CLADAG 2021

BOOK OF ABSTRACTS AND SHORT PAPERS

13th Scientific Meeting of the Classification and Data Analysis Group
Firenze, September 9-11, 2021

edited by
Giovanni C. Porzio
Carla Rampichini
Chiara Bocci

FIRENZE
UNIVERSITY
PRESS

PROCEEDINGS E REPORT

ISSN 2704-601X (PRINT) - ISSN 2704-5846 (ONLINE)

– 128 –

ANGULAR HALFSPACE DEPTH: CLASSIFICATION USING SPHERICAL BAGDISTANCES*

Houyem Demni[1], Davide Buttarazzi[1], Stanislav Nagy[2],
and Giovanni C Porzio[1]

[1] Department of Economics and Law, University of Cassino and Southern Lazio
(e-mail: houyem66@gmail.com, davidebuttarazzi@outlook.com,
porzio@unicas.it)

[2] Department of Probability and Mathematical Statistics, Charles University
(e-mail: nagy@karlin.mff.cuni.cz)

ABSTRACT: Directional data lies on the surface of the unit sphere. Exploiting new results on the computation and the properties of the angular halfspace depth, we introduce the spherical version of the bagdistance, applicable to directional data. A bagdistance-based classification method for directional data is considered. The proposed method will be compared with other directional classifiers by means of a simulation study.

KEYWORDS: angular depth, bagdistance, directional data, supervised learning.

1 Introduction

Depth functions are nonparametric tools that assess how "centrally located", or "inner" is a point with respect to (w.r.t.) a given probability distribution. They have been successfully adopted in supervised classification analysis. However, many depths suffer when evaluating points that lie in the tails of the distribution. This is because the depth functions are typically not robust at their lowest values, and also because they can easily assign constant zero depth to many points when evaluated w.r.t. datasets (the so-called outsider issue). An example of an important depth sharing all these shortcomings is the standard *halfspace depth* defined in Euclidean spaces $\Re^q$, $q \geq 1$.

Contrary to the depths, distance functions are much more powerful when dealing with points at the extremes of the distribution. Nevertheless, they generally suffer from robustness issues as well (unless some robustified versions

*The work of H. Demni and G.C. Porzio has been partially funded by the BiBiNet project (grant H35F21000430002) within the POR-Lazio FESR 2014-2020. The work of S. Nagy was supported by the grant 19-16097Y of the Czech Science Foundation, and by the PRIMUS/17/SCI/3 project of Charles University.

are adopted), and for a fruitful use of the distances in classification, certain assumptions on the data distribution typically need to be imposed (e.g., ellipticity of the underlying distribution in the case of the Mahalanobis distance).

For these reasons, and to introduce a supervised classification rule for Euclidean data, Hubert *et al.*, 2017 proposed to combine the information from these two approaches to obtain the so-called *bagdistance*, a function which joins the depth and the distance to obtain a measure of how close/inner is a point w.r.t. a given distribution. Bagdistances are robust, nonparametric, and able to manage information in the tails of the distribution.

In this work, we introduce the bagdistance for directional data. To do so, we use the angular halfspace depth, being the directional analogue of the standard halfspace depth from $\Re^q$. We also evaluate the performance of the bagdistance within the setting of supervised classification for directional data.

Our short paper is organized as follows. Section 2 provides some background on the bagdistance in the Euclidean case, while in Section 3, the spherical bagdistance and a directional classifier based on it are introduced.

2 The bagdistance for Euclidean data

Let Y be a random variable in $\Re^q$ with distribution P_Y, and let θ be its halfspace median (the point that maximizes the halfspace depth w.r.t. P_Y, or the barycentre of the set of such points if not a singleton). Denote by $B(Y) \subset \Re^q$ the smallest halfspace depth central region of P_Y (i.e., an upper level set of the halfspace depth of P_Y) that contains at least 50 % of the P_Y-probability mass. The bagdistance of x to Y is given by the ratio of the Euclidean distances of x to θ, and $c(x)$ to θ:

$$BD(x, P_Y) := \begin{cases} 0 & \text{if } c(x) = \theta, \\ \|x - \theta\| \,/\, \|c(x) - \theta\| & \text{otherwise,} \end{cases}$$

where $c(x)$ is the intersection of the boundary of the bag $B(Y)$ and the ray from the halfspace median θ passing through x.

3 The spherical bagdistance and a classification rule

Directional data can be viewed as realizations of a random variable X whose support is the unit hyper-sphere $S^{(q-1)} := \{x \in \Re^q \colon \|x\| = 1\}$. For directional data, the spherical bagdistance can be introduced in complete analogy with the bagdistance for Euclidean data.

