ON THE ROBUSTNESS OF THE COSINE DISTRIBUTION DEPTH CLASSIFIER

Houyem Demni[1] [2], Amor Messaoud[2] and Giovanni C. Porzio[1]

[1] Department of Economics and Law, University of Cassino and Southern Lazio, (e-mail: houyem66gmail.com, porzio@unicas.it)

[2] University of Tunis, Tunisia (e-mail: amor.messaoud@gmail.com)

ABSTRACT: Investigating how classifiers perform under some data contaminations is an important issue in robustness studies. While some research is available on the robustness of classifiers, a little is known about directional classifiers. This work thus investigates the robustness of the cosine depth distribution classifier, a classification technique recently introduced for directional data. This latter is a non-parametric method and it is based on the distribution function of the cosine depth.

KEYWORDS: directional data, supervised classification, unit vectors.

1 Introduction

Directional data occur when observations are recorded as directions. They can be described as unit vectors on the surface of the $(d-1)$ dimensional hypersphere $S^{(d-1)} := \{x : x^T x = 1\}$. This kind of data can be found in many scientific areas such as medicine, astronomy, biology and geology, to cite a few. Applications include cases with $d = 2$ (circular data), $d = 3$ (Mardia & Jupp, 2000) and in higher dimensions (Buchta *et al.*, 2012).

In this work, we consider the problem of classifying directional data according to some supervised classification technique, and in particular on a technique which relies on data depth.

Data depth functions provide basis for nonparametric inference given that they aim at ordering data in a d-dimensional space according to some centrality measures. The particular properties of directional data and the complexity of the sample space imply the need of specific methods to analyze them.

Within the framework of classification, the use of data depth has been extensively investigated and successfully applied. The max depth classifier has been firstly developed (Ghosh & Chaudhuri, 2005, after Liu *et al.*, 1990). Later, the idea has been extended and the DD-classifier has been introduced (Li *et al.*, 2012).

A recent interest arises on the use of depth based classifier for directional data: the use of the directional max-depth classifier based on some new depth functions has been investigated (Pandolfo *et al.*, 2018a), and the DD-plot classifier for circular data has been discussed (Pandolfo *et al.*, 2018b).

Even more recently, a depth based distribution classifier was introduced in the framework of supervised classification to assign points lying on the surface of hyper-spheres (spherical data) to groups (Demni *et al.*, 2019). It was based on the cosine depth, and called the cosine distribution depth classifier. Simulation results showed that the cosine depth distribution classifier outperforms the max depth classifier in term of average misclassification rate also in many settings.

In supervised classification, the presence of anomalous observations in the training set can greatly reduce the effectiveness of the classification method adopted (Vencalek & Pokotylo, 2018). For this reason, it is always of interest to investigate the robustness of these kind of techniques. Several works dealt with robust based classifiers (see Dutta & Ghosh, 2012; Li *et al.*, 2012). Pandolfo investigated some robustness aspects of the DD-classifier for directional distributions (Pandolfo, 2017).

Here, the focus will be on the cosine depth distribution classifier. By means of a simulation study, it will be investigated to what extent this classifier is able to deal with contaminated training sets. The rest of the work is organized as follows. Section 2 introduces the directional cosine depth distribution classifier, while in Section 3 the simulation scheme that will be used to assess its robustness is provided.

2 The cosine depth distribution classifier

Directions in d-dimensional spaces can be represented as unit vectors x on the sphere $S^{(d-1)} := \{x : x^T x = 1\}$ with unit radius and center at the origin. A distribution H with support $\Omega \subseteq S^{(d-1)}$ is called a directional distribution. By definition, the cosine depth of a point $x \in S^{(d-1)}$ with respect to H is given by:

$$D_{cos}(x,H) := 2 - E_H[(1 - x'W)],$$

where $E[.]$ is the expected value, and W is a random variable from H.

The cumulative distribution function of the cosine depth function $F_D^H(x)$ is given by:

$$F_D^H(x) := P(D_{cos}(X,H) \leq D_{cos}(x,H))$$

Suppose now observations come from either the distribution (group) H_1 or H_2. Then, the directional depth distribution classification rule (Demni *et al.*, 2019) is given by:

$$\begin{cases} F_D^{\hat{H}_1}(x) > F_D^{\hat{H}_2}(x) \Longrightarrow \text{assign } \ x \ \text{ to } \ \text{population 1} \\ F_D^{\hat{H}_1}(x) < F_D^{\hat{H}_2}(x) \Longrightarrow \text{assign } \ x \ \text{ to } \ \text{population 2}, \end{cases}$$

where $\hat{H}$ refers to the empirical distribution.

