Support Vector Machine for Interval-valued Data

Rui Malha[1*] and Paula Amaral[1,2]

[1*]Dep. Mathematics, University Nova de Lisboa, Portugal.
[2]Nova Math, University Nova de Lisboa, Campus de Caparica, Caparica, 2829-516, Portugal.

*Corresponding author(s). E-mail(s): rj.malha@campus.fct.unl.pt;
Contributing authors: paca@fct.unl.pt;

Abstract

In this work we propose a generalization of the Spherical Support Vector Machine method, in which the separator is a sphere, applied to Interval-valued data. This type of data belongs to a more general class, known as Symbolic Data, for which features are described by sets, intervals or histograms instead of classic arrays. This paradigm is raising interest in our days, specially in the context of Big Data. On the other end, SVM is a classic and well studied method for classification, and in the classical approach the separation is defined by an hyperplane. Generalizations of this concept have merged, with the use of non-linearity defined by Kernel functions. As an alternative to Kernell functions some authors have proposed non linear separation functions, in particular spherical separations, with the advantage of keeping the classification in the feature space. In this paper we present a quadratic optimization model for the spherical SVM for interval data and develop a relaxation linear problem. The performance of these two formulations for classification purposes is tested against standard methods for a collection of data sets.

Keywords: SVM; Interval-valued Data; Symbolic Data; SVM; Spherical separation

Acknowledgements. This work is funded by national funds through the FCT (Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UIDB/00297/2020 (https://doi.org/10.54499/UIDB/00297/2020) and UIDP/00297/2020 (https://doi.org/10.54499/UIDP/00297/2020) (Center for Mathematics and Applications)

1 Introduction

In the contemporary era of Big Data, the sheer volume, velocity, and variety of data being generated are unprecedented. This explosion of data, originating from numerous sources such as social media, sensors, transaction records, and more, poses significant challenges in terms of storage, processing, and, crucially, interpretation. In this landscape, Machine Learning (ML) methods are essential to derive meaningful information from this vast sea of data. They enable organizations to transform raw data into valuable insights, driving innovation and efficiency across various domains. As data continues to grow in complexity and scale, the development and application of robust ML techniques will remain a cornerstone of data science and analytics.

In particular, Automatic Classification methods help in sorting through heterogeneous data, enabling the understanding and identifications of inner structures in the data, that allows to characterize different classes, identify relevant patterns and trends, that might otherwise remain hidden. This facilitates more effective analysis, informed decision-making processes and strategic planning. In business, healthcare, finance, and various other fields, rapid and accurate decision-making is crucial.

One such method, the Support Vector Machine (SVM), has demonstrated considerable success in data classification tasks, either employed directly or serving as a foundational approach for the creation of related methodologies [1–8]. The SVM algorithm is valued for its ability to handle high-dimensional spaces. Numerous enhancements and generalizations of the SVM boundary function to improve its performance [9] and adapt it to various application domains has been proposed over the years. These innovations have expanded the applicability of SVMs, making them a crucial tool in the analysis of complex data sets generated by modern monitoring systems, such as sensors, which are widespread in agriculture, economics, security, and industry in general. Interpreting and classifying this type of data requires adaptation of previous methods.

Consider for instance that we have data collected from sensors for several units (machines, people, cities, etc) and we want to create a classification model that will allow to differentiate between the malfunction and operational units. The malfunction state can be interpreted in a broad sense, as anything in opposition to one or more expected states, that with an abuse of language can be defined as "normality". It can be a malfunction of a device, a disease, a fraud, and so on. We may consider to use the full array of data, for each unit, which is not a good option for several reasons. First we cannot define as a feature the i-th reading of the sensor. A feature is a measurement that has an unique and equal meaning for all the examples and that can be compared among them: price, weight, height, pressure, temperature. If our study has a chronological component and each measurement for all units is taken in the same period of time, then it may be meaningful to consider a feature that is the i-th reading of the units regarding some period of time. For instance, it can be the reading in the i-th day, the i-th week, or the i-th year. If not, then it is meaningless and can conduct to wrong interpretations. In addition we may have different numbers of readings for

each unit, their number may be too high as it is the case of sensors data. In these situations it is common the use, for unit $j \in N$, the average $\bar{x_j}$ instead of the original vector of data $x_j = (x_{1j}, \ldots, x_{Lj})$. Replacing the vector of data by a single measure of central tendency may work well in some cases, for instance, if what differentiates a good from a malfunction behavior in our devices is a shift in the average. However, different distributions of data can be an indication of a malfunction without affecting the mean as it is the case in Figure 1.

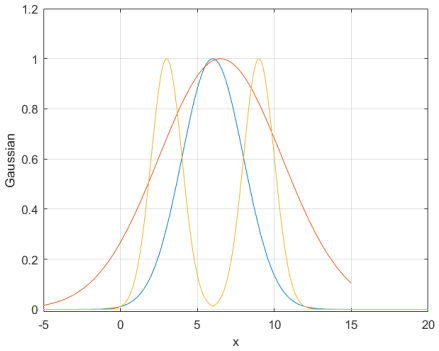


Fig. 1: Distribution of data.

In some cases the malfunction may induce a shift of values away from the average in both sides (increase of the variance), or values that are just low or high (bimodal distribution). In the first case, using the mean will not be effective to interpret a malfunction and if using the mean and variance may work for this case, in the bi-modal pattern malfunction case, that will not be enough.

Examples of bimodal distribution as an indicator of a clear differentiation between two distinct classes can be found in many contexts. For instance, a two peak in an age distribution of a cancer is an important distinction, often implying that what was thought as one type of cancer can actually be two different cancers, with similar morphology but with distinctive clinical features and different methods of treatment [10]. Another example arises in the context of spam and co-bursting detection [11], where reviewers' posting rates (number of reviews written in a period of time) in fake or spam reviews to secretly promote or demote some target products and services, are bimodal. In the scope of fraud detection, bimodal distributions are often and indication of fraud in exams or elections [12]. In [13] detection of climate changes using bimodal distributions is also reported.

Considering that the features of distinction between classes are not known *apriori*, methods that take into account as much information as possible about the distribution of the data are important. These methods should be able to receive as input

variables more complex than arrays, data with structures, such as intervals or histograms, as depicted in Figure 2, known as *Symbolic Data*. General methods, with

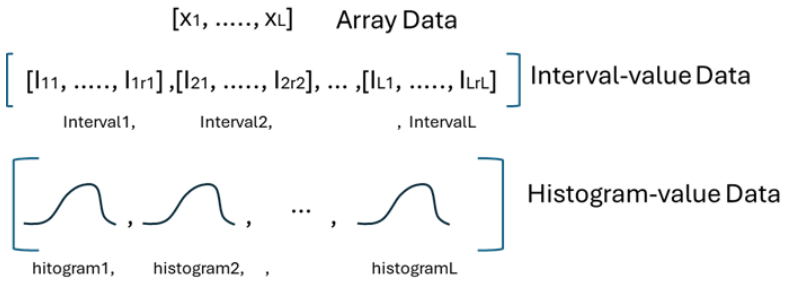


Fig. 2: Types of data.

a large spectrum of applicability, capable of performing automatic classification also for this type of data was our motivation in this work.

Symbolic Data is explained in detail in section 2 together with a review on the state of the art in symbolic data analysis. The remaining organization of the paper consists in the presentation of the classical and spherical SVM and its extension to interval valued data in section 3. Computational experience is reported in section 4 and conclusions are presented in the last section.

2 Symbolic Data

2.1 Basic concepts

Symbolic Data was proposed by Diday [14] to extend classic variables to more complex data, closer to reality, such as "events", "assertions","hordes" and "synthesis" objects. For instance in the study of birds, color is a variable of interest and different birds can be characterized by different conjunctions of colors. The concept of symbolic variables evolved to incorporate data defined by intervals and more recently by distributions-histograms. Symbolic data can be represented in complex tables where in each cell we may find not only values and categories but also sets, intervals or distributions. In these tables the rows refer to entities/units and columns are the corresponding features. We may have first or high order level units, depending if we are considering in the first case objects or individuals or in the former, classes. The features may be qualitative or quantitative, which in turn can be represented by single values (single-valued variables), sets of values (multi-valued variables), intervals (interval-valued variables) or probability/frequency/weight distributions (histogram-valued variables).

For instance, the units may be beaches as in Table 1 and the register data can be interval-valued (Temperature of water), qualitative (Life watcher), a set of values (Facilities), or histogram-valued (Number of visitors).

Table 1: Symbolic data example

Beaches	Temperature of water	Life watcher	Number of facilities	Number of visitors
A	[14,22]	No	3	{[0,50], 0.2; [51,100], 0.4;[101,200], 0.3;[201,500], 0.1;}
B	[13,25]	Yes	2	{[0,50], 0.1; [51,100], 0.35;[101,200], 0.35;[201,500], 0.2;}
C	[16,28]	Yes	4	{[0,100], 0.3; [101,200], 0.5;[201,500], 0.2;}

In this work we are particularly interested in variables in the form of intervals. As future work we intend to develop similar approaches to histogram-value data. Note that an interval can be considered an histogram with one class and frequency equal to 1, and in that sense histogram-valued variables are a generalization of interval-valued ones. One obvious importance of interval data in the era of Big Data, and that explains the increasing interest towards this subject, is related to the ability to summarize large sets of data from databases, the web or generated in streaming by sensors. Reducing this array of data to just the average, implies losing information, so statistical methods were developed to extend classic statistical analysis to symbolic variables, creating a new area of research defined as Symbolic Data Analysis (SDA).

