Distance based Depth-Depth classifier for directional data

DD-classifier per dati direzionali basati su funzioni di distance-depths

Giuseppe Pandolfo and Giovanni C. Porzio

Sommario The DD-classifier, which has been extended to the classification of directional objects, is here investigated in the case of some new distance-based directional depths. The DD-classifier is a non-parametric techniques based on the depth vs. depth (DD) plot. Its main advantages concern the flexibility and the independence from specific distribution parameters. The new depth functions adopted here allow using them for high-dimensional directional data sets.

Sommario *Il DD-classifier, che è stato esteso alla classificazioni di dati direzionali, viene qui analizzato nel caso di due nuove funzioni di depth recentemente introdotte. Il classifier è un metodo non parametrico basato sul depth vs. depth plot. I maggiori vantaggi riguardano la flessibilità e l'indipendenza da specifiche ipotesi distribuzionali. La nuove funzioni di depth, essendo particolarmente vantaggiose in termini computazionali, permettono di utilizzare questa metodologia anche nel caso di insiemi di dati direzionali a dimensioni elevate.*

Key words: Angular data depth, Chord depth, Classification, Cosine depth.

1 Introduction

Directional data arise when observations are measured as directions or angles, and observations are represented as points on the circumference of the unit circle, or on the surface of the unit hyper-sphere. Simple examples are directions on a compass $(0^\circ - 360^\circ)$, and times of the day $(0 - 24$ hours). Higher dimensional directional data arise in text mining and gene expression analysis.

Giuseppe Pandolfo
University of Naples Federico II, Napoli, e-mail: giuseppe.pandolfo@unina.it

Giovanni C. Porzio
University of Cassino and Southern Lazio, Cassino e-mail: porzio@unicas.it

They play thus an important role in many fields such as biology, meteorology, neurology and physics. Within the literature, directional data are presented and discussed in books by Mardia (2000) and Batschelet (1981). They provide a wide survey about specific features and problems we have to face when dealing with them.

Specific statistical methods are needed to analyse directional data. This is due to the periodicity and boundness of the sample space. To make it clearer, one may consider two angles of 10° and 350° which are 20° far away from each other. If treated as linear data, their arithmetic mean would be equal to 180°, while their correct directional mean is 0°. Hence, to avoid misleading results, appropriate techniques are necessary to perform directional data analysis, and this applies to classification in directional spaces as well.

In this work, the focus is on supervised classification (also called discriminant analysis), where the aim is to assign observations to classes (or groups), given a previous knowledge on their structures obtained through some preliminary observed data.

Specifically, we consider the directional non-parametric DD-classifier introduced in Pandolfo (2014, 2015), and further investigated in Pandolfo (2017).

The classifier evaluates the depth of a new observation with respect to some previously collected samples (a depth wrt each of the groups), and then uses its value as a basis for a classification rule. While Pandolfo (2014, 2015, 2017) investigated such a tool adopting different notions of depth functions already available within the literature, this work focuses on the advantages that can be obtained if some new depth functions are adopted. Pandolfo et al. (2017) recently introduced a new class of functions based on directional distances. It seems one of the main advantages of these functions is their computational feasibility. For this reason, they might be particularly suitable to deal with non-parametric classification problem in high-dimensional spaces. This essentially motivates this work, whose aim is to present a performance analysis of the directional DD-classifier under these two new depth functions.

The paper is organized as follows: Section 2 briefly recalls the definition of these new directional depth functions and presents the related classifier; Section 3 offers some concluding remarks.

2 Non-parametric classification using data depth

The literature on nonparametric discriminant analysis for directional data is not particularly extensive. Liu and Singh (1992) discussed methods for ordering directional data and suggested a possible solution for discriminant analysis. That is, an observation can be classified according to its center-outward ordering with respect to two given competing distributions. Later, following the linear approach by Stoller (1954) a circular discriminant analytical method was proposed by Ackermann (1997). This method, which looks for the pair of splitting angles which maximizes

the probability of correct classification. Unfortunately, this solution requires solving an NP-complex problem, and hence it is quite computationally infeasible.

Several tools are available for classification problems in standard multivariate analysis. Assuming normality, linear discriminant analysis is probably the most used. Many other methods, both parametric and non-parametric, are available as well. Non-parametric classifiers have the important advantage to be more flexible, given that they do not require any distribution assumptions. Amongst them, this work focuses on classifiers based on data depth. Considering their full non-parametric nature, they can be used in many different contests.

