Pattern Anal Applic (2018) 21:731–740
https://doi.org/10.1007/s10044-017-0601-3

THEORETICAL ADVANCES

Kernelized inner product-based discriminant analysis for interval data

D. C. F. Queiroz[1] · R. M. C. R. Souza[1] · F. J. A. Cysneiros[2] · M. C. Araujo[3]

Received: 3 March 2016 / Accepted: 18 January 2017 / Published online: 10 February 2017

Abstract This work presents an approach based on the kernelized discriminant analysis to classify symbolic interval data in nonlinearly separable problems. It is known that the use of kernels allows to map implicitly data into a high-dimensional space, called feature space; computing projections in this feature space results in a nonlinear separation in the input space that is equivalent to linear separating function in the feature space. In this work, the kernel matrix is obtained based on kernelized interval inner product. Experiments with synthetic interval data sets and an application with a Brazilian thermographic breast database demonstrate the usefulness of this approach.

Keywords Symbolic data analysis · Supervised classification · Kernel · Linear discriminant analysis

✉ D. C. F. Queiroz
dcfq@cin.ufpe.br

R. M. C. R. Souza
rmcrs@cin.ufpe.br

F. J. A. Cysneiros
cysneiros@de.ufpe.br

M. C. Araujo
marcus.araujo@ufpe.br

[1] Centro de Informática, Universidade Federal de Pernambuco, Av. Jornalista Anibal Fernandes, s/n - Cidade Universitaria (Campus Recife), Recife, PE 50.740-560, Brazil

[2] Departamento de Estatística - CCEN, Universidade Federal de Pernambuco, Av. Jornalista Anibal Fernandes, s/n - Cidade Universitaria (Campus Recife), Recife, PE 50.740-540, Brazil

[3] Departamento de Engenharia Mecânica, CTG, Universidade Federal de Pernambuco, Av. Prof. Moraes Rego, 1235 - Cidade Universitaria (Campus Recife), Recife, PE 50.670-901, Brazil

1 Introduction

With the increasing demand to storage and analyze huge data sets and in order to be able to manage them, it is essential to be able to summarize them while still maintaining as much knowledge inherent to the entire data set as possible. One direct consequence of this problem is that the data may no longer be formatted as single values such as is the case for classical data, but may be represented by lists, intervals, distributions, and the like instead. These summarized data are examples of symbolic data types. Table 1 shows part of a symbolic data set described by intervals.

The breast temperature interval data set was previously considered in [1]. In order to evaluate the feasibility of breast temperature abnormalities (malignant, benign and cyst) and detect breast cancer, they proposed a three-stage feature extraction approach in which breast interval data are extracted from a breast thermography data set, transformed into continuous features and then are used as input data for a classification task. It is composed by 50 breast thermograms of patients aged >35 years with a suspected mass, whose diagnoses were confirmed by clinical examination and followed by ultrasound, mammographic and biopsy exams. Here, the data set is scattered into two classes of different sizes: 14 elements of malignant masses class and 31 elements of non-malignant masses class (composed by elements belonging to benign masses and elements belonging to of cyst masses). Each patient is described by four interval variables that represent the temperature intervals obtained from the left breast (X_1) and the right breast, (X_2) the join between left and right breasts (X_3) and an interval obtained from a morphological processing with both breasts (X_4) as described in [1].

Symbolic data types were defined in symbolic data analysis (SDA) [2]. SDA aims to provide a set of

Table 1 Breast temperature interval data set

X_1	X_2	X_3	X_4	Class
Interval temperature variables				
[31.38, 33.81]	[31.09, 34.03]	[31.09, 34.03]	[0.98, 1.1]	1
[31.5, 35.24]	[30.95, 34.75]	[30.95, 35.24]	[0.69, 1.14]	1
[32.38, 35.93]	[33.11, 36.05]	[32.38, 36.05]	[0.48, 0.59]	1
⋮		⋯		⋮
[34.09, 35.61]	[33.63, 35.57]	[33.63, 35.61]	[0.44, 0.53]	2
[33.21, 35.45]	[32.84, 35.34]	[32.84, 35.45]	[0.9, 1.08]	2
[32.05, 33.35]	[31.65, 33.62]	[31.65, 33.62]	[0.42, 0.54]	2
⋮		⋯		⋮
[30.97, 34.36]	[30.54, 34.43]	[30.54, 34.43]	[1.63, 2.79]	2
[32.86, 35.15]	[32.85, 34.76]	[32.85, 35.15]	[0.65, 1.05]	2
[31.45, 33.8]	[31.49, 33.72]	[31.45, 33.8]	[0.49, 0.75]	2

suitable methods (clustering, factorial techniques, decision trees, etc.) for managing aggregated data described through many types of variables whose values can be sets of categories, intervals or probability distributions in the cells of a data table [2]. A symbolic variable is defined according to its type of domain, i.e., an interval variable takes, for its object, an interval of $\Re$ (the set of real numbers). A symbolic modal one takes, for its object, a nonnegative measure (a frequency or a probability distribution or a system of weights). If this measure is specified in terms of a histogram, the modal variable is called histogram variable.

Several supervised classification tools have been extended to handle interval data: Ichino et al. [3] introduced a symbolic classifier as a region-oriented approach for multi-valued data. In this approach, the classes of examples are described by a region (or set of regions) obtained through the use of an approximation of a mutual neighborhood graph (MNG) and a symbolic join operator. Souza et al. [4] proposed a *MNG* approximation to reduce the complexity of the learning step without losing the classifier performance in terms of prediction accuracy. D'Oliveira et al. [5] presented a region-oriented approach in which each region is defined by the convex hull of the objects belonging to a class.

Ciampi et al. [6] introduced a generalization of binary decision trees to predict the class membership of symbolic data. Rossi and Conan-guez [7] have generalized multi-perceptrons to work with interval data. Mali and Mitra [8] extended the fuzzy radial basis function network to work in the domain of symbolic data. Appice et al. [9] introduced a lazy-learning approach (labeled Symbolic Objects Nearest Neighbor) that extends a traditional distance weighted k-nearest neighbor classification algorithm to interval and modal data. Silva and Brito [10] proposed three approaches to the multivariate analysis of interval data, focusing on linear discriminant analysis and Souza et al. [11] introduced four pattern classifiers based on logistic regression methodology in which these classifiers differ on the way they represent each interval variable.

