Information Sciences 372 (2016) 591–607

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Distance-based linear discriminant analysis for interval-valued data

Ana B. Ramos-Guajardo[a,*], Przemyslaw Grzegorzewski[b]

[a] Department of Statistics, Operational Research and Mathematic Didactics, University of Oviedo, Spain
[b] Systems Research Institute, Polish Academy of Sciences and Faculty of Mathematics and Information Science, Warsaw University of Technology, Poland

ARTICLE INFO

Article history:
Received 14 January 2016
Revised 31 July 2016
Accepted 19 August 2016
Available online 24 August 2016

Keywords:
Classification
Discriminant analysis
Interval data
LDA
Random intervals

ABSTRACT

Interval-valued observations arise in many real-life situations as either the precise representation of the objective entity or the representation of incomplete knowledge. Thus given p features observed over a sample of objects belonging to one of two possible classes, each observation can be perceived as a non-empty closed and bounded hyperrectangle on $\mathbb{R}^p$. The aim of the paper is to suggest a p-dimensional classification method for random intervals when two or more classes are considered, by the generalization of Fisher's procedure for linear discriminant analysis. The idea consists of finding a directional vector which maximizes the ratio of the dispersion between the classes and within the classes of the observed hyperrectangles. A classification rule for new observations is also provided and some simulations are carried out to compare the behavior of the proposed classification procedure with respect to other methods known from the literature. Finally, the suggested methodology are applied on a real-life situation example.

1. Introduction

Interval-valued data have drawn an increasing interest in recent years. Quite often a real random variable is imprecisely observed or is so uncertain that the results are recorded as real intervals containing the precise outcomes of the experiment. Sometimes the exact value of a variable is hidden deliberately due to some confidentiality reasons. In such cases as mentioned above intervals are considered as disjunctive sets representing incomplete information (*epistemic view* according to [14]). There are also other situations in which the experimental data appear as essentially interval-valued data describing a precise information (as, e.g., ranges of fluctuations of some physical measurements, time interval spanned by an activity, etc.). Such intervals are called conjunctive and correspond to the *ontic view* (see [14]). Interval data appear also as a kind of the so-called *Symbolic Data Analysis* summarizing the information stored in large data sets (see [5]).

The problem of interval data in regression analysis, time series, hypothesis testing and principal component analysis has been extensively elaborated (see, e.g.,[6,7,9,10,16,17,43–45] for the most recent results). Different problems concerning clustering methods for interval-valued data have been previously tackled in the literature [18,20–23,29,30]. Some classification algorithms for interval-valued data have been developed based on the imprecise probability theory in [41,47,48]. Other classification and regression models that consider different distances between intervals or hyper-rectangles have been also

* Corresponding author.
E-mail addresses: ramosana@uniovi.es (A.B. Ramos-Guajardo), pgrzeg@ibspan.waw.pl (P. Grzegorzewski).

http://dx.doi.org/10.1016/j.ins.2016.08.068
0020-0255/

developed (see, e.g., [11,33,42]). In addition, in [13] a procedure for supervised classification of fuzzy data (also useful for interval data since they are a particular case of fuzzy data) has been introduced based on a non-parametric kernel density estimation procedure which considers conditional probabilities on certain balls.

The are only a few papers on discriminant analysis for interval-valued data, see [1,24,25,34,38,41,49]. On one hand, the extension of DEA-DA (i.e. a method based on mathematical programming and designed to identify the existence or non-existence of an overlap between two groups) upon interval data was considered in [34,38,49]. On the other hand, a knowledge-based support vector machine (SVM) linear classifier derived from convex optimization theory and inserting information into the standard SVM in the form of intervals was suggested in [1]. Finally, the discriminant analysis methods for interval data developed in [24] and [25] are described below.

In [24] Duarte Silva and Brito proposed three methods for handling interval data in the typical way for Symbolic Data Analysis. In the first approach they assume that each interval represents the possible values of an underlying real-valued variable which are uniformly distributed there. Then they derive the measures of dispersion and define appropriately linear combinations of intervals which maximize the usual discriminant criterion. It seems that the assumed hypothesis on the uniform mixture of variables might be considered too strong. Their second approach considers only the vertices of the hypercubes (but not the intervals themselves) and proceeds with a classical discriminant analysis based on the set of all interval description vertices. However, this method leads to a loss of the idea of interval-valued data. The third method is similar to the second one but the authors use there a different representation of the intervals: i.e. theirs midpoints and ranges. Then they perform two separate classical discriminant analysis on midpoints and ranges, respectively, and combine the results in some way. Nevertheless again the idea of interval-data is somehow lost. The methods mentioned above are carried out in the framework of Symbolic Data Analysis and might be considered as nonparametric exploratory approaches. In this context the approach proposed in [25] is a parametric one. It develops the idea of [8] which consists of representing intervals by a bivariate normal distribution (one variable responsible for the midpoints and the second for the logarithm of ranges) and then applying the classical linear or quadratic discriminant classification rules.

