Remark 2. For the standard training data case, $x_i^{\mathrm{L}} = x_i^{\mathrm{U}}$, KKT conditions lead to $\mu = -v$, with $\mu, v \geqslant 0$, and hence $\mu = v = 0$. Therefore, $\mathbf{Q}$ reduces to the standard Gram matrix, terms with μ disappear from the cost function to be minimized, and constraints on μ also disappears. In this way, the dual QP problem (34) is the usual dual QP problem associated with a standard SVM.

Remark 3. The dual QP problem obtained from convex optimization theory can be reduced to that obtained from interval analysis by observing that,

$$\begin{aligned} & \sum_i y_i\left(\alpha_i x_i^{\mathrm{L}} + v_i\right) \\ & = \sum_{\mathscr{X}_+} y_i(\alpha_i x_i^{\mathrm{L}} + v_i) + \sum_{\mathscr{X}_-} y_i(\alpha_i x_i^{\mathrm{U}} - \mu_i) \\ & = \sum_{\mathscr{X}_+} y_i \alpha_i x_i^{\mathrm{L}} + \sum_{\mathscr{X}_+} y_i v_i + \sum_{\mathscr{X}_-} y_i \alpha_i x_i^{\mathrm{U}} - \sum_{\mathscr{X}_-} y_i \mu_i \\ & = \sum_{\mathscr{X}_+} y_i \alpha_i x_i^{\mathrm{L}} + \sum_{\mathscr{X}_-} y_i \alpha_i x_i^{\mathrm{U}} + \sum_{\mathscr{X}_+} v_i + \sum_{\mathscr{X}_-} \mu_i, \end{aligned} \tag{37}$$

hence it can be defined that

$$\mu = \sum_{\mathscr{X}_+} v_i + \sum_{\mathscr{X}_-} \mu_i \geqslant 0 \tag{38}$$

and the dual QP problem (34) is recovered.

An example of use is illustrated in Fig. 1. The two top graphics show the result of a trained I-SVM using the interval information; whereas the two lower graphics are trained considering only points, without interval information, with a standard SVM. I-SVM recovers standard SVM when interval information is not critical (left-hand side); however it uses the overall interval information (parameter μ in (36) is nonzero) when it is necessary (right-hand side). This modification has been obtained by moving the lower support vector to the left.

4.4. Soft I-SVM

The QP formulation in (28) ensures that multi-dimensional interval restrictions lie on the correct side of the separating hyperplane, as can be appreciated in the top two graphics of Fig. 1. In order to allow some imprecisions in the knowledge sets, some slack variables ξ_i, $i = 1, \ldots, n$, ζ_j, $j = 1, \ldots, m$ are added in a similar way to the soft formulation of standard SVM. Hence, the soft I-SVM