However, these classification methods for symbolic data were not developed to solve nonlinearly separable problems, that is, problems where elements belonging to one class cannot be separated from elements belonging to another class by a hyperplane, and thus, another approach is needed to solve this family of problems when data are interval-valued. Generalized discriminant analysis [12] (GDA) is a generalization of the classical linear discriminant analysis (LDA) that obtains nonlinear discriminants through kernel functions. This is achieved by formulation as an eigenvalue resolution problem and applies kernel functions to find a feature vector space where the input data becomes linearly separable, similar to the underlying theory on Support Vector Machines.

This work addresses a way in which GDA is generalized for interval data. It changes the inner product used on the core matrix of GDA to the inner product for interval data and then introduces kernelized inner product, allowing the interval-valued data to be kept as intervals while still performing the nonlinear mapping into a feature space. In addition, the proposed approach is applied to a breast temperature abnormality classification problem regarding malignant versus non-malignant classes. Section 2 describes the proposed kernelized discriminant approach for interval data. Section 3 shows the synthetic data sets considered in this work. Section 4 presents the experimental evaluation regarding the synthetic data sets and the Brazilian's thermography breast database displayed in Table 1. Section 5 gives the conclusions.

2 Proposed model

In this section, we present an extension of the GDA [12] to treat interval data, called here Interval Kernel Discriminant Analysis (IKDA). The main idea is to obtain a classifier for interval data that should be able to solve classification problems for classes not linearly separable.

According to the GDA classifier, the IKDA one mainly consists of obtaining a kernel matrix whose elements are composed by the inner product between elements of each class against each other and then incorporates this matrix to the classical linear discriminant analysis, formulating it as an eigenvector problem, then data are projected into a space in which each test data point can be allocated.

Let $X = \{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, N$ be a set of training symbolic objects. Each object i of Ω is described by a set of p symbolic interval variables and a categorical discrete variable. A symbolic interval variable [2] is a

correspondence $\mathfrak{I} \rightarrow \mathfrak{R}$ such that each pattern i is represented by an interval $[a,b] \subseteq \mathfrak{I}$ where $\mathfrak{I} = \{[a,b] : a,b \in \mathfrak{R}, a \leq b\}$ is an interval. Here, the N training symbolic patterns $(\mathbf{x}_i, y_i)$ have $\mathbf{x}_i = (x_{i1} = [a_{i1}, b_{i1}], \ldots, x_{ip} = [a_{ip}, b_{ip}])$ as a vector of interval covariates and y_i as response variable which contains C class labels.

Let $\mathbf{K}$ be a $C \times C$ symmetric block matrix defined over the classes of the training set, whose elements are defined as being matrices themselves:

$$\begin{aligned} &\mathbf{k}_{gh} = (k_{gh})_{lm} \\ &g,h \in \{1,\ldots,C\}, \\ &l \in \{1,\ldots,n_g\}, m \in \{1,\ldots,n_h\} \end{aligned} \tag{1}$$

in which n_g is the number of elements in class g and n_h is the number of elements in class h. In order to propose a kernel matrix regarding p-dimensional interval data space, each pattern is split in p parts and p kernel functions are defined for these parts. Suppose that any point w over the interval $[a_j, b_j]$ for dimension j can be mapped from input data space to a high-dimensional feature space F through a nonlinear function $\phi(w)$:

$$\begin{aligned} &\phi : X \rightarrow F \\ &[a,b] \rightarrow \phi([a,b]) \end{aligned} \tag{2}$$

Consider ϕ as a monotonic nonlinear function defined on real numbers that compose the interval $[a_j, b_j]$. For all $w, r \subseteq [a_j, b_j]$ such that $w \leq r$, ϕ preserves or reverses the order ($\phi(w) \leq \phi(r)$ or $\phi(w) \geq \phi(r)$, respectively), and thus, we do not need to apply ϕ to all real numbers inside the interval, only to its boundaries. Here, $a_j \leq b_j$ so $\phi(a_j) \leq \phi(b_j)$ or $\phi(a_j) \geq \phi(b_j)$.

The main ways in which symbolic interval data arise are aggregation of large data sets. For example, in a breast temperatures matrix, the main interest is to evaluate the feasibility of temperature abnormalities for each breast. All temperature values for each breast are aggregated and their characteristics combined into a single object. In this way, all points inside of the interval $[a_j, b_j]$ can be mapped using the ϕ function. As ϕ is monotonic, interval structure can be preserved. Then, applying this function to the lower and upper bounds of the interval domain still remains in a nonlinear space. An interval in feature space can be defined as:

$$\begin{aligned} [a_j, b_j] &= [\phi(a_j), \phi(b_j)] \quad \text{if } \phi \text{ is monotonically nondecreasing} \\ [a_j, b_j]_\phi &= [\phi(b_j), \phi(a_j)] \quad \text{if } \phi \text{ is monotonically nonincreasing} \end{aligned}$$

2.1 Kernelized inner product for interval data

For data points, this nonlinear mapping is often replaced by an inner-product kernel to obtain the corresponding points in the transformed space. Here, for interval data, we consider to kernelize the interval inner product and to achieve a similar result to the original GDA.