2.2 Algebric operations with Histograms and Quantile functions

The development of methods for symbolic data requires the definition of algebraic operations and notions of distance between elements. In this section we will address arithmetic operations between histograms. We will discuss the difficulties in using such arithmetic and introduce quantile functions as a more convenient representation of histograms, simplifying the definition of these operations.
The outcome associated with a histogram-valued variable, may take the following form:

$$H_X = \{[\underline{l}_{X_1}, \bar{l}_{X_1}[, p_1; ...; [\underline{l}_{X_m}, \bar{l}_{X_m}[, p_m\} \tag{1}$$

where $\underline{l}_{X_i} \leq \bar{l}_{X_i}; \bar{l}_{X_i} \leq \underline{l}_{X_{i+1}}$ and $\sum_{i=1}^{m} p_i = 1$.

Example 1 Considering two histograms,

$$H_X = \{[1, 3[, 0.1; [3, 5[, 0.6; [5, 8[, 0.3\}$$
$$H_Y = \{[0, 1[, 0.8; [1, 4[, 0.2\}$$

We have the following operations between these two histograms:

$$\begin{aligned} H_X + H_Y =& \{[1, 2[, 0.0267; [2, 3[, 0.0307; [3, 4[, 0.1907; [4, 5[, 0.18807; [5, 6[, 0.2480; \\ & [6, 7[, 0.0980; [7, 9[, 0.1880; [9, 12[, 0.0300\} \\ H_X + 2 =& \{[3, 5[, 0.1; [5, 7[, 0.6; [7, 10[, 0.3\} \\ H_X - 2 =& \{[-1, 1[, 0.1; [1, 3[, 0.6; [3, 6[, 0.3\} \end{aligned}$$

$$
\begin{aligned}
2H_X &= \{[2,6[,0.1;[6,10[,0.6;[10,16[,0.3\} \\
-H_X &= \{[-8,-5[,0.3;[-5,-3[,0.6;[-3,-1[,0.1\} \\
H_X - H_X &= \{[-7,-5[,0.0120;[-5,-4[,0.0420;[-4,-3[,0.0570;[-3,-2[,0.0720;[-2,0[,0.3170; \\
&\quad [0,2[,0.3170;[2,3[,0.0720;[3,4[,0.0570;[4,5[,0.0420;[5,7[,0.0120\}.
\end{aligned}
$$

For instance, to construct $H_X + H_Y$ we must first consider the possible combination of the pairs corresponding to the limits of H_X and H_Y, respectively $\{(1,3),(3,5),(5,8)\}$ and $\{(0,1),(1,4)\}$, and for each sub-interval, assuming a uniform distribution, we assign the weight given by the product of the weights of each interval. Next we must divide that weight by the number of unitary segments contained in that interval, and obtain the weight of each unitary range sub-interval. For $[1,3[+[0,1[=[1,4[$, with weight $0,1 \times 0.8 = 0.08$, each unitary range interval has the weight $0.08/3 = 0,0267$. For $[1,3[+[1,4[=[2,7[$ the weight is $0.1 \times 0.2 = 0.02$ and for each unit range interval we obtain $0.02/5 = 0.004$. For $[3,5[+[0,1[=[3,6[$, for each unitary sub-interval we have the weight 0.16. Repeating the reasoning, for $[3,5[+[1,4[=[4,9[$ we have $\frac{0.6\times0.2}{5} = 0.024$. For $[5,8[+[0,1[=[6,9[$ and $[5,8[+[1,4[=[6,12[$ the unitary weights are respectively, $\frac{0.3\times0.2}{3} = 0.06$ and $\frac{0.3\times0.2}{6} = 0.01$. Next for each sub-interval we must sum all the weights assign to that sub-interval. For instance, for $[5,6[$ the weight is given by $0.16 + 0.06 + 0.004 + 0.024 = 0248$ and for the interval $[9,12[$ the weight is $3 \times 0.01 = 0.03$.

These operations are not suited to be used in an algorithm that implies complex operations and steps. Besides they present undesirable properties as for instance the fact that $\frac{H+H}{2} \neq H$. As alternative, quantile functions were proposed [15], [16], to represent histograms with a much simpler system of basic operations. A quantile function is a piece-wise function given by the inverse of the cumulative function.

$$
\Psi_X^{-1}(t) = \begin{cases}
\underline{l}_{X_1} + \frac{(t-w_0)}{w_1-w_0}\left(\bar{l}_{X_1} - \underline{l}_{X_1}\right) & \text{if } w_0 \leq t < w_1 \\
\underline{l}_{X_2} + \frac{(t-w_1)}{w_2-w_1}\left(\bar{l}_{X_2} - \underline{l}_{X_2}\right) & \text{if } w_1 \leq t < w_2 \\
\vdots & \\
\underline{l}_{X_m} + \frac{(t-w_{m-1})}{w_m-w_{m-1}}\left(\bar{l}_{X_m} - \underline{l}_{X_m}\right) & \text{if } w_{m-1} \leq t < w_m \\
\vdots &
\end{cases}
$$

where $w_0 = 0$ and $w_j = \sum_{i=1}^{j} p_i$, so $w_m = 1$.

Using the centers $c_{X_i} = \frac{\underline{l}_{X_i} + \bar{l}_{X_i}}{2}$, and half rays $r_{X_i} = \frac{\underline{l}_{X_i} - \bar{l}_{X_i}}{2}$, to define the histogram

we have:

$$\Psi_X^{-1}(t) = \begin{cases} c_{X_1} + \left(\frac{2(t-w_0)}{w_1-w_0} - 1\right) \times r_{X_1} & \text{if } w_0 \leq t < w_1 \\ c_{X_2} + \left(\frac{2(t-w_1)}{w_2-w_1} - 1\right) \times r_{X_2} & \text{if } w_1 \leq t < w_2 \\ \vdots & \\ c_{X_m} + \left(\frac{2(t-w_{m-1})}{w_m-w_{m-1}} - 1\right) \times r_{X_m} & \text{if } w_{m-1} \leq t \leq w_m \end{cases}$$

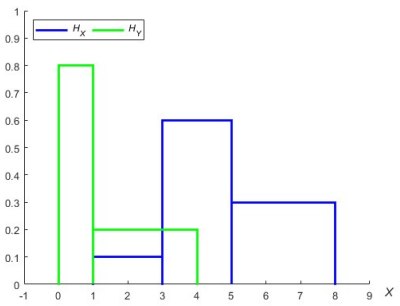
Given the histograms in Example 1, the cumulative functions are given by:

$$\Psi_X(x) = \begin{cases} 0.1\frac{x-1}{2} & \text{if } 1 \leq x < 3 \\ 0.1 + 0.6\dfrac{x-3}{2} & \text{if } 3 \leq x < 5 \\ 0.7 + 0.1\dfrac{x-5}{3} & \text{if } 5 \leq x \leq 8 \end{cases}, \quad \Psi_Y(x) = \begin{cases} 0.8x & \text{if } 0 \leq x < 1 \\ 0.8 + 0.2\dfrac{x-1}{3} & \text{if } 1 \leq x \leq 4 \end{cases}$$

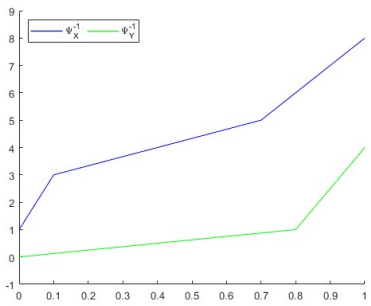
and their inverse the quantile functions are,

$$\Psi_X^{-1}(t) = \begin{cases} 1 + \dfrac{t}{0.1} \times 2 & \text{if } 0 \leq t < 0.1 \\ 3 + \dfrac{t-0.1}{0.6} \times 2 & \text{if } 0.1 \leq t < 0.7 \\ 5 + \dfrac{t-0.7}{0.3} \times 3 & \text{if } 0.7 \leq t \leq 1 \end{cases}, \Psi_Y^{-1}(t) = \begin{cases} \dfrac{t}{0.8} & \text{if } 0 \leq t < 0.8 \\ 1 + \dfrac{t-0.8}{0.2} \times 3 & \text{if } 0.8 \leq t \leq 1 \end{cases}$$

These histograms and quantile functions are represented in Figure 3.



Histograms H_X and H_Y



Quantile functions Ψ_X^{-1}, Ψ_Y^{-1}

Fig. 3: Histograms and quantile functions of Example 1

For quantile functions the basic operations of sum and scalar product, are defined as:

- Addition: $\Psi_X^{-1}(t) + \Psi_Y^{-1}(t) = (\Psi_X^{-1} + \Psi_Y^{-1})(t)$
- Product by a real number: $\alpha\Psi_X^{-1}(t) = (\alpha\Psi_X^{-1})(t)$

We must note however that the set of quantile functions with the above operations is not a subspace. In fact if $\alpha < 0$ then $\alpha\Psi_X^{-1}$ is not a quantile function. To allow for general linear combinations of quantile functions it was proposed [17] to use a quantile function given by $-\Psi_X^{-1}(1-t)$ representing $-H$ the symmetric of histogram H.

2.3 Distance measures

Most of the methods for supervised or unsupervised classification require the definition of a distance metric between objects. In the case of histogram value data, for which intervals are a particular case, several measures were proposed in the literature to define the distance between two quantile functions (see e.g. [18]).