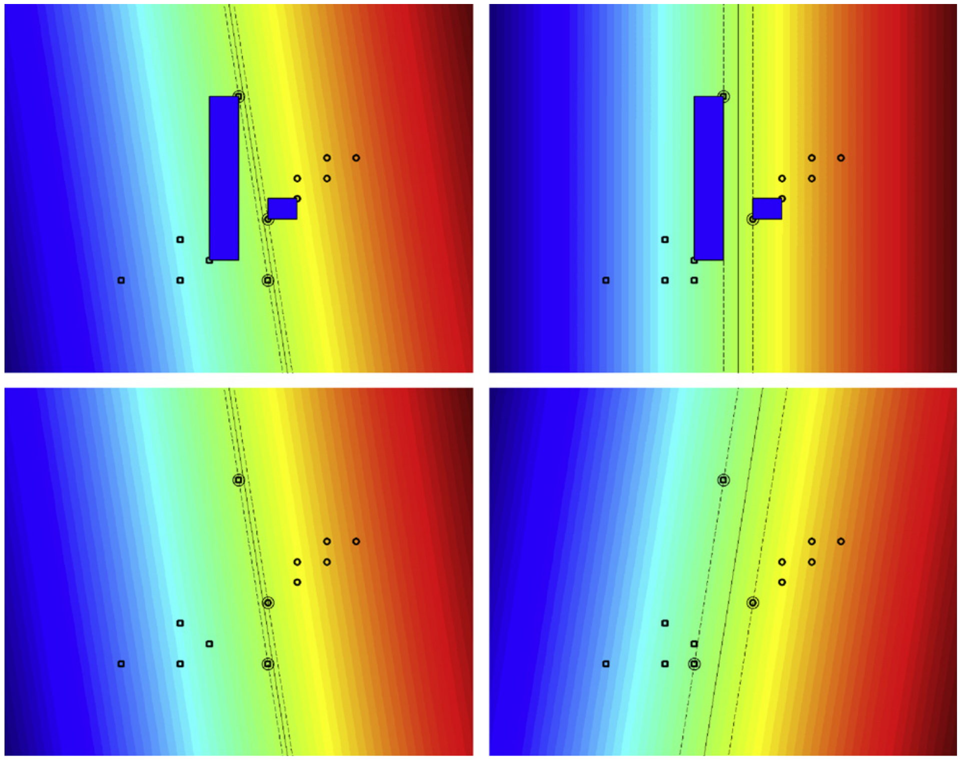


Fig. 1. The two top graphics illustrate trained SVM using the interval information (I-SVM), whereas the two lower graphics are trained with a standard SVM. I-SVM recovers standard SVM when interval information is not critical (left-hand side) and it uses the interval information (parameter μ is nonzero) when it is necessary (right-hand side).

becomes the primal QP problem,

$$\begin{aligned}\min_{w^+,w^-\in\mathbb{R}^m} \quad & \tfrac{1}{2}\|w^+ - w^-\|^2 + C\sum_{i=1}^{n}\xi_i + D\sum_{j=1}^{m}(\zeta_{j1}+\zeta_{j2})\\ & y_i\cdot((w^+)'\cdot x_i^{\mathrm{L}} - (w^-)'\cdot x_i^{\mathrm{U}})\geqslant 1-\xi_i, \quad y_i = 1,\\ \text{s.t.} \quad & y_i\cdot((w^+)'\cdot x_i^{\mathrm{U}} - (w^-)'\cdot x_i^{\mathrm{L}})\geqslant 1-\xi_i, \quad y_i=-1,\\ & w^+ + \zeta_1, \quad w^- + \zeta_2 \geqslant \mathbf{0}\end{aligned} \tag{39}$$

with $\zeta_k = (\zeta_{jk})_j$, $k=\{1,2\}$, $C,D\geqslant 0$ and $w = w^+ - w^-$ as the solution.

Using the Lagrangian function, similar to (30), the dual formulation can be obtained,

$$\begin{aligned}\min_{\alpha,\mu} \quad & \tfrac{1}{2}(\alpha)',\mu'\cdot\mathbf{Q}\cdot\begin{pmatrix}\alpha\\ \mu\end{pmatrix} - (\mathbf{1}',\mathbf{0})'\cdot\begin{pmatrix}\alpha\\ \mu\end{pmatrix}\\ \text{s.t.} \quad & C\geqslant\alpha_i\geqslant 0, \quad i=1,\ldots,n,\\ & D\geqslant\mu_j\geqslant 0, \quad j=1,\ldots,m\end{aligned} \tag{40}$$

with $\mathbf{Q}$ as defined in (35). The weight vector w solution can be written as

$$w = \mathbf{q}\cdot\alpha + \mu \tag{41}$$

and bias b is obtained as usual.