According to [13], given any interval-valued variables $x_r = ([a_{r1}, b_{r1}], \ldots, [a_{rp}, b_{rp}])$ and $x_s = ([a_{s1}, b_{s1}], \ldots, [a_{sp}, b_{sp}])$, the inner product for interval data is given by:

$$\langle \mathbf{x}_r, \mathbf{x}_s \rangle = \begin{cases} \frac{1}{4} \sum_{j=1}^{p} (a_{rj} + b_{rj})(a_{sj} + b_{sj}), & \text{if } \mathbf{x}_r \neq \mathbf{x}_s \\ \frac{1}{3} \sum_{j=1}^{p} (a_{rj}^2 + a_{rj} b_{rj} + b_{rj}^2), & \text{if } \mathbf{x}_r = \mathbf{x}_s \end{cases} \tag{3}$$

Using the Eq. (3) the kernelized inner product can be defined as

$$\langle \mathbf{x}_r, \mathbf{x}_s \rangle_\phi = \begin{cases} \frac{1}{4} \sum_{j=1}^{p} \{\phi(a_{rj}) \cdot \phi(a_{sj}) + \phi(a_{rj}) \cdot \phi(b_{sj}) \\ \quad + \phi(b_{rj}) \cdot \phi(a_{sj}) + \phi(b_{rj}) \cdot \phi(b_{sj})\} \\ \frac{1}{3} \sum_{i=1}^{p} \{\phi(a_{rj}) \cdot \phi(a_{rj}) + \phi(a_{rj}) \cdot \phi(b_{rj}) \\ \quad + \phi(b_{rj}) \cdot \phi(b_{rj})\} \end{cases} \tag{4}$$

under the same restrictions as Eq. (3), that is: if $\mathbf{x}_r \neq \mathbf{x}_s$ and $\mathbf{x}_r = \mathbf{x}_s$, respectively.

Regarding the properties that the sum of kernel functions under the same points input space is a kernel function [14], we can say that $\langle \mathbf{x}_r, \mathbf{x}_s \rangle_\phi$ is a valid kernel.

If $a_{rj} = b_{rj}$ and $a_{sj} = b_{sj}$, we have a particular case

$$\langle \mathbf{x}_r, \mathbf{x}_s \rangle_\phi = \begin{cases} \sum_{i=1}^{p} \phi(a_{rj}) \cdot \phi(a_{sj}), & \text{if } \mathbf{x}_r \neq \mathbf{x}_s \\ \sum_{i=1}^{p} \phi(a_{rj}) \cdot \phi(a_{rj}), & \text{if } \mathbf{x}_r = \mathbf{x}_s \end{cases} \tag{5}$$

The kernel model $\langle \mathbf{x}_r, \mathbf{x}_s \rangle_\phi$ in Eq. (5) is the sum of univariate kernels as a combined kernel for point data. Thus, the kernel model $\langle \mathbf{x}_r, \mathbf{x}_s \rangle_\phi$ in Eq. (4) allows to generalize the traditional kernel to treat interval data

$$\begin{aligned} &\mathbf{k}_{gh} = (k_{gh})_{lm} = \langle \mathbf{x}_l(g), \mathbf{x}_m(h) \rangle_\phi \\ &\qquad g,h \in \{1,\ldots,C\}, \\ &l \in \{1,\ldots,n_g\}, m \in \{1,\ldots,n_h\}. \end{aligned} \tag{6}$$

2.2 Optimization problem

The kernel operator $\mathbf{K}$ allows the construction of nonlinear separating function in the input space that is equivalent to linear separating function in the feature space F. The construction of this function is formulated by maximizing the inter-classes inertia and minimizing the intra-classes inertia.

According to [12], the formulation of this optimizing problem is to need to find eigenvalues λ and eigenvectors $\mathbf{v}$, which are the solutions of the equation:

$$\lambda V\mathbf{v} = B\mathbf{v} \tag{7}$$

The largest eigenvalue of the previous equation gives the maximum of the following quotient of inertia:

$$\lambda = \frac{\mathbf{v}^t\mathbf{B}\mathbf{v}}{\mathbf{v}^t\mathbf{V}\mathbf{v}} \tag{8}$$

where $\mathbf{V}$ and $\mathbf{B}$ represent the total and inter-classes inertia matrices, respectively, in the feature space F.

Because the eigenvectors are linear combinations of elements in F, there exist coefficients $\alpha_{gq}(g = 1, \ldots, C)$ and $(q = 1, \ldots, n_g)$ such that:

$$\mathbf{v} = \sum_{g=1}^{C}\sum_{q=1}^{n_g} \alpha_{gq}\,\phi(x_{g(q)}) \tag{9}$$

The general coefficient vector $\boldsymbol{\alpha} = (\alpha_{gq})$ can be written as $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_g)_{g\in\{1,\ldots,C\}}$ where $\boldsymbol{\alpha}_g = (\alpha_{gq})_{q=1,\ldots,n_g}$; $\boldsymbol{\alpha}_g$ is the coefficient vector of the class g into $\mathbf{v}$.

From Appendix B of [12], the Eq. (8) is equivalent to

$$\lambda = \frac{\boldsymbol{\alpha}^t\mathbf{K}\mathbf{W}\mathbf{K}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^t\mathbf{K}\mathbf{K}\boldsymbol{\alpha}} \tag{10}$$

in which $\mathbf{W}$ is a block diagonal matrix where each one of its elements $\mathbf{W}_g$ is square $n_g \times n_g$ matrices with terms all equal to $1/n_g$ $(g \in \{1, \ldots, C\})$.

The elements of the matrix $\mathbf{K}$ are centered in the feature space according to [12], and the solution of the system in Eq. (10) is given using the eigenvectors decomposition of the matrix $\mathbf{K}$

$$\mathbf{K} = \mathbf{U}\boldsymbol{\Gamma}\mathbf{U}^t \tag{11}$$

where $\boldsymbol{\Gamma}$ is the diagonal matrix of nonzero eigenvalues and $\mathbf{U}$ the matrix of normalized eigenvectors associated to $\boldsymbol{\Gamma}$.