Table 2: Distances between functions

Divergency measures	
Kullback-Leibler	$D_{KL}(f,g) = \int_{\mathbb{R}} log\left(\frac{f(x)}{g(x)}\right) f(x)dx$
Jeffreys	$D_J(f,g) = D_{KL}(f,g) + D_{KL}(g,f)$
χ^2	$D_{\chi^2}(f,g) = \int_{\mathbb{R}} \frac{\lvert f(x)-g(x)\rvert^2}{g(x)}dx$
Hellinger	$D_H(f,g) = \left[\int_{\mathbb{R}} \left(\sqrt{f(x)} - \sqrt{g(x)}\right) dx\right]^{\frac{1}{2}}$
Total variation	$D_{var}(f,g) = \int_{\mathbb{R}} \lvert f(x) - g(x)\rvert dx$
Kolmogorov	$D_K(f,g) = \max_{\mathbb{R}} \lvert F(x) - G(x)\rvert$
Wasserstein	$D_W(f,g) = \int_0^1 \lvert F^{-1}(t) - G^{-1}(t)\rvert dt$
Mallows	$D_M(f,g) = \sqrt{\int_0^1 (F^{-1}(t) - G^{-1}(t))^2\, dt}$

Of the several distances displayed in Table 2, the Mallows distance has been considered as an adequate measure to evaluate the similarity between distributions.

$$D_M\left(\Psi_X^{-1}, \Psi_Y^{-1}\right) = \sqrt{\int_0^1 (\Psi_X^{-1}(t) - \Psi_Y^{-1}(t))^2 dt} \tag{2}$$

This distance has been successfully used in cluster analysis for histogram data [16], in forecasting histogram time series [15], and in linear regression with histogram/interval-valued variables [15], [19], [20], [21].

For histogram-valued data, under the uniformity hypothesis, and considering a fixed weight decomposition (same weights for the different intervals), the formula in (2) can be rewritten (see [16]) using the centers and half rays to define the classes of the histograms as:

$$D_M^2(\Psi_X^{-1}, \Psi_Y^{-1}) = \sum_{i=1}^{m} p_i \left[(c_{X(i)} - c_{Y(i)})^2 + \frac{1}{3}(r_{X(i)} - r_{Y(i)})^2 \right] \tag{3}$$

2.4 State of the art

In general, most existing published research articles related to interval-valued data mainly focuses on clustering analysis, regression analysis and feature selection, and less on classification tasks.
The first proposals of basic univariate and bivariate statistics for histogram-valued data were due to Bertrand and Goupil [22] and later integrated and extended by Billard and Diday [23].
Irpino and Verde [19] proposed a novel set of univariate and bivariate statistics that better take into account the sources of variability in data and extend some properties of the classic basic statistics to those for multi-valued numeric data. In the context of discriminant analysis Silva and Brito [24] made contribution for interval-valued data and Brito and Dias [20] and Dias, Brito and Amaral [21] for histogram-valued variables. In [25] it was presented a new approach for constructing regression and classification models for interval-valued data including the extension of the support vector machine method for interval-valued data. In [17] the author proposed a linear regression models for histogram-valued data and interval-valued data represented by their quantile functions. An one-class classification support vector machine model by interval-valued training data was presented in [26]. An interval-valued data classification method based on the Unified Representation Frame, which takes in account not only the mid point and radius of the interval-valued data but also the relationship between them was proposed in [27]. Recently [28] developed a method for interval-valued data classification based on multi-view learning which is a machine learning paradigm that deals with learning from multiple data representations (views) of the same underlying information. The aim of the proposed algorithm is to classify multi-view information extracted from the interval-valued observations, using SVM, random forests and neural networks, as the basic structures of the multi-view classifier.

In our work we are going to present a new method for interval classification using a SVM-type approach based on a spherical separation.

3 A Spherical SVM for Interval Classification

In this section it is presented the Spherical SVM for interval data after a short introductory review of this approach for classical data as proposed in [9].

3.1 The SVM and SSVM problems

SVM is a supervised classification method. It uses two sets of labeled data, one (the training examples) to find a frontier that separates two or more classes of data, and the other (the testing examples) to test the accuracy of the classification. It may use a third set for validation, but this process is not always necessary. In the classical SVM the frontier is defined by an hyper-plane as in Figure 4. The frontier is defined so that the examples of the separate categories are divided by a clear gap (the margin in Figure 4) that is as wide as possible. New points are automatically predicted to belong to a category based on the side of the gap on which they fall. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

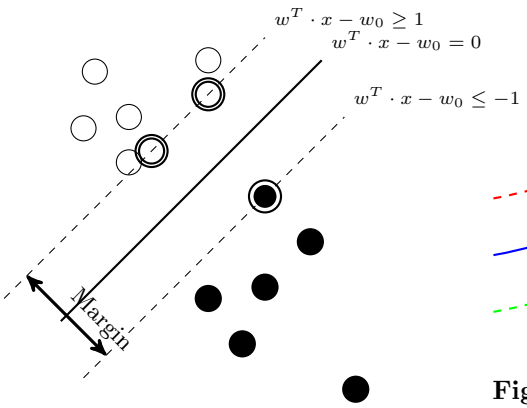


Fig. 4: SVM in $\mathbb{R}^2$ - two features

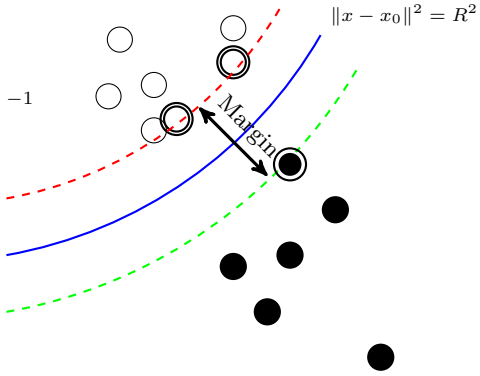


Fig. 5: SSVM in $\mathbb{R}^2$ - two features

Given a set of n training examples (x_j, y_j), with $x_j \in \mathbb{R}^L$ and $y_j \in \{-1, 1\}$, for $j \in N = \{1, \ldots, n\}$, corresponding to two classes (one class with $y_j = 1$ for $j \in N_1$ and the other with $y_j = -1$ for $j \in N_2$), SVM aims to find an hyperplane defined by a vector w that maximizes the margin and such that the points from distinct classes lie on opposite sides of the hyperplane. For that we need to solve the following optimization problem:

$$SVM: \; v = \max_{j \in N} \min_{w} D_j \tag{4}$$

$$s.t. \; w^T x_j - w_0 \geq 1 \text{ for } j \in N_1 \tag{5}$$

$$w^T x_j - w_0 \leq -1 \text{ for } j \in N_2 \tag{6}$$

$$D_j = \frac{|w^T x_j - w_0|}{\|w\|} \text{ for } j \in N. \tag{7}$$

This maxmin problem can be reformulated as

$$SVM: \ v = \max \delta \tag{8}$$
$$s.t. \ w_1 x_j^1 + w_2 x_j^2 - w_0 \geq 1 \text{ if } y_j = 1 \tag{9}$$
$$w_1 x_j^1 + w_2 x_j^2 - w_0 \leq -1 \text{ if } y_j = -1 \tag{10}$$
$$\frac{|w_1 x_j^1 + w_2 x_j^2 - w_0|}{\|w\|} \geq \delta \text{ for } j \in N \tag{11}$$

which is equivalent to the well known SVM problem formulation

$$SVM: \ v = \min \|w\| \tag{12}$$
$$s.t. \ y_j(w^T x_j - w_0) \geq 1 \text{ for } j \in N. \tag{13}$$

The points, in each class for which the margin in Figure 4 is attained are the support vectors and are signaled in the figure.
As an alternative to the linear mapping a spherical separation was proposed in [9] as depicted in Figure 5.

$$SSVM: \ v_0 = \max \delta \tag{14}$$
$$s.t. \ \|x_j - x_0\|^2 \leq (R - \delta)^2, \ \text{ if } j \in N_1 \tag{15}$$
$$\|x_j - x_0\|^2 \geq (R + \delta)^2, \ \text{ if } j \in N_2 \tag{16}$$
$$R \geq \delta \tag{17}$$
$$\delta \geq 0 \, , x_0 \in \mathbb{R}^L.$$

The parameters of the model are now the center x_0 and the radius R of the sphere and the goal is to maximize the margin δ. This model SSVM can be seen as a generalization of the linear case SVM, because when $R \to +\infty$, locally the spherical surface approaches a line. For classification purposes a new point is classified as Class 1 if it lies inside the sphere and as Class 2 if it lies outside. While the linear separation problem is symmetric (irrelevant what is Class 1 or 2) this is not the case in SSVM where the sphere parameters change depending on which class of points are inside the sphere. Problem 14-17 is equivalent to

$$SSVM: \ v_0 = \max \delta$$
$$s.t. \ y_j\|x_j - x_0\|^2 \leq y_j(R - y_j\delta)^2, \ \text{ if } j \in N \tag{18}$$
$$R \geq \delta$$
$$\delta \geq 0 \, , x_0 \in \mathbb{R}^L.$$

This model represents the hard-margin approach. When the classes are not linearly separable, the problem becomes infeasible and lacks a solution. To address this, soft margins are introduced, allowing some misclassification to achieve a workable solution.