5. Complexity analysis and numerical testing

In order to analyse the complexity of the different approaches for completeness, an I-SVM can also be defined by translating any interval to its punctual vertices and using a standard SVM on these vertices. In this case, the resulting QP problem can be formulated as,

$$\begin{aligned}\min_{w\in\mathbb{R}^m} \quad & \tfrac{1}{2}\|w\|^2\\ & y_i\cdot w'\cdot x_i^j\geqslant 1, \quad i=1,\ldots,n, \; j=1,\ldots,2^m,\end{aligned} \tag{42}$$

where x_i^j is any of the 2^m vertices of the interval $I_i\in\mathscr{X}\mathscr{T}$. Hence, in this case, the formulation considers m parameters to be optimized, however $n2^m$ constraints must be accomplished by the optimization problem.

Summarized in Table 1. The results about the complexity of the different approaches are it can be agreed that the interval analysis approach (28) or (34) leads to the more balanced QP problem, which holds a low number of parameters to be optimized on a little-constrained optimization problem.

Numerical testing for accuracy comparison was carried out on two datasets from the UCI Repository [18], the Wisconsin Prognostic Breast Cancer dataset WPBC and the Promoter Recognition dataset, with the intention of comparing our results with those reported in [5] using the original full optimization approach.

Table 1
The complexity comparison between QP problems associated to the different approaches for the I-SVM

Approach	Parameters	Constraints	m, n	
Primal				
Convex (Eq. (11))	$(2n+1)m$	$(3m+1)n$	804	1300
Convex (Eq. (12))	$2nm$	$(2m+1)n$	800	900
Interval (Eq. (28))	$2m$	$n+2m$	8	108
Standard (Eq. (42))	m	$n2^m$	4	1600
Dual				
Convex (Eq. (17))	$(m+1)n$	$(m+1)n$	500	500
Interval (Eq. (34))	$n+m$	$n+m$	104	104
Standard	$n2^m$	$n2^m$	1600	1600

An illustrative example considers an $m=4$ dimension problem with $n=100$ training patterns.

For the WPBC dataset, a 60-month cutoff for the prediction of the recurrence or nonrecurrence of the disease was considered. The prior knowledge consisted of two prognosis rules used by doctors depending on the tumor size (T) and the lymph node status (L):

$$(L \geqslant 5) \wedge (T \geqslant 4) \Longrightarrow (R)ECUR,$$

$$(L = 0) \wedge (T \leqslant 1.9) \Longrightarrow (N)ONRECUR.$$

When the above rules are applied directly to the 110 given points (64 R, 46 N) of the training set, only 32 points (14 R, 18 N) are covered and correctly classify 23 (9 R, 14 N) of these 32 points. Hence, only an accuracy of 20% is obtained when rules are applied as a classifier. In [5] rules were converted into linear inequalities and used in their KSVM algorithm without any use of the data, which resulted in a linear classifier achieving 66.4% accuracy, more than the 66.2% correctness achieved by standard SVM using all the data. In our case, rules are converted into intervals, where the non-described bound (upper or lower) on the rules is assigned to the maximal or minimal value of the feature. The values of C and D associated with the soft I-SVM are obtained by a tuning procedure which consists of varying these values on a square grid: $\{2^{-6.0}, 2^{-5.9}, \ldots, 2^{6.0}\} \times \{2^{-6.0}, 2^{-5.9}, \ldots, 2^{6.0}\}$. In this way, a 70.0% accuracy is obtained without using training examples and with a simple linear classifier.

The second dataset, the Promoter Recognition dataset, is from the domain of DNA sequence analysis. This example includes 106 training points not matching 14 prior rules. These rules by themselves cannot serve as a classifier, but they do capture significant information about promoters, thus the effect of our assumptions on the training set will be completely tested in this way. These 14 prior rules were converted into 52 knowledge sets in a lightly different manner[4] that made in [5]. Following the methodology cited therein, we tested our algorithm using a 'leave-one-out' cross validation procedure, the values of C and D associated with the soft I-SVM being obtained by a similar tuning procedure to the precedent experiment.

[4]The reader is referred to Appendix Appendix A for more details.