Substituting $\mathbf{K}$ in Eq. (10)

$$\lambda = \frac{\boldsymbol{\alpha}^t\mathbf{U}\boldsymbol{\Gamma}\mathbf{U}^t\mathbf{W}\mathbf{U}\boldsymbol{\Gamma}\mathbf{U}^t\boldsymbol{\alpha}}{\boldsymbol{\alpha}^t\mathbf{U}\boldsymbol{\Gamma}\mathbf{U}^t\mathbf{U}\boldsymbol{\Gamma}\mathbf{U}^t\boldsymbol{\alpha}} = \frac{(\boldsymbol{\Gamma}\mathbf{U}^t\boldsymbol{\alpha})^t\mathbf{U}^t\mathbf{W}\mathbf{U}(\boldsymbol{\Gamma}\mathbf{U}^t\boldsymbol{\alpha})}{(\boldsymbol{\Gamma}\mathbf{U}^t\boldsymbol{\alpha})^t\mathbf{U}^t\mathbf{U}(\boldsymbol{\Gamma}\mathbf{U}^t\boldsymbol{\alpha})} \tag{12}$$

Consider $\boldsymbol{\beta} = \boldsymbol{\Gamma}\mathbf{U}^t\boldsymbol{\alpha}$. So, the Eq. (12) can be rewritten as

$$\lambda\boldsymbol{\beta} = \mathbf{U}^t\mathbf{W}\mathbf{U}\boldsymbol{\beta} \tag{13}$$

For a given $\boldsymbol{\beta}$, there is at least one $\boldsymbol{\alpha}$ satisfying $\boldsymbol{\beta} = \boldsymbol{\Gamma}\mathbf{U}^t\boldsymbol{\alpha}$ in the form:

$$\boldsymbol{\alpha} = \mathbf{U}(\boldsymbol{\Gamma})^{-1}\boldsymbol{\beta}.$$

The coefficients $\boldsymbol{\alpha}$ are normalized by requiring that the corresponding vectors $\mathbf{v}$ be normalized $\mathbf{v}^t\mathbf{v} = 1$ in F. So,

$$\boldsymbol{\alpha} = \frac{\boldsymbol{\alpha}}{\sqrt{\boldsymbol{\alpha}^t\mathbf{K}\boldsymbol{\alpha}}} \tag{14}$$

Given the normalized eigenvectors $\mathbf{v}$, we can obtain the projection vector of an element represented by $\mathbf{x}$ on $\mathbf{v}$ as

$$\mathbf{z}(\mathbf{x}) = \mathbf{v}^t\phi(\mathbf{x}) = \sum_{g=1}^{C}\sum_{l=1}^{n_g} \alpha_{gl}\,\langle\mathbf{x}_l(g), \mathbf{x}\rangle_\phi. \tag{15}$$

2.2.1 The algorithm

The IKDA algorithm is summarized as follows:

(1. Input Data)

Consider a C-class interval data set of n objects described by $(\mathbf{x}_i, y_i)$ as: $\mathbf{x}_i = ([a_{i1}, b_{i1}], \ldots, [a_{inv}, b_{inv}])$ $y_i = 1, \ldots, C$.

- Obtain the symmetric block matrix $\mathbf{K}$ where each element $\mathbf{k}_{gh}$ is a matrix whose elements $(k_{gh})_{lm} = \langle\mathbf{x}_l(g), \mathbf{x}_m(h)\rangle_\phi$ are obtained by the inner product given in Equation (4);
- Obtain the diagonal block matrix $\mathbf{W}$, where each element $(\mathbf{W})_g$, $g = 1, \ldots, C$ is a matrix with elements all equal to $1/n_g$, which is the number of elements in each class.

(2. Obtaining the projection onto discriminant space)

- Compute the eigenvector decomposition of the matrix $\mathbf{K}$ as in Equation (11);
- Compute the $\boldsymbol{\beta}$ eigenvectors of the problem and each corresponding coefficient $\boldsymbol{\alpha}$ solutions in Equation (13).;
- Compute $\mathbf{v}$ and normalize them by computing α as given in Equation(14)
- Compute projection vector $\mathbf{z}_i$ $(i = 1, \ldots, N)$ of Ω using Equation (15);
- Compute the mean project vector $\mathbf{L}_g$ of the class $g \in \{1, \ldots, C\}$ as

$$\mathbf{L}_g = \frac{1}{n_g}\sum_{m=1}^{n_g} \mathbf{z}(\mathbf{x}_m).$$

(3. Allocation step)

- Given an test element, compute its projection vector $\mathbf{z}$ using Equation (15).
- Affect $\mathbf{z}$ is to class g if and only if:
  - $D(\mathbf{z}, \mathbf{L}_g) \leq D(\mathbf{z}, \mathbf{L}_h)\forall g, h \in \{1, \ldots, C\}$

where D is a distance measure.

3 Three synthetic interval data sets

In this section, three different data sets are presented: two synthetic interval data sets with synthetic seeds and one synthetic data set with real data seeds.

3.1 Two synthetic interval data sets with synthetic seeds

The procedure to generate synthetic interval data sets based on synthetic seeds consists of two steps:

- To obtain a seed data set with classical variables.
- To consider variability for seed data in order to generate a synthetic interval data set.

To obtain these synthetic interval data set, two standard synthetic quantitative data sets are generated and used as seeds to obtain the synthetic interval data sets. With regard to the two standard synthetic quantitative data sets, both are

generated in $\Re$, and therefore, they have two standard continuous quantitative variables.

The first data set has 100 points scattered among two classes. Each class is a defined as an upper and lower halves of a circumference generated from data in the same uniform distribution plus Gaussian noise, then the upper class was shifted to increase the proximity between classes. The second data set has 150 points distributed in two classes of unequal sizes, the first class has 100 points, and the second has 50. Both classes were designed as circumferences with the same origin, but each class has a different radius and is generated from data in an independent uniform distribution with Gaussian noise.

The quantitative data set 1 is generated by the following parameters:

$$\begin{aligned}
&\textit{Class 1} \quad X_1 \sim U(5,25)\\
&\qquad\quad X_2 = \sqrt{100-(X_1-15)^2}+20\\
&\qquad\quad \text{noise} \sim N(0,1)\\
&\qquad\quad S_{X_1} = 10\\
&\qquad\quad S_{X_2} = -3\\
&\textit{Class 2} \quad X_1 \sim U(5,25)\\
&\qquad\quad X_2 = \sqrt{100-(X_1-15)^2}+20\\
&\qquad\quad \text{noise} \sim N(0,1)
\end{aligned}$$

The quantitative data set 2 is generated by the following parameters:

$$\begin{aligned}
&\textit{Class 1} \quad X_1 \sim U(0,40)\\
&\qquad\quad X_2 = \sqrt{400-(X_1-20)^2}+20\\
&\qquad\quad \text{noise} \sim N(0,1)\\
&\textit{Class 2} \quad X_1 \sim U(15,25)\\
&\qquad\quad X_2 = \sqrt{25-(X_1-20)^2}+20\\
&\qquad\quad \text{noise} \sim N(0,1)
\end{aligned}$$

X_1 is the first coordinate, X_2 is given by the circle equation, S_{X_1} and S_{X_2} are the values added to each coordinate to force class 1 closer to class 2 and *noise* is a value added to the X_2 coordinate. Now to generate symbolic data sets from these two standard quantitative data sets, a procedure where each variable is expanded to form an interval is used.