The depth of a point $x \in \mathbb{R}^q$ (for dimension $q \geq 1$) is a function that measures the "centrality" or "deepness" of it with respect to a multivariate distribution F or a multivariate data cloud, and it is denoted by $D(x,F)$.

Some notions of angular data depth are available within the literature: the *angular simplicial depth*, the *angular Tukey's depth* and the *arc distance depth*. All these were given in Liu and Singh (1992). More recently Pandolfo et al. (2017) introduced a class of depth functions for directional data based on angular distances. The class is defined below.

Let $\mathscr{S}^{q-1}$ be the unit sphere in a $q-1$ dimensional space. A particular member of the class will be obtained by fixing a particular (bounded) distance $d(\cdot,\cdot)$ on $\mathscr{S}^{q-1}$. For such a distance, $d^{\sup} := \sup\{d(\theta,\psi) : \theta,\psi \in \mathscr{S}^{q-1}\}$ will denote the upper bound of the distance between any two points on $\mathscr{S}^{q-1}$. We have the following definition (Pandolfo et al., 2017).

Definition 1 (Directional distance-based depths) *Let $d(\cdot,\cdot)$ be a bounded distance on $\mathscr{S}^{q-1}$ and H be a distribution on $\mathscr{S}^{q-1}$. Then the* directional d-depth of $\theta(\in \mathscr{S}^{q-1})$ with respect to H *is*

$$D_d(\theta,H) := d^{\sup} - E_H[d(\theta,W)], \tag{1}$$

where E_H is the expectation under the assumption that W has distribution H.

Accordingly, two new easy-to-compute directional depth functions can be obtained if the distance in (1) is chosen to be the chord or the cosine distance, respectively. While the main properties of these functions have been investigated in Pandolfo et al. (2017), their use within a DD-classifier is investigated here.

2.1 Depth-based classifiers and the DD-plot

After the first suggestion in Liu and Singh (1992), the use of data depth to perform supervised classification has been suggested and investigated by many authors. Two main approaches have been adopted in the literature: the *maximum depth classifier* and the *Depth vs Depth* (*DD*) classifier. The latter (that is a refinement of the former) is based on the *DD*-plot (Depth vs Depth plot), introduced by Liu et al. (1999).

The *DD*-plot is a two-dimensional scatterplot where each data point is represented with coordinates given by its depth evaluated with respect to two distributions. A classification rule $r(\cdot)$ is then directly applied in this latter. The same procedure can be applied to any kind of data, providing that a corresponding depth function exists. For istance, DD-plot for functional data have been developed.

For directional data, the following questions arise. First, it is of interest to investigate if some depths perform better than others. Second, the classification rule to be adopted within the plot may also affect performances. A proper simulation study has been thus developed in order to suggest under which conditions one depth function should be preferred over the other.

3 Concluding remarks

The depth vs. depth classification method extended to directional data has been investigated here when some new distance based depth functions are adopted. The idea of depth provides a criterion to order a directional sample from center-outward, providing a new way to classify directional objects. The performance in terms of average misclassification rate of the classifiers is evaluated by means of a simulation study. Furthermore, their use is illustrated through a real data example. First results are promising, calling for further investigation on the performance under different directional settings.

Riferimenti bibliografici

1. Ackermann, H.: A note on circular nonparametrical classification. Biometrical J **5**, 577–587 (1997)
2. Batschelet, E.: Circular statistics in biology. Academic Press, London, (1981)
3. Liu, R.Y., Singh, K.: Ordering directional data. Concepts of data depth on circles and spheres. Ann Stat **20**, 1468–1484 (1992)
4. Liu, R.Y., Singh, K.: Multivariate analysis by data depth: descriptive statistics, graphics and inference. Ann Stat **27**, 783–1117 (1999)
5. Mardia, K.V., Jupp, E.P.: Statistics of directional data. Academic Press, London (1972)
6. Pandolfo, G.: On depth functions for directional data. Ph.D. Thesis. Department of Economics and Law, University of Cassino and Southern Lazio (2014)
7. Pandolfo, G.: A depth-based classifier for circular data. Mola F., Conversano C. eds., CLADAG 2015 BOOK of Abstracts, CUEC Editrice, 324–327 (2015)
8. Pandolfo, G., Paindaveine, D., Porzio, G.C.: Distance-based depths for directional data. Working Papers ECARES 2017 – 35 (2017) Available at https://ideas.repec.org/p/eca/wpaper/2013-258549.html
9. Stoller, D.S.: Univariate two-population distribution-free discrimination. J Am Statist Assoc **49**, 770–777 (1954)