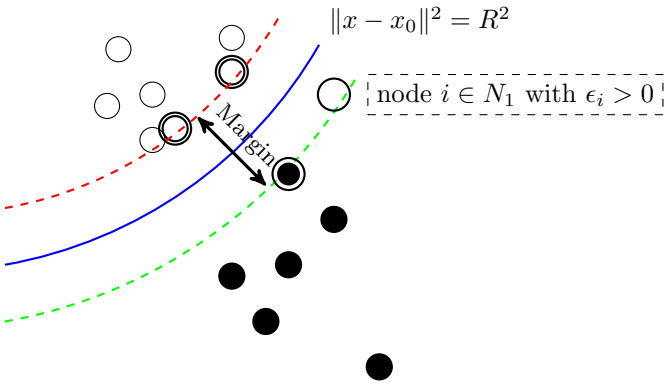


Fig. 6: Soft margins in $\mathbb{R}^2$ - two features

3.1.1 Soft Margins

If the data is not linearly separable, we introduce soft margins by incorporating one slack variables ϵ_i, for each example $i \in N$, as follows:

$$SSSVM: \; v_0 = \max \delta - \mu \sum_{i=1}^{n} \epsilon_i \tag{19}$$

$$s.t. \; \|x_i - x_0\|^2 \leq (R - \delta + \epsilon_i)^2, \quad \text{if } i \in N_1 \tag{20}$$

$$\|x_i - x_0\|^2 \geq (R + \delta - \epsilon_i)^2, \quad \text{if } i \in N_2 \tag{21}$$

$$R \geq \delta$$

$$\delta \geq 0 \, , x_0 \in \mathbb{R}^L.$$

Figure (6) illustrates a point positioned outside the boundary defined by the red dashed line. Note that with this formulation we may define that just one class can be missclassified.

Problem $SSVM$ (as well as $SSSVM$) is not convex. For convexity to hold in inequalities constraints we must have the scheme,

$$\text{convex} \; \leq \; \text{concave} \quad \text{or} \quad \text{concave} \; \geq \; \text{convex} \; .$$

Since $\|x_i - x_0\|^2$, $(R - \delta + \epsilon_i)^2$ and $(R + \delta - \epsilon_i)^2$ are convex these conditions do not hold. Besides considering for $SSVM$ (similar for $SSSVM$) the quadratic component of the constraints 20 and 21 in matricial form are the same for all $i \in N_1$ and $j \in N_2$ and equal to $[xo, R, \delta]^T Q[xo, R, \delta]$ with

$$Q = \begin{bmatrix} I & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

which is indefinite. Due to the non-convexity of Problem $SSVM$, finding an optimal solution becomes challenging in high-dimensional settings. Therefore, we developed a relaxed version of the problem, easier to solve, from which a feasible and potentially high-quality solution to the original problem can be retrieved.

3.1.2 Relaxation

In [9] is also presented a linear relaxation for the SSVM problem.

$$
\begin{aligned}
SRL_0 : \max\ & \theta \\
\text{s.t. } & 4\theta + 2(x_j - x_i)^T x_0 \leq \|x_j\|^2 - \|x_i\|^2, \quad \text{for } i \in N_1, j \in N_2 && (22)\\
& \theta \geq 0. && (23)
\end{aligned}
$$

It can be shown that this problem may be unbounded. We present a formal proof that was not included in [9].

Theorem 1 *Consider problem SRL_0, (22)-(23) and a direction*

$$d^T = [\, d_1, d_0^T \,] \tag{24}$$

with $d_1 \in R^+$, $d_0 \in R^n$. If the optimal solution of the following problem:

$$
\begin{aligned}
\min v(x_0, \theta) = & \sum \beta^{i,j} \\
& s.t.\ \ \alpha^{ij} d_0 - s^{i,j} = 1 - \beta^{i,j}, \quad for\ \forall i \in N_1, \forall ij \in N_2 && (25)\\
& \beta^{i,j}, s^{i,j} \geq 0 i \in N_1, j \in N_2
\end{aligned}
$$

is equal to zero ($v(x_0^*, \theta*) = 0$), with $\alpha^{ij} = -1/2(x_j - x_i)^T$, then d is a direction of unboundedness.

Proof: To prove that d is a direction of unboundedness we must establish that for $\forall \gamma \in R_0^+$, γd is a feasible direction, there is (22) and (23) hold:

$$
\begin{aligned}
4(\theta + \gamma d_1) + 2(x_j - x_i)^T (x_0 + \gamma d_0) \leq \|x_j\|^2 - \|x_i\|^2, \quad \text{for } i \in N_1, j \in N_2 && (26)\\
\theta + \gamma d_1 \geq 0 && (27)
\end{aligned}
$$

Now regarding the unboundedness, if:

$$4d_1 + 2(x_j - x_i)^T d_0 \leq 0, \quad \text{for } i \in N_1, j \in N_2 \tag{28}$$

then γd is a feasible direction when $\gamma \to +\infty$. Considering $\alpha^{ij} = -1/2(x_j - x_i)^T$ we must have:

$$d_1 - \alpha^{ij} d_0 \leq 0, \Leftrightarrow \alpha^{ij} d_0 \geq d_1, \ \forall i \in N_1, \forall j \in N_2 \tag{29}$$

Without loosing generality we can make $d_1 = 1$ and obtain the condition:

$$\alpha^{ij} d_0 \geq 1\ \forall i \in N_1, \forall j \in N_2 \tag{30}$$

It is straightforward to see that this system of inequalities is feasible if the optimal value of problem (25) is zero, where $s^{i,j}$ are the slack variables of the system. The variables $\beta^{i,j}$ are artificial variables to ensure that (25) is always feasible. □

To prevent the solution to be unbounded, in [9] it is proposed an indirect regularization technique introducing a new variable ρ and a new constraint, in the above formulation, obtaining:

$$
\begin{aligned}
SRL_1(k): \ & \max \theta + k\rho && (31)\\
& \text{s.t. } 4\theta + 2(x_j - x_i)^T x_0 \leq \|x_j\|^2 - \|x_i\|^2 \text{ for } i \in N_1, j \in N_2 && (32)\\
& \quad 4\theta + 2(x_j - x_i)^T x_0 \geq \rho \ \forall i \in N_1, \forall j \in N_2 && (33)\\
& \quad \theta \geq 0
\end{aligned}
$$

In the next section we generalize these spherical models to interval-valued data.

3.2 The Spherical SVM for Interval data

Developing classification methods for interval-valued data is essential due to the growing need to handle data with inherent variability. Interval-valued data arises in many fields - such as sensor measurements, economic forecasting, and environmental monitoring - where it is necessary to aggregate readings or measurements to define one feature. Although we may apply intervals to data where exact values are often unknown or fluctuating, this is not the correct theoretical ground for symbolic data analysis. Instead of relying on single-point estimate, like the average, to aggregate an array of data, interval-valued data provides a range, within which the values are expected to lie, offering a more comprehensive representation of the data variability. Effective classification methods for this type of data enable better decision-making by accurately capturing and interpreting the underlying information, reducing the risk of misclassification and improving the reliability of predictive models.
Within this framework, in this section, we propose an extension of SSVM for interval-valued data.

3.2.1 One feature - SSVMI

We will start with the univariate case $L = 1$ of just one variable X which is a random interval-valued variable. For each observation i, X_i can be represented by an interval with center c_{X_i} and radius r_{X_i}:

$$I_{X_i} = [c_{X_i} - r_{X_i}, c_{X_i} + r_{X_i}], \ i \in N \tag{34}$$

and the quantile function $\Psi^{-1}_{X_i}$.

$$\Psi^{-1}_{X_i}(t) = c_{X_i} + (2t-1)r_{X_i}, \quad \text{for } 0 \leq t \leq 1 \tag{35}$$

The classification model, analogue to (18) is addressed by an optimization problem where the goal is to find $\Psi^{-1}_{X_0}$ and positive R that maximizes non-negative δ and the

formulation is as follows:

$$SSVMI : \max \delta \tag{36}$$

$$\begin{aligned} \text{s.t. } & D^2(\Psi_{X_i}^{-1}, \Psi_{X_0}^{-1}) \leq (R-\delta)^2, \quad \text{for } i \in N_1 \\ & D^2(\Psi_{X_j}^{-1}, \Psi_{X_0}^{-1}) \geq (R+\delta)^2, \quad \text{for } j \in N_2 \qquad (37) \\ & R \geq \delta \\ & \delta \geq 0. \end{aligned}$$

Using for the distance D, the Mallows distance

$$D_M^2(\Psi_{X_i}^{-1}, \Psi_{X_0}^{-1}) = (c_{X_i} - c_{X_0})^2 + \frac{1}{3}(r_{X_i} - r_{X_0})^2 \tag{38}$$

where c_{X_0} and r_{X_0} are respectively the center and the radius of the quantile function $\Psi_{X_0}^{-1}$ representing the center of the sphere with radius R, we obtain the following problem formulation:

$$SSVMI : \max \delta$$

$$\begin{aligned} \text{s.t. } & (c_{X_i} - c_{X_0})^2 + \frac{1}{3}(r_{X_i} - r_{X_0})^2 \leq (R-\delta)^2, \quad \text{for } i \in N_1 \qquad (39) \\ & (c_{X_j} - c_{X_0})^2 + \frac{1}{3}(r_{X_j} - r_{X_0})^2 \geq (R+\delta)^2, \quad \text{for } j \in N_2 \qquad (40) \\ & R \geq \delta \\ & r_{X_0} \geq 0 \\ & \delta \geq 0 \end{aligned}$$