Each data point (x_1, x_2) of each one of these synthetic quantitative data sets is a seed for a vector of intervals (rectangle) through the following procedure:

$$([x_1-\gamma_1/2, x_1+\gamma_1/2], [x_2-\gamma_2/2, x_2+\gamma_2/2])$$

where these parameters γ_1 and γ_2 are randomly selected from a predefined interval $[1,5]$, $[1,10]$ or $[1,15]$.

Therefore, from each element in the standard data set, we generate an interval element on the synthetic interval data set. Figure 1 presents an example of the generation of the symbolic data sets described in this section, on the left side, synthetic data set 1 and its corresponding symbolic counterpart, on the right side, synthetic data set 2 and its corresponding symbolic counterpart. These examples were generated by choosing both γ_1 and γ_2 from the $[1,5]$ interval.

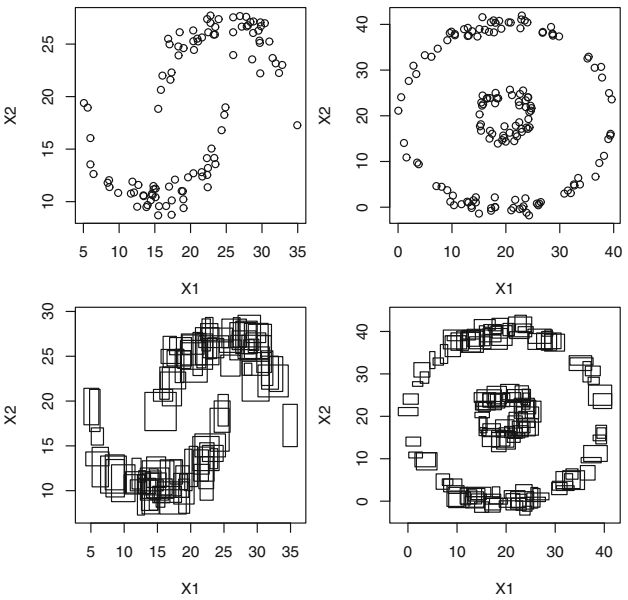

Fig. 1 Quantitative data sets 1 and 2 and their correspondent Symbolic data sets

3.2 A synthetic interval data set with real seeds: interval Iris data

As a different study case for our method, we analyze Fisher's Iris flower data set which is a typical test case used by the machine learning community. This classical data set consists of 3 classes described by 4 continuous variables that correspond to the sepal and petal length and width of each element.

Given the different nature of data our classifier is supposed to address, we subject the database to the same procedure used to generate synthetic interval-valued data in previous subsection.

That is, the original Iris data set is subjected to the same procedure as the synthetic data sets 1 and 2 to generate a symbolic Iris data set, whose variables are interval variables. The generation parameters γ_1 and γ_2 were chosen from the same intervals used on the synthetic data sets. Table 2 shows partially the resulting data set.

4 Experimental evaluation

In this section, the experimental evaluation is presented. The proposed classifier (IKDA) is evaluated and compared against three other classifiers:

Table 2 Interval iris data set from original variables

X_1	X_2	X_3	X_4	Class
Iris data set				
[3.3857, 6.8143]	[1.6483, 5.3517]	[0.6034, 2.1966]	$[-3.1042, 3.5042]$	setosa
[2.0835, 7.7165]	$[-0.3997, 6.3997]$	[0.0071, 2.7929]	$[-0.7503, 1.1503]$	setosa
[0.0586, 9.3414]	$[-0.4843, 6.8843]$	[0.1645, 2.4355]	$[-0.8342, 1.2342]$	setosa
[3.392, 5.808]	[1.5703, 4.6297]	$[-0.0466, 3.0466]$	$[-2.7881, 3.1881]$	setosa
[2.0601, 7.9399]	[0.5454, 6.6546]	$[-2.2488, 5.0488]$	$[-3.8565, 4.2565]$	setosa
[3.2657, 7.5343]	$[-0.2905, 8.0905]$	[0.0098, 3.3902]	$[-1.581, 2.381]$	setosa
[3.5732, 5.6268]	$[-0.6209, 7.4209]$	$[-2.0674, 4.8674]$	$[-2.626, 3.226]$	setosa
[0.8362, 9.1638]	$[-0.8858, 7.6858]$	$[-3.249, 6.249]$	$[-1.8626, 2.2626]$	setosa
⋮	⋮	⋮	⋮	⋮
[4.2421, 9.7579]	[0.7487, 5.6513]	[3.8281, 5.5719]	$[-2.675, 5.475]$	versicolor
[4.4519, 8.3481]	$[-1.219, 7.619]$	[2.4064, 6.5936]	$[-1.1289, 4.1289]$	versicolor
[3.2846, 10.5154]	[1.3495, 4.8505]	[1.4284, 8.3716]	[0.4523, 2.5477]	versicolor
[3.1032, 7.8968]	$[-1.9712, 6.5712]$	[1.6533, 6.3467]	[0.5772, 2.0228]	versicolor
[4.668, 8.332]	$[-1.2205, 6.8205]$	[1.6073, 7.5927]	$[-3.3725, 6.3725]$	versicolor
[2.9017, 8.4983]	[1.1164, 4.4836]	[0.943, 8.057]	[0.1527, 2.4473]	versicolor
[1.7103, 10.8897]	[1.4133, 5.1867]	[0.7691, 8.6309]	$[-2.4553, 5.6553]$	versicolor
[3.2687, 6.5313]	$[-0.5674, 5.3674]$	[1.4328, 5.1672]	$[-0.4272, 2.4272]$	versicolor
⋮	⋮	⋮	⋮	⋮
[2.6082, 10.9918]	[1.8836, 4.5164]	[2.6022, 9.1978]	[1.6735, 2.9265]	virginica
[3.9923, 9.4077]	[2.7687, 3.8313]	[2.404, 8.996]	[1.7886, 3.2114]	virginica
[4.2811, 9.1189]	$[-1.4358, 7.4358]$	[2.4587, 7.9413]	$[-0.856, 5.456]$	virginica
[3.6448, 8.9552]	$[-1.2468, 6.2468]$	[3.0853, 6.9147]	$[-1.7393, 5.5393]$	virginica
[5.6545, 7.3455]	$[-1.8079, 7.8079]$	[4.0121, 6.3879]	$[-0.3619, 4.3619]$	virginica
[4.6869, 7.7131]	[0.207, 6.593]	[2.7139, 8.0861]	[1.7564, 2.8436]	virginica
[4.7226, 7.0774]	$[-0.1174, 6.1174]$	[2.4278, 7.7722]	[0.4576, 3.1424]	virginica