If $F_D^{\hat{H}_1}(x) = F_D^{\hat{H}_2}(x)$, the classification rule will randomly assign the observation to one of the two groups with equal probability.

3 A simulation scheme to study the robustness of the cosine depth distribution classifier

To investigate the robustness properties of the cosine depth distribution classifier for directional data, the following simulation setting will be used.

Let H_1 and H_2 be two von Mises-Fisher distributions (vMF). That is, their corresponding density functions $h()$ are given by

$$h(x;\mu,c) := \left(\frac{c}{n}\right)^{d/2-1} \frac{1}{\Gamma^{(d/2)} I_{d/2-1}(c)} \exp\{c\mu^T x\},$$

where $c \geq 0$, $||\mu|| = 1$, and I_v denotes the modified Bessel function of the first kind and order v. The parameters μ and c are the mean direction and the concentration parameter, respectively.

The training set size will be 1000 (500 from each group), while the size of the testing set will be 500. The number of replications will be set equal to 150 times. For the concentration parameters c_1 and c_2 of H_1 and H_2, we consider two cases: equal concentration ($c_1 = c_2 = 5$), and different concentration ($c_1 = 2$ and $c_2 = 6$).

The location parameters for μ_1 and μ_2 are set to be equal to $(0,0,1)$, $(1,0,0)$ in dimension $d = 3$, respectively. The training observations from H_1 are contaminated with observations generated from VmF with location parameter equal to $\mu = (0,0,-1)$ and concentration parameter $c = 8$.

The location parameters are set to be equal to $\mu_1 = (0,0,0,0,0,0,0,0,0,1)$, and $\mu_2 = (1,0,0,0,0,0,0,0,0,0)$ in dimension $d = 10$.
Contaminated observations are generated from VmF with location parameter $\mu = (0,0,0,0,0,0,0,0,0,-1)$ and concentration parameter $c = 8$.

Finally, the contamination levels will be set equal to 0%, 10%, 20%.

References

BUCHTA, C, KOBER, M, FEINERER, I, & HORNIK, K. 2012. Spherical k-means clustering. *Journal of Statistical Software*, **50**(10), 1-22.

DEMNI, H, MESSAOUD, A, & PORZIO, G.C. 2019. The Cosine depth distribution classifier for directional data. *In:* ICKSTADT K, TRAUTMANN H, SZEPANNEK G LÜBKE K, & N, BAUER (eds), *Applications in Statistical Computing, Chapter 4*. Springer Nature Switzerland AG, in press. https://doi.org/10.1007/978-3-030-25147-5_4.

DUTTA, S, & GHOSH, A. K. 2012. On robust classification using projection depth. *Annals of the Institute of Statistical Mathematics*, **64**(3), 657-676.

GHOSH, A. K, & CHAUDHURI, P. 2005. On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, **32**(2), 327-350.

LI, J, CUESTA-ALBERTOS, J. A, & LIU, R. Y. 2012. DD-classifier: Nonparametric classification procedure based on DD-plot. *Journal of the American Statistical Association*, **107**(498), 737-753.

LIU, R. Y, *et al.* 1990. On a notion of data depth based on random simplices. *The Annals of Statistics*, **18**(1), 405-414.

MARDIA, K.V, & JUPP, P.E. 2000. Directional Statistics, John Wiley and Sons, London.

PANDOLFO, G. 2017. Robustness aspects of DD-classifiers for directional data. *In:* GRESELIN, F, MOLA F, & ZENGA, M (eds), *CLADAG 2017 Book of Short Papers*. Universitas Studiorum S.r.l. Casa Editrice, Mantova.

PANDOLFO, G, PAINDAVEINE, D, & PORZIO, G. C. 2018a. Distance-based depths for directional data. *Canadian Journal of Statistics*, **46**(4), 593-609.

PANDOLFO, G, D'AMBROSIO, A, & PORZIO, G.C. 2018b. A note on depth-based classification of circular data. *Electronic Journal of Applied Statistical Analysis*, **11**(2), 447-462.

VENCALEK, O., & POKOTYLO, O. 2018. Depth-weighted Bayes classification. *Computational Statistics and Data Analysis*, **123**, 1-12.