Table 2
Comparison of I-SVM leave-one-out total error with classification algorithms reported in [5]

Method	# errors (out of 106)
KBANN	4
KSVM, I-SVM	5
BP	8
SVM_1	9
O'Neil	12
NN	13
ID3	19

The number of times that a test element is misclassified for each compared method reported in [5] and our method (I-SVM) is counted as an error and reported in Table 2. Note that our proposed algorithm, I-SVM, has a good performance behaviour among all the considered algorithms, and it is as accurate as the original full optimization approach, KSVM.

6. Conclusions

Imprecision in the input information, incompleteness in the patterns, discretization procedures, prior knowledge insertion or speed-up learning can motivate arriving interval represented data. Unlike existing SVM approaches working on interval data, a new formulation for a linear SVM classifier is derived from convex optimization theory, called I-SVM, which directly inserts information in the form of intervals. The new approach drastically reduces the complexity of the original convex-based approach, and can be interpreted as based on interval arithmetic. Numerical testing has been carried out on two datasets from the UCI Repository, the Wisconsin Prognostic Breast Cancer dataset WPBC, with prior knowledge of two prognosis rules used by doctors, and the Promoter Recognition dataset, including training points not matching prior rules. Similar or higher accuracy than precedent works has been obtained using an I-SVM simple linear classifier. A future line of research is to address nonlinear classifiers determined by nonlinear kernels.

Acknowledgements

We would like to thank the anonymous reviewers and editors for their useful comments and helpful advice when preparing the final manuscript.

Appendix A. About the promoter recognition dataset

When the experimentation in [5] for the DNA sequence analysis Promoter Recognition dataset was replicated by using the rules proposed in [4], some inconsistency between rules R_6, R_7 and R_8 with R_{10} were found. Since inconsistent

$$\begin{array}{ccccccc} M35: & & M10: & & CONF: & & \\ \begin{bmatrix} R_1 \\ or \\ R_2 \\ or \\ R_3 \\ or \\ R_4 \end{bmatrix} & \wedge & \begin{bmatrix} R_5 \\ or \\ R_6 \\ or \\ R_7 \\ or \\ R_8 \end{bmatrix} & \wedge & \begin{bmatrix} R_9 \\ or \\ R_{10} \\ or \\ R_{11} \\ or \\ R_{12} \end{bmatrix} & \Longrightarrow & PROMOTER \\ 4 & \times & 4 & \times & 4 & = & \text{64 Rules.} \end{array}$$

Fig. A1. Converting rules in a matrix of inequalities in [4].

Table 3: A Domain Theory For Promoters

promoter	:- contact, conformation.
contact	:- minus_35, minus_10.
minus_35	:- @-37 "cttgac".
minus_35	:- @-36 "ttgxca".
minus_35	:- @-36 "ttgaca".
minus_35	:- @-36 "ttgac".
minus_10	:- @-14 "tataat".
minus_10	:- @-13 "taxaxt".
minus_10	:- @-13 "tataat".
minus_10	:- @-12 "taxxxt".
conformation	:- @-45 "aaxxa".
conformation	:- @-45 "axxxa", @-4 "t", @-28 "txxxtxaaxxtx".
conformation	:- @-49 "axxxxt", @-1 "a", @-27 "txxxxaxxtxtg".
conformation	:- @-47 "caaxttxac", @-22 "gxxxtxc", @-8 "gcgccxcc".

$$R_{11}: \begin{array}{l} (p_{-49} = A) \wedge (p_{-44} = T) \wedge (p_{-27} = T) \wedge \\ (p_{-22} = T) \wedge (p_{-18} = T) \wedge (p_{-16} = A) \wedge \\ (p_{-15} = C) \wedge (p_{-01} = C) \end{array} \Longrightarrow CONF$$

Fig. A2. Left: Table 3 from [26]. Right: R_{11} according to [4].

rules R_6, R_7 and R_8 with R_{10} were obviated and knowledge is derivated following the combination of rules in Fig. A1, final knowledge sets were reduced from 64 to 52.