- Logistic Regression classifier (LOGIT) where two regressions are adjusted for each class, one regarding the lower bounds of the interval variables and another regarding the upper bounds of the interval variables, allocation is given by the average of the response obtained by each regression.
- Linear Discriminant Analysis for Interval Data (ILDA), using the distributional approach with either definitions A or B found in [10] and Hausdorff distance for interval data (ILDA-A refers to ILDA using definition A and ILDA-B refers to ILDA using definition B).

In our experiments with IKDA proposed method, the following elements were considered:

- Polynomial kernel with degree $d = 1, 2, 3, 4, 5$ and Gaussian kernel with width $\sigma = 0.5, 1, 3, 5, 7$.
- Euclidean distance in the allocation step.

Prediction accuracy is measured by the error rate of classification which is estimated by a Monte Carlo simulation for the simulated data set with 500 replications, through a tenfold cross-validation for the synthetic data set with real seed and through the leave-one-out method for the real data sets. On the framework of a Monte Carlo simulation, test and learning sets are randomly selected from each synthetic interval data set. The learning set corresponds to 75% of the original data, and the test data set corresponds to 25%.

4.1 Synthetic data sets with synthetic seeds

Tables 3 and 4 present the average and standard deviation (in parenthesis) of the error rate for IKDA method and interval data set 1 and Tables 5 and 6 for IKDA method and interval data set 2. Table 7 shows error rate averages for LOGIT and ILDA methods. From the results in these tables, some remarks are listed.

- For interval data set 1, it can also be seen that the small increase in the value of the parameters did not cause an increase in performance when the polynomial kernel was used; however, it was the opposite when the Gaussian kernel was used.
- For interval data set 1, the best result regarding polynomial kernel is with $d = 1$ and the best result

Table 3 Average (in %) and standard deviation of the error rate for IKDA approach, synthetic data set 1 and polynomial kernel

Chosen γ	$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 5$
[1, 5]	3.33	5.56	6.14	6.36	7.42
	(3.26)	(3.67)	(4.38)	(4.41)	(4.47)
[1, 10]	2.00	4.00	4.62	7.45	8.67
	(2.56)	(3.36)	(3.80)	(4.49)	(4.59)
[1, 15]	2.14	3.49	5.38	6.15	7.03
	(2.63)	(3.14)	(4.07)	(4.19)	(4.50)

Table 4 Average (in %) and standard deviation of the error rate for IKDA approach, synthetic data set 1 and Gaussian kernel

Chosen γ	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 3$	$\sigma = 5$	$\sigma = 7$
[1, 5]	4.28	3.79	2.32	2.09	2.29
	(3.69)	(3.48)	(2.79)	(2.67)	(2.68)
[1, 10]	4.45	4.62	1.37	0.90	0.48
	(3.78)	(3.93)	(2.61)	(2.39)	(1.74)
[1, 15]	9.21	7.80	3.00	2.50	2.45
	(5.20)	(4.99)	(3.33)	(2.85)	(2.77)

Table 5 Average (in %) and standard deviation of the error rate for IKDA approach, synthetic data set 2 and polynomial kernel

Chosen γ	$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 5$
[1, 5]	50.22	35.97	32.52	33.28	33.09
	(9.04)	(6.26)	(5.90)	(5.72)	(5.96)
[1, 10]	49.60	35.40	31.76	31.62	31.65
	(9.03)	(6.40)	(5.54)	(5.78)	(5.96)
[1, 15]	51.39	17.07	13.54	13.70	13.93
	(4.83)	(2.87)	(2.44)	(2.47)	(2.55)

Table 6 Average (in %) and standard deviation of the error rate for IKDA approach, synthetic data set 2 and Gaussian kernel

Chosen γ	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 3$	$\sigma = 5$	$\sigma = 7$
[1, 5]	7.37	4.87	2.71	2.71	2.43
	(4.01)	(3.73)	(2.90)	(2.84)	(2.63)
[1, 10]	15.27	10.56	5.48	3.14	2.28
	(5.68)	(5.21)	(3.74)	(3.12)	(2.80)
[1, 15]	8.01	7.62	4.85	3.78	3.27
	(2.40)	(2.33)	(2.11)	(1.80)	(1.71)

regarding Gaussian kernel is with $\sigma = 7$. Under these conditions the Gaussian kernel is slightly superior polynomial kernel. This is expected since the interval data set 1 has weak nonlinear separation when compared to interval data set 2.

- For interval data set 2, which shows a greater degree of nonlinear separation than that of the interval data set 1, the Gaussian kernel is superior to the polynomial kernel for any value of $\sigma \in \{0.5, 1, 3, 5, 7\}$.
- The linear classifiers obtained bad performance, overall only comparable with the worse results from the nonlinear classifiers.