We first define the directional variant of the halfspace depth. Let X be a directional random variable with distribution P_X. The *angular halfspace depth* ahD of a point $x \in S^{(q-1)}$ w.r.t. P_X can be defined considering the collection $\mathscr{H}_0$ of closed halfspaces in $\Re^q$ whose boundary contains the origin:

$$ahD(x,P_X) := \inf\{P_X(H) \colon H \in \mathscr{H}_0,\ \ x \in H\} \in [0,1].$$

Denote by $aB(X) \subset S^{(q-1)}$ the *angular bag* of X, defined as the smallest angular depth central region containing at least 50 % of the P_X-probability mass. Such a region always exists; its properties are detailed in the contribution of *P. Laketa* and *S. Nagy* in the present book of short papers. The *spherical bagdistance* from $x \in S^{(q-1)}$ to X is defined as the ratio of the arc distance between x and the angular halfspace median $\tilde{\theta}$ (a maximizer of the angular halfspace depth of X), and the arc distance between $c_{aB}(x)$ and $\tilde{\theta}$. Here, $c_{aB}(x)$ is the intersection between the boundary of the angular bag $aB(X)$ and the geodesic from $\tilde{\theta}$ to x. Altogether, we define

$$SBD(x,P_X) := \begin{cases} 0 & \text{if } c_{aB}(x) = \tilde{\theta}, \\ \arccos(x^\top \tilde{\theta}) / \arccos(c_{aB}(x)^\top \tilde{\theta}) & \text{otherwise.} \end{cases}$$

Similarly as the usual bagdistance in $\Re^q$, the spherical bagdistance can be exploited for supervised classification of directional objects. Formally, considering K directional distributions on $S^{(q-1)}$, a directional classifier is defined as the function $class : S^{(q-1)} \to \{1,\ldots,K\}$. Given a training set composed of K empirical distributions $\hat{P}_{X_i}$, $i = 1,...,K$, the directional bagdistance classifier is then defined as the rule $class_{bag}$ such that:

$$class_{bag}(x) := u(SBD(x;\hat{P}_{X_1}),...,SBD(x;\hat{P}_{X_i}),...,SBD(x;\hat{P}_{X_K})),$$

where $u : \Re^K \to \{1,...,i,...,K\}$ is some discriminating function. That is, the classifier is a rule defined on a Euclidean space given by the bagdistances of the training set values w.r.t the directional distributions defined on a Riemannian manifold. For the choice of the discriminating function, we refer to the literature available for depth based classifiers, which includes the linear (LDA), quadratic (QDA) and k-NN classifiers (see e.g., Demni *et al.*, 2021).

In line with such a strategy, a simulation study with data generated according to a Kent distribution for each group has been performed. First results are promising: the spherical bagdistance classifier reaches the same level of correct classification as achieved by the empirical Bayes, at least under some circumstances. To exemplify, boxplots of the misclassification rates of the proposed classifier and of the empirical Bayes classifier under Kent are reported in Figure 1. The two Kent distributions have equal locations and ovalness, and different concentrations (the simulation setting described in Setup 2 in Demni & Porzio, 2021 has been adopted). The training set size is 400 (200 from each group), while the size of the testing set is 200; the number of replications is 100. Misclassification errors are essentially equivalent, with some preference to be given to the LDA and QDA solution. Performances under other simulation settings and comparison with other directional classifiers are under investigation.

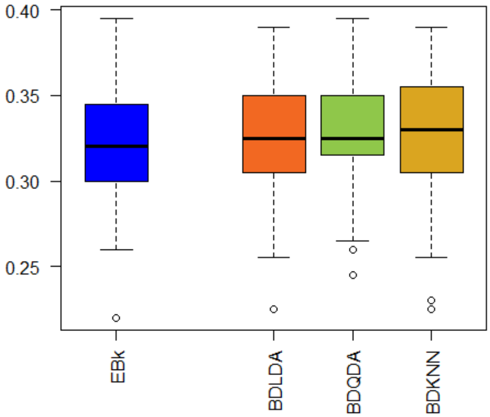

Figure 1: Misclassification rates of the empirical Bayes under Kent (EBk), and the spherical Bagdistance classifier (BD) when associated with the LDA, QDA, and k-NN classification rule. Data generated according to Kent distributions.

References

DEMNI, HOUYEM, & PORZIO, GIOVANNI C. 2021. Directional DD-classifiers under non-rotational symmetry. *IEEE Xplore, submitted.*

DEMNI, HOUYEM, MESSAOUD, AMOR, & PORZIO, GIOVANNI C. 2021. Distance-based directional depth classifiers: a robustness study. *Communications in Statistics – Simulation and Computation, in press.*

HUBERT, MIA, ROUSSEEUW, PETER, & SEGAERT, PIETER. 2017. Multivariate and functional classification using depth and distance. *Advances in Data Analysis and Classification*, **11**(3), 445–466.