Moreover, according to article [26] whose results were reported for comparison in [5], it can be found that the third rule for *conformation*(Fig. A2(left)) is not exactly the same that of rule R_{11} in [4] (Fig. A2(right)).

In our experimentation, R_{11} was modified in the sense of [26]. The 106 training points from the dataset were maintained. As a last modification, we used a real-valued encoding to represent the nucleotides $\{A, G, C, T\}$ instead of binary encoding. Considering these modifications, we obtained 5 errors out of 106 for our algorithm.

References

[1] H.H. Bock, Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, Springer, New York, Secaucus, NJ, USA, 2000.

[2] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, Cambridge, 2000.

[3] T.-N. Do, F. Poulet, Kernel methods and visualization for interval data mining, in: Proceedings of the International Symposium on Applied Stochastic Models and Data Analysis, Brest, France, 2005, pp. 345–354.

[4] G. Fung, O.L. Mangasarian, J. Shavlik, Knowledge-based support vector machine classifiers, Technical Report 01-09, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, November 2001.

[5] G.M. Fung, O.L. Mangasarian, J.W. Shavlik, Knowledge-based support vector machine classifiers, in: S.T.S. Becker, K. Obermayer (Eds.), Advances in Neural Information Processing Systems, vol. 15, MIT Press, Cambridge, MA, 2003, pp. 521–528.

[6] L. González, F. Velasco, F. Cuberos, J.A. Ortega, C. Angulo, A kernel to use with a discretization of continuous features, in: Proceedings of Learning'04, Elche, Spain, 2004.

[7] L. González, C. Angulo, F. Velasco, A. Catala, Dual unification of bi-class support vector machine formulations, Pattern Recognition 39 (7) (2006) 1325–1332.

[8] L. Gonzalez-Abril, C. Angulo, F. Velasco, J. Ortega, A note on the bias in SVMs for multi-classification, IEEE Trans. on Neural Networks, accepted for publication, doi:10.1109/TNN.2007.914138.

[9] E. Hansen, Global Optimization using Interval Analysis, Marcel Dekker, New York, 1992.

[10] R. Hebrich, Learning Kernel Classifiers: Theory and Algorithms, MIT Press, Cambridge, MA, 2002.

[11] D.H. Hong, C. Hwang, Interval regression analysis using quadratic loss support vector machine, IEEE Trans. Fuzzy Systems 13 (2) (2005) 229–237.

[12] C. Hwang, D.H. Hong, K.H. Seok, Support vector interval regression machine for crisp input and output data, Fuzzy Sets and Systems 157 (8) (2006) 1114–1125.

[13] L. Jaulin, M. Kieffer, O. Didrit, E. Walter, Applied Interval Analysis, Springer, London, 2001.

[14] V. Kreinovich, F. Modave, S.A. Starks, G. Xiang, Towards real world applications: interval-related talks at nafips'05, Reliable Comput. 12 (1) (2006) 73–77.

[15] O.L. Mangasarian, Nonlinear Programming, SIAM, Philadelphia, PA, 1994.

[16] T. Martinetz, K. Schulten, A "neural gas" network learns topologies, Artificial Neural Networks, Elsevier, Amsterdam, 1991, pp. 397–402.

[17] R.E. Moore, Interval Analysis, Prentice-Hall, Englewood Cliffs, 1966.

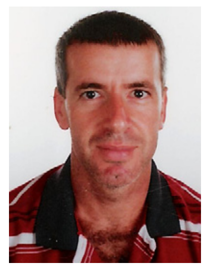
[18] C.B.D.J. Newman, S. Hettich, C. Merz, UCI repository of machine learning databases ⟨http://www.ics.uci.edu/~mlearn/MLRepository.html⟩.