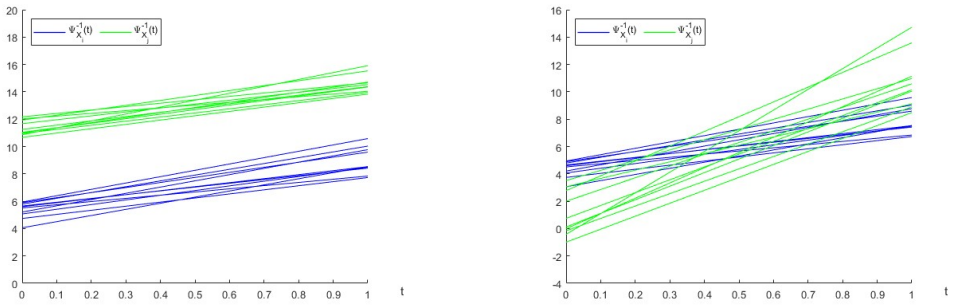
using $y_i = 1$ for class 1 and $y_i = -1$ for class 2 we obtain the equivalent formulation

$$SSVMI : \max \delta$$

$$\begin{aligned} \text{s.t. } & y_i(c_{X_i} - c_{X_0})^2 + \frac{y_i}{3}(r_{X_i} - r_{X_0})^2 \leq y_i(R - y_i\delta)^2, \quad \text{for } i \in N \qquad (41) \\ & R \geq \delta \\ & r_{X_0} \geq 0 \\ & \delta \geq 0. \end{aligned}$$

Example 2 To illustrate the application of $SSVM1$ we present an example that will be fully explained and developed in the computational experience. The purpose now is just to show the graphical representation, in Figure 7, of the quantile functions of two data sets, each with two classes of intervals (green and blue), and one feature. Sim1 represent an example where the classes are separable while in Sim2 they are not.

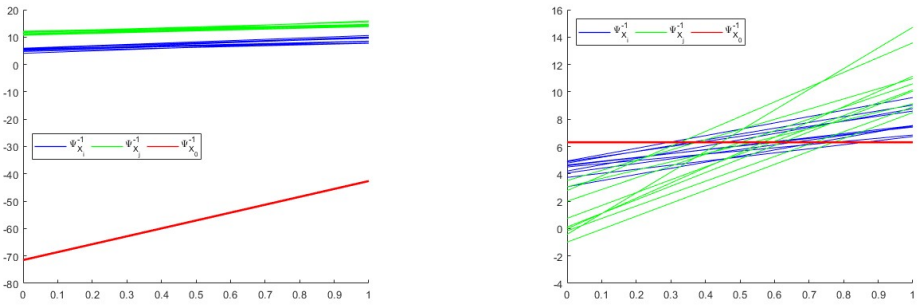
In Figure 8 is shown Sim1 and Sim2 quantile functions and the quantile function representing the separation classifier $\Psi_{X_0}^{-1}$ when using the complete set for training. It is clear that for



Sim1 Sim2

Fig. 7: Sim1 and Sim2 quantile functions

Sim1 the blue class is closer to $\Psi_{X_0}^{-1}$ (Mallows distance smaller than R), than the green class. Given a new unit, it will be classified as blue if the Mallows distance from $\Psi_{X_0}^{-1}$ is smaller than R. For Sim2, since the classes are not separable in less clear how the separation is defined.



Sim1 with classifier Sim2 with classifier

Fig. 8: Sim1 and Sim2 quantile functions with classifier

3.2.2 Multiple features - SSVMMI

Now to extend the previous formulation to more than one feature, each observation $i \in N$ incorporates L intervals, $I_{X_i(1)}$ to $I_{X_i(L)}$, each one defined by a center $c_{X_i(l)}$ and radius $r_{X_i(l)}$.

$$I_{X_i(l)} = [c_{X_i(l)} - r_{X_i(l)}, c_{X_i(l)} + r_{X_i(l)}],\ i \in N,\ l = 1, ..., L \tag{42}$$

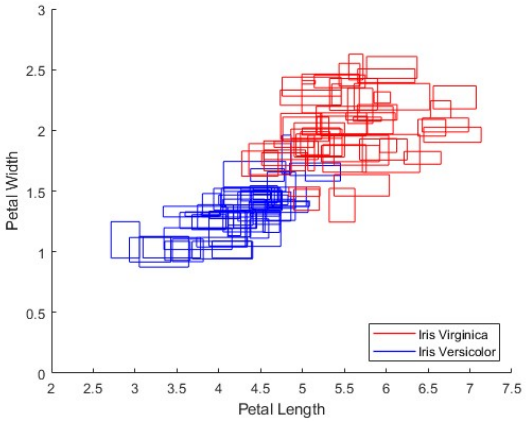


Fig. 9: Iris intervals (two features)

Example 3 In Figure 9 it is represented the well known Iris dataset adapted to interval-value data. Only two classes (Iris Virginica and Iris Versicolor) and two features (petal length and width) where considered. Now each unit/example is represented by a rectangle, where the base of the rectangle is the interval of one feature (petal lenght limits defined in the horizontal axis) and the height represents the interval of the second feature, in the vertical axis (petal width). The separation sphere has a center to be determined, defined by an array of L intervals $I_{X_0(1)}, \ldots, I_{X_0(L)}$ and a radius R. Let $c_{X_0(l)}$ and $r_{X_0(l)}$ be respectively the center and radius of interval $I_{X_0(l)}$.

Using the Mallows distance for the multivariate case with intervals

$$D_M^2(\Psi_X^{-1}, \Psi_Y^{-1}) = \sum_{l=1}^{L} \left[(c_{X(l)} - c_{Y(l)})^2 + \frac{1}{3}(r_{X(l)} - r_{Y(l)})^2 \right], \; l = 1, ..., L \tag{43}$$

we have the following formulation for the Spherical SVM for Multiple Interval (SSVMMI)

$SSVMMI$: max δ

$$\text{s.t.} \quad \sum_{l=1}^{L} \left[(c_{X_i(l)} - c_{X_0(l)})^2 + \frac{1}{3}(r_{X_i(l)} - r_{X_0(l)})^2 \right] \leq (R - \delta)^2, \quad \text{for } i \in N_1 \tag{44}$$

$$\sum_{l=1}^{L} \left[(c_{X_j(l)} - c_{X_0(l)})^2 + \frac{1}{3}(r_{X_j(l)} - r_{X_0(l)})^2 \right] \geq (R + \delta)^2, \quad \text{for } j \in N_2 \tag{45}$$

$$R \geq \delta$$

$$\delta \geq 0$$
$$r_{X_0(l)} \geq 0., \ l = 1, ..., L$$

As in (40), identifying the class membership using $y_i = \pm 1$ labels, we obtain a more compact description of the model:

$$SSVMMI : \max \delta$$
$$\text{s.t. } y_i \sum_{l=1}^{n} \left[(c_{X_i(l)} - c_{X_0(l)})^2 + \frac{1}{3}(r_{X_i(l)} - r_{X_0(l)})^2 \right] \leq y_i(R - y_i\delta)^2, \quad \text{for } i \in N \tag{46}$$
$$R \geq \delta$$
$$\delta \geq 0$$
$$r_{X_0(l)} \geq 0, \ l = 1, ..., L.$$

3.2.3 Soft Margins in the Multiple features - SSSVMMI

If data, distributed in two classes C_1 and C_2, such that $N_1 = \{1, 2, \ldots, |C_1|\}$ and $N_2 = \{1, 2, \ldots, |C_2|\}$, is not separable, we have to consider the concept of soft margins and the introduction of slack variables ϵ_i and ϵ_j as follows (with parameter $\mu > 0$):

$$SSSVMMI : \max \delta - \mu(\sum_{i=1}^{n_1} \epsilon_i + \sum_{j=1}^{n_2} \epsilon_j), \ n_1 = \left|C_1\right|, \ n_2 = \left|C_2\right| \tag{47}$$
$$\text{s.t. } \sum_{l=1}^{L} \left[(c_{X_i(l)} - c_{X_0(l)})^2 + \frac{1}{3}(r_{X_i(l)} - r_{X_0(l)})^2 \right] \leq (R - \delta + \epsilon_i)^2, \quad \text{for } i \in N_1 \tag{48}$$
$$\sum_{l=1}^{L} \left[(c_{X_j(l)} - c_{X_0(l)})^2 + \frac{1}{3}(r_{X_j(l)} - r_{X_0(l)})^2 \right] \geq (R + \delta - \epsilon_j)^2, \quad \text{for } j \in N_2 \tag{49}$$
$$R \geq \delta$$
$$\delta \geq 0$$
$$r_{X_0(l)} \geq 0, \ l = 1, ..., L$$
$$\epsilon_i \geq 0, \ \text{for } i \in N_1$$
$$\epsilon_j \geq 0, \ \text{for } i \in N_2.$$

The variables ϵ_i and ϵ_j allow for points in Class 1 to be outside the sphere, and points of Class 2 to fall inside the sphere, respectively, counting in both cases with the margin.
Instead, without loosing generality and maintaining the previous considerations, we will use the following alternative formulation, which allows simpler calculations for

the linear relaxation presented in the next section.

$$SSSVMMI: \max \delta - \mu(\sum_{i=1}^{n_1}\epsilon_i + \sum_{j=1}^{n_2}\epsilon_j),\ n_1 = \left|C_1\right|,\ n_2 = \left|C_2\right| \tag{50}$$

$$\text{s.t.} \quad \sum_{l=1}^{L}\left[(c_{X_i(l)} - c_{X_o(l)})^2 + \frac{1}{3}(r_{X_i(l)} - r_{X_o(l)})^2\right] \leq (R-\delta)^2 + \epsilon_i, \quad \text{for } i \in N_1 \tag{51}$$

$$\sum_{l=1}^{L}\left[(c_{X_j(l)} - c_{X_o(l)})^2 + \frac{1}{3}(r_{X_j(l)} - r_{X_o(l)})^2\right] \geq (R+\delta)^2 - \epsilon_j, \quad \text{for } j \in N_2 \tag{52}$$

$$R \geq \delta$$
$$\delta \geq 0$$
$$r_{X_o(l)} \geq 0,\ l = 1, ..., L$$
$$\epsilon_i \geq 0, \text{ for } i \in N_1$$
$$\epsilon_j \geq 0, \text{ for } j \in N_2.$$

3.2.4 A linear relaxation of the SSSVMMI

The quadratic Soft Margin Spherical SVM for Multiple Interval (SSSVMMI) problem is nonconvex, so for large scale problems a global optimal solution may be difficult to achieve computationally even with the best solvers. In that sense we developed a relaxation problem, from which a feasible solution for the original problem can be retrieved. The relaxation obtained is a linear programming problem and so can be efficiently solved even for large scale problems. The relaxation can also be used to find an upper bound for the optimal value. In an heuristic approach this value can be used to define optimality gaps. The upper bound is a key feature in a branch and bound approach in a future work.