4.2 Synthetic data sets with real seeds

Tables 8 and 9 present the average and standard deviation of the error rate for the IKDA classifier regarding the synthetic interval data set with real seeds using for γ the intervals [1, 5], [1, 10] and [1, 15]. Table 10 shows the average and standard deviation of the error rate for LOGIT, ILDA-A and ILDA-B classifiers.

The results in these tables show that the IKDA method with polynomial kernel had better performance than Gaussian kernel for parameter $d = 1$ for each interval of uncertainty introduced in the data set; however, despite the polynomial kernel obtaining the best results, overall the Gaussian kernel was more successful. This effect can be due to the fact that the original iris data set has linearly separable classes and the polynomial kernel is similar to a linear model, being well adjusted for the parameters chosen. The linear classifiers had overall lower accuracy than both methods using kernels.

4.3 Real breast temperature interval data set

As stated in [15], "Most work on the analysis of breast thermal images provide classification results using the accuracy, specificity and sensitivity measures or/and also present the corresponding ROC curves of their methods," this is mostly due to a type I error approach, that is, most works are interested in classifying correctly malignant abnormalities class more than other classes (also reflected in our representation of this problem as a binary problem). Global misclassification/accuracy alone analyzes the overall correctness of classification, but cannot identify if the class of interest has a good detection rate, which justifies other measures being calculated and presented together with accuracy/misclassification values. Researchers in the medical field value sensitivity [16–18] because classifying wrongly patients that should be allocated to the malignant abnormalities class may lead directly to their death.

Therefore, in our analysis we prioritize sensitivity followed by global misclassification rate in this specific order. Table 11 presents confusion matrices for the IKDA proposed method using polynomial kernel with parameter $d = 1$, $d = 2$, $d = 3$, $d = 4$ and $d = 5$ and Table 12 presents confusion matrices for the IKDA proposed method using Gaussian kernel with $\sigma = 0.5$, $\sigma = 1$, $\sigma = 3$, $\sigma = 5$

Table 7 Average (in %) and standard deviation of the error rate for LOGIT and ILDA classifiers and synthetic data sets 1 and 2

Chosen γ	Synthetic data set 1			Synthetic data set 2		
	LOGIT	ILDA-A	ILDA-B	LOGIT	ILDA-A	ILDA-B
[1, 5]	50.00	50.00	50.00	32.45	66.65	66.23
	(0.00)	(0.00)	(0.00)	(0.38)	(5.59)	(6.72)
[1, 10]	50.00	50.00	50.00	32.49	64.26	63.28
	(0.00)	(0.00)	(0.00)	(0.87)	(10.26)	(11.51)
[1, 15]	50.00	50.00	50.00	71.27	28.73	31.37
	(0.00)	(0.00)	(0.00)	(0.20)	(0.00)	(10.26)

Table 8 Average (in %) and standard deviation of the error rate for IKDA approach, synthetic interval data set with real seeds and polynomial kernel

Chosen γ	$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 5$
[1, 5]	6.46	10.53	11.86	15.93	19.80
	(5.48)	(7.68)	(7.14)	(8.17)	(10.56)
[1, 10]	4.06	29.26	27.73	38.13	35.46
	(5.43)	(11.25)	(11.70)	(13.19)	(12.89)
[1, 15]	6.46	50.26	33.40	50.00	40.20
	(5.48)	(13.49)	(11.01)	(10.93)	(11.03)

Table 9 Average (in %) and standard deviation of the error rate for IKDA approach, synthetic interval data set with real seeds and Gaussian kernel

Chosen γ	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 3$	$\sigma = 5$	$\sigma = 7$
[1, 5]	24.20	21.86	16.80	14.80	13.60
	(6.96)	(7.45)	(7.17)	(7.42)	(7.54)
[1, 10]	17.40	11.80	9.86	9.80	8.93
	(9.14)	(8.16)	(7.91)	(7.13)	(7.58)
[1, 15]	18.26	14.46	9.40	8.86	9.20
	(9.44)	(7.70)	(6.38)	(6.18)	(6.21)

Table 10 Average (in %) and standard deviation of the error rate for LOGIT and ILDA classifiers and synthetic interval data set with real seeds

Chosen γ	LOGIT	ILDA-A	ILDA-B
[1, 5]	65.20	66.67	66.67
	(3.21)	(0.00)	(0.00)
[1, 10]	66.67	66.67	66.67
	(0.00)	(0.00)	(0.00)
[1, 15]	66.67	66.67	66.67
	(0.00)	(0.00)	(0.00)

and $\sigma = 7$, respectively. The best performances of the IKDA method are achieved with $d = 1$ and $\sigma = 5$ and 7 for polynomial and Gaussian kernels, respectively.

Table 13 displays the confusion matrices for the LOGIT, ILDA-A and ILDA-B classifiers. The LOGIT method is inferior to the ILDA-A and ILDA-B ones in terms of correct predicted classifications of malign abnormalities class, but superior in terms of overall correct predicted classifications.

Table 11 Confusion matrix for the IKDA classifier with polynomial kernel

Class	Predict		Total
	Non-malign	Malign	
$d = 1$			
Non-malign	26	10	36
Malign	4	10	14
Total	30	20	50
$d = 2$			
Non-malign	27	9	36
Malign	6	8	14
Total	33	17	50
$d = 3$			
Non-malign	27	9	36
Malign	6	8	14
Total	33	17	50
$d = 4$			
Non-malign	27	9	36
Malign	6	8	14
Total	33	17	50
$d = 5$			
Non-malign	26	10	36
Malign	5	9	14
Total	31	19	50

The global misclassification rate and sensitivity index are computed from the previous tables. Sensitivity index represents the proportion of actual positives samples which are correctly identified as such and plays an important role in medical field as it related to the ratio between the true positive and true negative observations. The sensitivity can be calculated as

$$\mathrm{Sen}(i) = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \tag{16}$$

where TP_i = True positive for class i and FN_i = False Negative for class i.