[19] P. Nivlet, F. Fournier, J. Royer, Interval discriminant analysis: an efficient method to integrate errors in supervised pattern recognition, in: 2nd International Symposium on Imprecise Probabilities and their Applications, 2001.

[20] Y. Ohta, Nonconvex polygon interval arithmetic as a tool for the analysis and design of robust control systems, Reliable Comput. 6 (2000) 247–279.
[21] F. Palumbo, Editorial, Comput. Stat. 21 (2) (2006) 183–185.
[22] S. Pikorski, L. Lacassagne, M. Kieffer, D. Etiemble, Efficient 16-bit floating point interval processor for embedded systems and applications, in: 12th GAMM-IMACS International Symposium on Scientific Computing, Computer Arithmetic and Validated Numerics, 2006.
[23] X. Rovira, N. Agell, M. Sánchez, F. Prats, X. Parra, An approach to qualitative radial basis function networks over orders of magnitude, in: Proceedings of the 18th International Workshop on Qualitative Reasoning (QR'04), Evanston, Illinois, USA, 2004.
[24] F.J. Ruiz, N. Agell, C. Angulo, A kernel intersection defined on intervals, in: Recent Advances in Artificial Intelligence Research and Development, vol. 113, in: Frontiers in Artificial Intelligence and Applications, IOS Press, 2004, pp. 103–110.
[25] H. Tanaka, H. Lee, Interval regression analysis by quadratic programming approach, IEEE Trans. Fuzzy Systems 6 (4) (1998) 473–481.
[26] G.G. Towell, J.W. Shavlik, M.O. Noordenier, Refinement of Approximate Domain Theories by Knowledge Based Neural Network, vol. 2, 1990, pp. 861–866.
[27] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[28] Y. Zhao, Q. Chen, Q. He, An interval set classification based on support vector machines, in: Proceedings of the Joint International Conference on Autonomic and Autonomous Systems 2005/International Conference on Networking and Services 2005 (ICAS/ICNS 2005), IEEE Computer Society, Papeete, Tahiti, 2005, pp. 81–86.

Cecilio Angulo (Barcelona, Spain, 1969) received his M.Sc. degree in Mathematics in 1993 from the University of Barcelona and the Ph.D. degree in 2001 from the Technical University of Catalonia. He is currently an Associate Professor at the Technical University of Catalonia and a member of the Knowledge Engineering Research Group, GREC. His research interests include machine learning, intelligent control and ambient intelligence.

Davide Anguita (Genova, Italy, 1963) obtained the 'Laurea' degree in Electronic Engineering in 1989 and the Ph.D. in Electronic Engineering and Computer Science in 1994 from University of Genova. From 1989 to 1991 he worked at Esacontrol SPA in the field of distributed systems for traffic control. From 1993 to 1994 he was postdoc research assistant at ICSI http://www.icsi.berkeley.edu (Berkeley, USA), working on modelling and simulation of artificial neural networks. Since 1996 he is Assistant Professor at DIBE (Department of Biophysical and Electronic Engineering—Faculty of Engineering—University of Genova) working in the Electronic Systems and Networking Group. His current research activity focuses on Smart Electronic Systems.



Luis Gonzalez-Abril is a Lecturer in Department of Applied Economics I in the University of Seville. He obtained the M.Sc. degree in Mathematics and the Ph.D. degree in Economics from the University of Seville. His research lines are about learning machines and bifurcations in dynamical systems with applications in economic problems.

Juan Antonio Ortega (Conil, Spain, 1968) received his Ph.D. in Computer Engineering. Coordinator of the Research Group ARCA about Qualitative reasoning and its applications, from December 2004 he is the Head Director of the CICA (Scientific Computing Centre in Andalusia). He has participated and is responsible of a number of research projects in the area of Computing Systems, in particular, in diagnosis and qualitative reasoning, hypermedia systems, information systems and web-based ubiquitous systems.