So from 51 and 52we get

$$SSSVMMI: \max \delta - \mu(\sum_{i=1}^{n_1}\epsilon_i + \sum_{j=1}^{n_2}\epsilon_j),\ n_1 = \left|C_1\right|,\ n_2 = \left|C_2\right| \tag{53}$$

$$\text{s.t.} \quad \sum_{\ell=1}^{L}\left[c^2_{X_i(l)} - 2c_{X_i(l)}c_{X_0(l)} + c^2_{X_0(l)} + \frac{1}{3}(r^2_{X_i(l)} - 2r_{X_i(l)}r_{X_0(l)} + r^2_{X_0(l)})\right]$$
$$\leq R^2 + \delta^2 - 2R\delta + \epsilon_i, \quad \text{for } i \in N_1 \tag{54}$$

$$\sum_{\ell=1}^{L}\left[c^2_{X_j(l)} - 2c_{X_j(l)}c_{X_0(l)} + c^2_{X_0(l)} + \frac{1}{3}(r^2_{X_j(l)} - 2r_{X_j(l)}r_{X_0(l)} + r^2_{X_0(l)})\right]$$
$$\geq R^2 + \delta^2 + 2R\delta - \epsilon_j, \quad \text{for } j \in N_2 \tag{55}$$

$$
\begin{aligned}
& R \geq \delta \\
& \delta \geq 0 \\
& r_{X_o(l)} \geq 0,\ l = 1, ..., L \\
& \epsilon_i \geq 0,\ \text{for } i \in N_1 \\
& \epsilon_j \geq 0,\ \text{for } j \in N_2
\end{aligned}
$$

and rearranging the inequalities we get

$$
SSSVMMI : \max \delta - \mu(\sum_{i=1}^{n_1} \epsilon_i + \sum_{j=1}^{n_2} \epsilon_j),\ n_1 = \left|C_1\right|,\ n_2 = \left|C_2\right| \tag{56}
$$

$$
\begin{aligned}
\text{s.t. } & \sum_{\ell=1}^{L}(c_{X_i(l)}^2 + \frac{1}{3}r_{X_i(l)}^2) + \sum_{\ell=1}^{L}(c_{X_0(l)}^2 + \frac{1}{3}r_{X_0(l)}^2) - 2(\sum_{\ell=1}^{L} c_{X_i(l)}c_{X_0(l)} + \frac{1}{3}r_{X_i(l)}r_{X_0(l)}) \\
& \leq R^2 + \delta^2 - 2R\delta + \epsilon_i, \quad \text{for } i \in N_1 \qquad (57) \\
& \sum_{\ell=1}^{L}(c_{X_j(l)}^2 + \frac{1}{3}r_{X_j(l)}^2) + \sum_{\ell=1}^{L}(c_{X_0(l)}^2 + \frac{1}{3}r_{X_0(l)}^2) - 2(\sum_{\ell=1}^{L} c_{X_j(l)}c_{X_0(l)} + \frac{1}{3}r_{X_j(l)}r_{X_0(l)}) \\
& \geq R^2 + \delta^2 + 2R\delta - \epsilon_j, \quad \text{for } j \in N_2 \qquad (58) \\
& R \geq \delta \\
& \delta \geq 0 \\
& r_{X_o(l)} \geq 0,\ l = 1, ..., L \\
& \epsilon_i \geq 0,\ \text{for } i \in N_1 \\
& \epsilon_j \geq 0,\ \text{for } j \in N_2.
\end{aligned}
$$

Let us introduce some notation that will help to simplify the writing of the model. We have:

$$
\|x\|_M^2 = \sum_{\ell=1}^{L}(c_{X(l)}^2 + \frac{1}{3}r_{X(l)}^2) \tag{59}
$$

$$
(x^T y)_M = \sum_{\ell=1}^{L} c_{X(l)}c_{Y(l)} + \frac{1}{3}\sum_{\ell=1}^{L} r_{X(l)}r_{Y(l)} \tag{60}
$$

and so from (57) and (58), taking $\theta = R\delta$ we have (with parameter $\mu > 0$):

$$
SSSVMMI : \max \delta - \mu(\sum_{i=1}^{n_1} \epsilon_i + \sum_{j=1}^{n_2} \epsilon_j),\ n_1 = \left|C_1\right|,\ n_2 = \left|C_2\right| \tag{61}
$$

$$
\text{s.t. } \|x_i\|_M^2 - 2(x_i^T x_0)_M + 2\theta - \epsilon_i \leq R^2 + \delta^2 - \|x_0\|_M^2, \quad \text{for } i \in N_1 \tag{62}
$$

$$
\|x_j\|_M^2 - 2(x_j^T x_0)_M - 2\theta + \epsilon_j \geq R^2 + \delta^2 - \|x_0\|_M^2, \quad \text{for } j \in N_2 \tag{63}
$$

$$\delta \geq 0$$
$$R \geq 0$$
$$\epsilon_i \geq 0, \text{ for } i \in N_1$$
$$\epsilon_j \geq 0, \text{ for } j \in N_2.$$

Now, since the right sides of both inequalities 62 and 63 are the same, we have

$$\|x_i\|_M^2 - 2(x_i^T x_0)_M + 2\theta - \epsilon_i \leq R^2 + \delta^2 - \|x_0\|_M^2 \leq \|x_j\|_M^2 - 2(x_j^T x_0)_M - 2\theta + \epsilon_j, \quad \text{for } i \in N_1, j \in N_2 \tag{64}$$

Because each feasible solution of $SSSVMMI$ fulfills condition

$$\|x_i\|_M^2 - 2(x_i^T x_0)_M + 2\theta - \epsilon_i \leq \|x_j\|_M^2 - 2(x_j^T x_0)_M - 2\theta + \epsilon_j, \quad \text{for } i \in N_1, j \in N_2 \tag{65}$$

we have that

$$4\theta + 2(x_j^T x_0)_M - 2(x_i^T x_0)_M - \epsilon_j - \epsilon_i \leq \|x_j\|_M^2 - \|x_i\|_M^2, \quad \text{for } i \in N_1, j \in N_2 \tag{66}$$

equivalent to:

$$4\theta + 2[(x_j - x_i)^T x_0]_M - \epsilon_j - \epsilon_i \leq \|x_j\|_M^2 - \|x_i\|_M^2, \quad \text{for } i \in N_1, j \in N_2. \tag{67}$$

So, we come to the linear relaxation problem formulation as follows

$$LSSSVMMI : \max \theta - \mu(\sum_{i=1}^{n_1} \epsilon_i + \sum_{j=1}^{n_2} \epsilon_j),\ n_1 = \left|C_1\right|,\ n_2 = \left|C_2\right| \tag{68}$$

$$\text{s.t. } 4\theta + 2[(x_j - x_i)^T x_0]_M - \epsilon_j - \epsilon_i \leq \|x_j\|_M^2 - \|x_i\|_M^2, \quad \text{for } i \in N_1, j \in N_2 \tag{69}$$

$$\theta \geq 0$$
$$\epsilon_i \geq 0, \text{ for } i \in N_1$$
$$\epsilon_j \geq 0, \text{ for } j \in N_2.$$

Because this problem may be unbounded (the proof is similar to the one presented in Theorem 1) so as in 3.1.2 we propose an indirect regularization technique introducing a new variable ρ and a new constraint in the above formulation (with parameters $\mu > 0$ and $k > 0$)

$$LSSSVMMI : \max \theta + k\rho - \mu(\sum_{i=1}^{n_1} \epsilon_i + \sum_{j=1}^{n_2} \epsilon_j),\ n_1 = \left|C_1\right|,\ n_2 = \left|C_2\right| \tag{70}$$

$$\text{s.t. } 4\theta + 2[(x_j - x_i)^T x_0]_M - \epsilon_j - \epsilon_i \leq \|x_j\|_M^2 - \|x_i\|_M^2, \quad \text{for } i \in N_1, j \in N_2 \tag{71}$$

$$4\theta + 2[(x_j - x_i)^T x_0]_M - \epsilon_j - \epsilon_i \geq \rho, \quad \text{for } i \in N_1, j \in N_2 \tag{72}$$

$$
\begin{aligned}
&\theta \geq 0\\
&\rho \geq 0\\
&\epsilon_i \geq 0, \text{ for } i \in N_1\\
&\epsilon_j \geq 0, \text{ for } j \in N_2.
\end{aligned}
$$

After solving the relaxation, a solution for the original problem can be produced. The solution of the linear problem $LSSSVMMI$ is $(\theta, \rho, x_0, \epsilon)$. Therefore, we must find R and δ, and for that we propose two possible procedures.