In this contribution we propose a general approach to the problem of interval-valued data classification. In contrast to those approaches where intervals are replaced by points to make classical linear discriminant analysis, we consider intervals themselves, without loosing any information delivered by the interval structure. The main advantage of our method is that it does not involve any function to transform the spreads in interval variables to take values in $\mathbb{R}$, but the main idea consists of finding a directional vector which maximizes the ratio of the between-classes dispersion and the within-classes dispersion of the observed hyperrectangles inspired by the ideas of Fisher's linear discriminant analysis [27]. Due to the non-linearity of the class of closed intervals, both the between-classes dispersion and the within-classes dispersion are defined on the basis of a generalized distance firstly introduced in [46] instead of the usual differences. A classification rule for new observations is also established and a comparative study with respect the three methods proposed in [24] and the classical LDA based on the mid-points method is also carried out.

The paper is organized as follows: in Section 2 we introduce some basic concepts and notation related to interval data analysis including random intervals and their features. Then, in Section 3 we recall briefly the principles of the classical two-dimensional Fisher's linear discriminant analysis. Our proposal about the extension of the Fisher's linear discriminant to random intervals is addressed in Section 4. Next, we verify the behavior of the suggested classification procedure with respect other procedures on LDA for random intervals by means of some comparative simulation studies (Section 5). Finally, in Section 6, we illustrate the proposed methodology in a real-life situation, i.e. in the Greek wine classification example.

2. Interval data and random intervals

Let $\mathcal{K}_c(\mathbb{R})$ be the family of all non-empty closed and bounded intervals of $\mathbb{R}$. Each interval $A \in \mathcal{K}_c(\mathbb{R})$ is usually parametrized by means of a two-dimensional value, defined in terms of its endpoints, $(\inf A, \sup A) \in \mathbb{R}^2$ with $\inf A \leq \sup A$. Other characterization of an interval, which is in some situations more operative, is based on the point $(\mathrm{mid}\, A, \mathrm{spr}\, A) \in \mathbb{R} \times \mathbb{R}^+$, where $\mathrm{mid}\, A = (\sup A + \inf A)/2$ is the midpoint of the interval, and $\mathrm{spr}\, A = (\sup A - \inf A)/2$ denotes the spread or radius. Thus, A can be represented as $A = [\inf A, \sup A] = [\mathrm{mid}\, A \pm \mathrm{spr}\, A]$.

When dealing with intervals, a natural arithmetic is defined on $\mathcal{K}_c(\mathbb{R})$ based on the Minkowski addition [40] and the product by scalars. These operations are settled as

$$A + B = \{a + b : a \in A, b \in B\} \quad \text{and} \quad \lambda A = \{\lambda a : a \in A\},$$

for all $A, B \in \mathcal{K}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$, respectively. Using the mid/spr notation the above operations can be jointly expressed as follows

$$A + \lambda B = [(\mathrm{mid}\, A + \lambda\, \mathrm{mid}\, B) \pm (\mathrm{spr}\, A + |\lambda|\, \mathrm{spr}\, B)]. \tag{1}$$

It should be noted that the space $(\mathcal{K}_c(\mathbb{R}), +, \cdot)$ is not linear but semilinear, due to the lack of the inverse element with respect to the Minkowski addition. To overcome this problem, sometimes it is possible to consider the so-called Hukuhara difference [32] $A -_H B$ between the intervals A and B, defined by such element $C \in \mathcal{K}_c(\mathbb{R})$ that $B + C = A$, if it exists, and which always exists whenever $\mathrm{spr}\, B \leq \mathrm{spr}\, A$.

Let $(\Omega, \mathcal{A}, P)$ be a probability space. A *random interval* (for short RI) is a Borel measurable mapping $X : \Omega \longrightarrow \mathcal{K}_c(\mathbb{R})$, with the Borel measurability being intended w.r.t. the well-known Hausdorff metric on $\mathcal{K}_c(\mathbb{R})$ (see [39]). Equivalently, a mapping