Table 12 Confusion matrix for the IKDA classifier with Gaussian kernel

Class	Predict		Total
	Non-malign	Malign	
$\sigma = 0.5$			
Non-malign	25	11	36
Malign	8	6	14
Total	33	17	50
$\sigma = 1$			
Non-malign	25	11	36
Malign	9	5	14
Total	34	16	50
$\sigma = 3$			
Non-malign	26	10	36
Malign	8	6	14
Total	34	16	50
$\sigma = 5$			
Non-malign	28	8	36
Malign	3	11	14
Total	31	19	50
$\sigma = 7$			
Non-malign	28	8	36
Malign	3	11	14
Total	31	19	50

Table 13 Confusion matrix for the LOGIT, ILDA-A and ILDA-B classifiers

Class	Predict		Total
	Non-malign	Malign	
LOGIT			
Non-malign	30	6	36
Malign	4	10	14
Total	34	16	50
ILDA-A			
Non-malign	26	10	36
Malign	3	11	14
Total	29	21	50
ILDA-B			
Non-malign	26	10	36
Malign	3	11	14
Total	29	21	50

The overall misclassification rate and sensitivity index for the malignant class and IKDA, LOGIT, ILDA-A and ILDA-B methods are presented in Table 14. The results show that the best value of sensitivity index, which is extremely important for medical studies, is achieved with the IKDA method using Gaussian kernel ($\sigma = 5$ and $\sigma = 7$) and ILDA-A and ILDA-B models. Among these three methods, the IKDA one had the best overall misclassification rate.

Table 14 Misclassification rate and sensitivity index for malignant class and IKDA, ILDA-A, ILDA-B and LOGIT methods

Classifier	Misclassification rate (%)	Sensitivity index (%)
$IKDA_p$ ($d = 1$)	28	71
$IKDA_g$ ($\sigma = 5$ and 7)	22	78
LOGIT	20	71
ILDA-A	26	78
ILDA-B	26	78

5 Conclusions

This work introduced a kernelized classifier for interval-valued data. It was based on the generalized discriminant analysis (GDA) for its ability to solve nonlinearly separable problems. Here, the inner product for interval is kernelized as a resulting summation of multiple identical kernel functions applied to different bounds of each interval-valued variable. The proposed method is a generalization of the GDA to treat symbolic interval data regarding nonlinearly separable classes.

Two types of kernel functions were used to evaluate the behavior of the proposed classifier. Its performance was assessed by the global error rate based on different configurations of synthetic interval data sets. An application with a Brazilian's thermography breast database was considered, and the performance was assessed by the sensitivity index, which is extremely important for medical studies and global misclassification rate. The study of performance analysis allowed to confirm the usefulness of the proposed method in regard to interval data in nonlinearly separable class problems when compared with other classifiers of the symbolic data analysis literature.

Acknowledgements The authors wish to thank the Editor-in-Chief, Professor Sameer Singh and two anonymous referees for their constructive comments on an earlier version of this manuscript. This research work was partially supported by a CNPq, CAPES and FACEPE agency from Brazil.

References

1. Araujo MC, Lima RCF, Souza RMCR (2014) Interval symbolic feature extraction for thermography breast cancer detection. Expert Syst Appl 41:6728–6737
2. Bock HH, Diday E (2000) Analysis of symbolic data: exploratory methods for extracting statistical information from complex data. Springer, Heidelberg

3. Ichino M, Yaguchi H, Diday E (1996) A fuzzy symbolic pattern classifier. In: Ordinal and symbolic data analysis. Springer, pp 92–102
4. Souza RMCR, De Carvalho FAT, Frery AC (1999) Symbolic approach to SAR image classification. In: IEEE international geoscience and remote sensing symposium
5. D'Oliveira ST, de Carvalho FAT, Souza RMCR (2004) Classification of SAR images through a convex hull region oriented approach. In: Pal N, Kasabov N, Mudi R, Pal S, Parui S (eds) Neural information processing, volume 3316 of Lecture Notes in Computer Science. Springer, Berlin, pp 769–774
6. Ciampi A, Diday E, Lebbe J, Prinel E, Vignes R (2000) Growing a tree classifier with imprecise data. Pattern Recogn Lett 21(9):787–803
7. Rossi F, Conan-guez B (2002) Multi-layer perceptron on interval data. In: Classification, clustering and data analysis (IFCS, 2002), pp 427–434
8. Mali K, Mitra S (2005) Symbolic classification, clustering and fuzzy radial basis function network. Fuzzy Sets Syst 152(3):553–564
9. Appice A, D'Amato C, Esposito F, Malerba D (2006) Classification of symbolic objects: a lazy learning approach. Intell Data Anal 10(4):301–324
10. Duarte Silva AP, Brito P (2006) Linear discriminant analysis for interval data. Comput Stat 21:289–308
11. Souza RMCR, Queiroz DCF, Cysneiros FJA (2011) Logistic regression-based pattern classifiers for symbolic interval data. Pattern Anal Appl 14(3):273–282
12. Baudat G, Anouar F (2000) Generalized discriminant analysis using a kernel approach. Neural Comput 12(10):2385–2404
13. Wang H, Guan R, Wu J (2012) Linear regression of interval-valued data based on complete information in hypercubes. J Syst Sci Syst Eng 21(4):422–442
14. Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines: and other kernel-based learning methods. Cambridge University Press, New York
15. Borchartt TB, Conci A, Lima RCF, Resmini R, Sanchez A (2013) Breast thermography from an image processing viewpoint: a survey. Signal Image Process Tech Detect Breast Dis 93(10):2785–2803
16. Wishart GC, Campisi M, Boswell M, Chapman D, Shackleton V, Iddles S, Hallett A, Britton PD (2010) The accuracy of digital infrared imaging for breast cancer detection in women undergoing breast biopsy. Eur J Surg Oncol 36(6):535540
17. Acharya UR, Ng EY, Tan JH, Sree SV (2012) Thermography based breast cancer detection using texture features and support vector machine. J Med Syst 36:1503–1510
18. Mookiah MRK, Acharya UR, Ng E (2012) Data mining technique for breast cancer detection in thermograms using hybrid feature extraction strategy. Quant InfraRed Thermogr J 9(2):151–165