<u>R **and** δ **recover 1**</u>: Since x_0 is known, to get R and δ, we solve the following linear problem (with $\mu > 0$):

$$
\begin{aligned}
SRLI: \ \max\ & \delta - \mu(\sum_{i=1}^{n_1}\varphi_i + \sum_{j=1}^{n_2}\varphi_j),\ n_1 = \left|C_1\right|,\ n_2 = \left|C_2\right| && (73)\\
\text{s.t. } & R - \delta + \varphi_i \geq \|x_i - x_{0_{optim}}\|_M \text{ for } i \in N_1, j \in N_2 && (74)\\
& R + \delta - \varphi_j \leq \|x_j - x_{0_{optim}}\|_M, \text{ for } i \in N_1, j \in N_2 && (75)\\
& R \geq \delta\\
& \delta \geq 0\\
& \varphi_i \geq 0, \text{ for } i \in N_1\\
& \varphi_j \geq 0, \text{ for } j \in N_2.
\end{aligned}
$$

<u>R **and** δ **recover 2**</u>: In alternative to previous procedure, and considering that at the support vectors, we have:

$$\theta_{optim} + 2[(x_{j_{sv}} - x_{i_{sv}})^T x_{0_{optim}}]_M - \epsilon_{i_{sv}} - \epsilon_{j_{sv}} = \|x_{j_{sv}}\|_M^2 - \|x_{i_{sv}}\|_M^2, \text{ for } i \in N_1, j \in N_2$$

after finding the support vectors, we calculate R and δ by solving the equations:

$$
\begin{aligned}
R =& \frac{\sqrt{\|x_{j_{sv}} - x_{0_{optim}}\|_M^2 + \epsilon_{j_{sv}}} + \sqrt{\|x_{i_{sv}} - x_{0_{optim}}\|_M^2 - \epsilon_{i_{sv}}}}{2}\\
\delta =& \frac{\sqrt{\|x_{j_{sv}} - x_{0_{optim}}\|_M^2 + \epsilon_{j_{sv}}} - \sqrt{\|x_{i_{sv}} - x_{0_{optim}}\|_M^2 - \epsilon_{i_{sv}}}}{2}
\end{aligned}
$$

Although the second procedure seems simpler, the first procedure was more efficient in practice.

4 Computational Experience

The goal of the computational experience was to test the classification accuracy of the exact quadratic (SSSVMMI) and the linear relaxation (LSSSVMMI) models by comparing them against the accuracy results obtained with all the 34 classifiers available in Matlab Classification Learner App (Version R2024a) **??**, for real-world and generated datasets. In Figure 10 it is possible to identify in the horizontal axis all the Matlab classifiers.
Experiments were run on a computer with:
- Intel(R) Core(TM) i7-8550U processor, 1.80GHz
- 8.00 GB of RAM
- 64-bit operating system running on Windows 10, version 20H2.
- We code our method on Matlab R2024a, and we used BARON software, version 13.0.1, running on Matlab with MATLAB/BARON interface version v1.89.
- The BARON solver was used to solve the quadratic formulation.
- To run all 34 Matlab classifiers we used the Machine Learning and Deep Learning APP from MATLAB.

We used three types of interval-valued datasets. The properties of these datasets are summarized in Table 3.

Table 3: Attributes and Number of Data Set Samples

Data Set	Attributes	class1	class 2
Sim1	1	10	10
Sim2	1	10	10
Iris	4	50	50
Thyroid	5	150	35
Seeds	7	70	140
Twonorm	20	200	200
Mushroom	5	30	20

Sim1 and Sim2 are simulated interval-valued datasets. Sim1 was built as follows: for class 1 we made a 10×10 array of random real numbers taken from N(7,1), then for each row we took the minimum value for the lower interval limit and the maximum value for the upper interval limit. For class 2 the procedure is the same but with real numbers taken from N(13,1). So we have two classes with different average values and similar variance values. Sim2 was built the same way as Sim1 but for class 1 we used a 10x10 array of random real numbers taken from N(6,1) and for class 2, real numbers taken from N(6,3). So we have two classes with similar average values and different variance values. In Figure 7 we can see the quantile functions of both data sets.

The Mushroom dataset was extracted from site - The Funji of California [29], and is a real-world interval-valued dataset. It contains 50 species of two fungi genera, 30 of genera Agaricus and 20 of genera Amanita. There are five interval-valued features: the pileus cap width - Pw, the stipe length - Sl, the stipe thickness - St, the spores major axis length - Sma, and the spores minor axis length - Smi. Some instances of the mushroom dataset are shown in Table 4. The goal of our experiment on this dataset is to predict the genera.

Table 4: Mushroom interval-valued data

Genera	Species	Pw (cm)	Sl (cm)	St (cm)	Sma (cm)	Smi (μm)
Agaricus	Moronii	[6; 12]	[2; 7]	[1.5; 3]	[6; 7.5]	[4; 5]
Agaricus	Arorae	[3; 8]	[4; 9]	[0.5; 2.5]	[4.5; 5]	[3; 3.5]
Agaricus	Fissuratus	[6; 21]	[4; 14]	[1; 3.5]	[6.5; 9]	[4.5; 6]
Amanita	Augusta	[4; 12]	[5; 15]	[1; 2]	[8; 12]	[6; 8]
Amanita	Calyptroderma	[8; 25]	[10; 20]	[1.5; 4]	[8; 11]	[5; 6]

Table 5: Iris Interval-valued Data

	sepal length	sepal width	petal length	petal width
Iris versicolor 1	[6.49; 7.57]	[3.16; 3.46]	[4.39; 4.75]	[1.35; 1.49]
⋮	...	...	...	...
Iris versicolor 50	[5.19; 5.86]	[2.63; 2.81]	[3.89; 4.25]	[1.27; 1.33]
Iris virginica 1	[6.04; 6.36]	[3.13; 3.44]	[5.66; 6.34]	[2.39; 2.51]
⋮	...	...	..	...
Iris virginica 50	[5.34; 6.40]	[2.79; 3.01]	[5.06; 5.14]	[1.67; 1.96]

Iris, Thyroid, Seeds and Twonorm interval-valued datasets were constructed from real-world numerical datasets available from UCI Machine Learning Repository [30]. The procedure to generate an interval-valued data from numerical data was done as follows: the lower and upper limit of intervals, for each feature, were obtained as reducing or increasing the original numerical data by a random value between zero and 10 percent of the average of all values for that feature. From the three classes of Iris dataset we used only two, versicolor and virginica. In Table 5 we show some instances of the new generated interval-valued Iris dataset
In Figure 9 we show the intervals for two features (petal length and petal width) of classes versicolor and virginica from Iris dataset.

For comparison, three methods of validation were applied to the datasets: 5-fold cross-validation; classification using the complete set; train with 90% test with 10%,

only one run.

For classification with Matlab we used the mid point of each interval-valued data. The accuracy results from the 34 Matlab classifiers were condensed in intervals in which the lower limit refers to the minimum accuracy and the upper limit to the maximum accuracy value obtained. Also for SSSVMMI and LSSSVMMI we calculate the gap from the best Matlab accuracy value. The accuracy results obtained are summarized in Tables 6, 7 and 8.

Using 5-fold cross-validation (see Table 6) the gap to the best Matlab classifier

Table 6: Test set 5-fold cross validation accuracy results

Data Set	Accuracy%			Gap%	
	MATLAB Classifiers	SSSVMMI	LSSSVMMI	SSSVMMI	LSSSVMMI
Sim1	[50,00 ; 100,00]	100.00	100.00	0	0
Sim2	[50,00 ; 65,00]	100.00	95.00	-	-
Iris	[50,00 ; 96,00]	91.00	92.00	5.21	4.17
Seeds	[66.67 ; 96,67]	95.72	94.76	0.98	1.98
Thyroid	[81.08 ; 98.92]	98.38	94.05	3.91	3.91
Twonorm	[69.00 ; 98.25]	96.50	94.25	1.78	4.07
Mushrooms	[60.00 ; 94.00]	90.00	84.00	4.26	10.64

Table 7: Test set with complete set validation accuracy results

Data Set	Accuracy%			Gap%	
	MATLAB Classifiers	SSSVMMI	LSSSVMMI	SSSVMMI	LSSSVMMI
Sim1	[50.00 ; 100.00]	100.00	100.00	0	0
Sim2	[50.00 ; 100.00]	100.00	100.00	0	0
Iris	[50.00 ; 100.00]	99.00	95.00	1.00	5.00
Seeds	[66.67 ; 100.00]	93.81	93.33	6.19	6.67
Thyroid	[81.08 ; 100.00]	100.00	98.92	0.00	1.08
Twonorm	[56.25 ; 100.00]	100.00	97.25	0.00	2.75
Mushrooms	[60.00 ; 100.00]	100.00	94.00	0.00	6.00

accuracy varies from 0.98% to 5.21% with model SSSVMMI and from 1.98% to 10.64 with model LSSSVMMI

With complete set validation (see Table 7) we got 100% accuracy in three datasets with SSSVMMI.

Table 8: 90% train 10% test accuracy results

Data Set	Accuracy%			Gap%	
	MATLAB Classifiers	SSSVMMI	LSSSVMMI	SSSVMMI	LSSSVMMI
Iris	[50.00 ; 90.00]	90.00	90.00	0.00	0.00
Seeds	[66.67 ; 100.00]	95.24	95.24	4.76	4.76
Thyroid	[78.95 ; 100.00]	94.74	100.00	2.26	0.00
Twonorm	[65.00 ; 100.00]	100.00	92.50	0.00	7.50
Mushrooms	[60.00 ; 100.00]	80.00	80.00	20.00	20.00

In Table 8 the 20% gap in mushroom dataset is due to the fact that we used a very few number of instances (only 5) for the test set and therefore one misclassified instance represents 20%.
Figure 10 shows a comparison graph of the SSSVMMI and LSSSVMMI using complete set validation accuracy for thyroid dataset, with all the 34 classifiers available in Matlab. From this computational experience we may conclude that this methodology

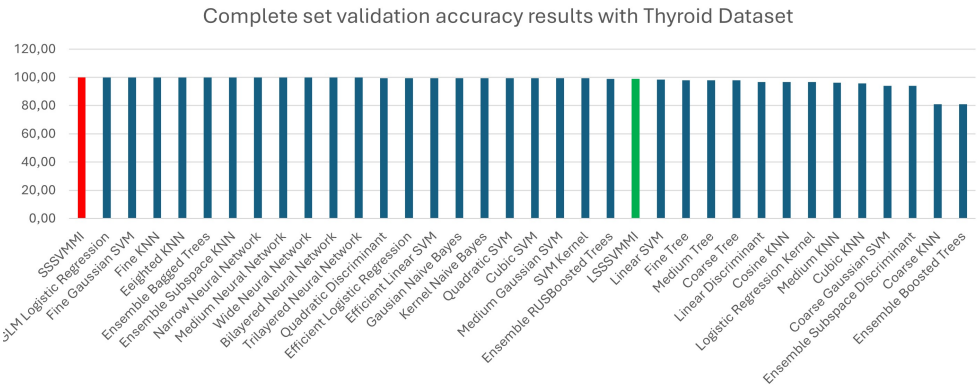


Fig. 10: Thyroid comparison results with complete set validation

can produce good results for classification of interval-value data. The alternative to reduce the data to an array data by aggregating values using the average or another central statistic has the disadvantage of ignoring important information about the data. The relaxation LSSSVMM1 produces worse accuracy but it is easier to solve than the quadratic exact formulation SSSVMM1 and the difference between the two is small enough to allow considering the relaxation an interesting solution. On the other hand both methods compare well with a full battery of well establish state of the art methods.

5 Conclusions

In this paper, we developed a model for interval-valued data classification inspired by SVM with spherical separation. This approach benefits from defining the separation using a center represented as an interval and a radius, rather than a hyperplane. Defining a hyperplane poses significant challenges, as the linear combination of quantile functions does not yield a quantile function, and multiplying a non-decreasing quantile function by a negative coefficient disrupts its quantile properties. By defining a hypersphere centered on an interval, we overcome these limitations.

The proposed model incorporates the SVM-inspired concepts of a maximized margin and soft margins. It is formulated as a quadratic optimization model, for which we provide a linear relaxation. The solution of this linear relaxation enables the construction of a feasible solution for the original model.

The results of this model were compared with several classification models in Matlab. Both simulated and real datasets, commonly used in classification literature, were tested. Of the instances used, only one was truly defined by intervals. For the rest, we applied a methodology followed by other authors of interval classification models, which consists of transforming classical data into intervals.

The results were very satisfactory, especially considering that the model has no more than a single multi-parameter (to account for deviations and margin in the objective function) and allows for working within the original data space. The comparison with state of the art methods encourage the use of this proposal specially the one based on the relaxation model.

As future work we intend to use the method in a real application, where the raw data has features with array data with different sizes and some missing data. A situation where this is common is in medical data where units are patients and features are defined by several reading of clinical indicators (Heart rate (pulse), Blood pressure Respiratory rate, Body temperature, Oxygen saturation) over a period of time. We also want to extend the method to histogram value data.

References

[1] Tax, D.M.J., Duin, R.P.W.: Support vector data description. Machine Learning **54**(1), 45–66 (2004)

[2] Abe, S.: Support Vector Machines for Pattern Classification. Springer (2005). https://doi.org/10.1007/1-84628-219-5

[3] Hao, P.-Y., Chiang, J.-H., Lin, Y.-H.: A new maximal-margin spherical-structured multi-class support vector machine. Applied Intelligence **30**(2), 98–111 (2009) https://doi.org/10.1007/s10489-007-0101-z

[4] Fathi, M., Nemati, M., Mohammadi, S.M., Abbasi-Kesbi, R.: A machine learning approach based on svm for classification of liver diseases. Biomedical Engineering: Applications, Basis and Communications **32**(03), 2050018 (2020)

[5] Chandra, M.A., Bedi, S.: Survey on svm and their application in image classification. International Journal of Information Technology **13**(5), 1–11 (2021)

[6] D'Onofrio, F., Grani, G., Monaci, M., Palagi, L.: Margin optimal classification trees. Computers & Operations Research **161**, 106441 (2024)

[7] Bomze, I., D'Onofrio, F., Palagi, L., Peng, B.: Feature selection in linear svms via hard cardinality constraint: a scalable sdp decomposition approach. arXiv preprint arXiv:2404.10099 (2024)

[8] Dias, T., Amaral, P.: A classification method based on a cloud of spheres. EURO Journal on Computational Optimization **11**, 100077 (2023)

[9] Malha, R., Amaral, P.: A maximal margin hypersphere svm. In: International Conference on Computational Science and Its Applications, pp. 304–319 (2021). Springer

[10] Berman, J.j.: Logic and Critical Thinking in Biomedical Sciences, (2020)

[11] Li, H., Fei, G., Wang, S., Liu, B., Shao, W., Mukherjee, A., Shao, J.: Bimodal distribution and co-bursting in review spam detection. In: Proceedings of the 26th International Conference on World Wide Web. WWW '17, pp. 1063–1072. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2017). https://doi.org/10.1145/3038912.3052582 . https://doi.org/10.1145/3038912.3052582

[12] Klimek, P., Yegorov, Y., Hanel, R., Thurner, S.: Statistical detection of systematic election irregularities. In: Natl Acad Sci U S A., vol. 109, pp. 16469–16473 (2012)

[13] Fallah, B., Sodoudi, S.: Bimodality and regime behavior in atmosphere–ocean interactions during the recent climate change. Dynamics of Atmospheres and Oceans **70**, 1–11 (2015) https://doi.org/10.1016/j.dynatmoce.2015.02.002

[14] Diday, E.: The symbolic approach in clustering and related methods of data analysis. Proceedings of IFCS, Classification and Related Methods of Data Analysis, 1988, 673–384 (1988)

[15] Arroyo, J., Maté, C.: Forecasting histogram time series with k-nearest neighbours methods. International Journal of Forecasting **25**(1), 192–207 (2009)

[16] Irpino, A., Verde, R.: A New Wasserstein Based Distance for the Hierarchical Clustering of Histogram Symbolic Data

[17] Dias, S.: Linear regression with empirical distributions. Phd thesis, University of Porto (2014)

[18] Bock, H.-H., Diday, E.: Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, (2000). https://doi.org/

10.1007/978-3-642-57155-8

[19] Irpino, A., Verde, R.: Basic statistics for distributional symbolic variables: a new metric-based approach. Advances in Data Analysis and Classification **9**(2), 143–175 (2015) https://doi.org/10.1007/s11634-014-0176-4

[20] Dias, S., Brito, P.: Linear regression model with histogram-valued variables. Statistical Analysis and Data Mining: The ASA Data Science Journal **8**(2), 75–113 (2015)

[21] Dias, S., Brito, P., Amaral, P.: Discriminant analysis of distributional data via fractional programming. European Journal of Operational Research **294**(1), 206–218 (2021)

[22] Bertrand, P., Goupil, F.: Descriptive statistics for symbolic data. In: Analysis of Symbolic Data, pp. 106–124. Springer, ??? (2000)

[23] Billard, L., Diday, E.: Symbolic Data Analysis: Conceptual Statistics and Data Mining John Wiley. Chichester (2006)

[24] Duarte Silva, A.P., Brito, P.: Linear discriminant analysis for interval data. Computational Statistics **21**(2), 289–308 (2006)

[25] Utkin, L., Coolen, F.: Interval-valued regression and classification models in the framework of machine learning. ISIPTA 2011 - Proceedings of the 7th International Symposium on Imprecise Probability: Theories and Applications (2011)

[26] Utkin, L.V., Zhuk, Y.A.: An one-class classification support vector machine model by interval-valued training data. Knowledge-Based Systems **120**, 43–56 (2017) https://doi.org/10.1016/j.knosys.2016.12.022

[27] Qi, X., Guo, H., Artem, Z., Wang, W.: An interval-valued data classification method based on the unified representation frame. IEEE Access **8**, 17002–17012 (2020) https://doi.org/10.1109/ACCESS.2020.2967780

[28] Ma, G., Lu, J., Fang, Z., Liu, F., Zhang, G.: Multiview classification through learning from interval-valued data. IEEE Transactions on Neural Networks and Learning Systems, 1–15 (2024) https://doi.org/10.1109/TNNLS.2024.3421657

[29] Wood, M., Stevens, F.: California Mushrooms. https://www.mykoweb.com/CAF/ Accessed 2024-05-25

[30] Nottingham, K., Longjohn, R., Kelly, M.: UCI Machine Learning Repository. https://archive.ics.uci.edu/ Accessed 2024-05-25