Directional density-based clustering

Paula Saavedra-Nieves[1], *Martín Fernández-Pérez*[2]

[1]paula.saavedra@usc.es, CITMAga, Universidade de Santiago de Compostela
[2]martin.fernandez.perez0@rai.usc.es, Universidade de Santiago de Compostela

Clustering for directional data has achieved a considerable relevance over the last decades, specially amongst the machine learning community. The most popular approaches are spherical $k-$means with cosine similarity (see [1]) and the use of finite mixture models with von Mises-Fisher components (see [2]). A nonparametric alternative to $k-$means is modal clustering. This approach that associates the notions of cluster and mode, does not require to specify the number of groups in advance (see [3]). The connection between clusters and modes is also present in density-based clustering methodology introduced in [4]. Under this perspective, clusters are identified with the connected components of density level sets. This topic has received remarkable attention in the literature but only for densities supported on an Euclidean space. Concretely, the computational problem of determining the connected components of level sets in high dimensional spaces was addressed in [5] and [6]. As a natural consequence, the empirical mode function and the cluster tree (under the generated hierarchical structure) were defined and an unsupervised classification method was proposed. The main goal of this work is to generalize density-based clustering techniques for directional data. Specifically, we present a novel algorithm for determining the connected components of level sets of densities supported on a unit hypersphere. An extensive simulation study shows the performance of the resulting classification methodology.

Keywords: Density level sets, directional clustering, unsupervised classification.

Acknowledgements. P. Saavedra-Nieves acknowledges the financial support of Ministerio de Ciencia e Innovación of the Spanish government under grants PID2020-118101GBI00 and PID2020-116587GBI00 and ERDF (Grupos de Referencia Competitiva ED431C 2021/24).

1. Bibliography

[1] Dhillon, I. S., and Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. Machine learning, 42(1), 143-175.

[2] Banerjee, A., Dhillon, I. S., Ghosh, J., Sra, S., and Ridgeway, G. (2005). Clustering on the unit hypersphere using von mises-fisher distributions. Journal of Machine Learning Research, 6(9).

[3] Oba, S., Kato, K., and Ishii, S. (2005). Multi-scale clustering for gene expression profiling data. In Fifth ieee symposium on bioinformatics and bioengineering (pp. 210-217).

[4] Hartigan, J. A. (1975). Clustering algorithms. John Wiley & Sons, Inc.

[5] Azzalini, A., and Torelli, N. (2007). Clustering via nonparametric density estimation. Statistics and Computing, 17(1), 71-80.

[6] Menardi, G., and Azzalini, A. (2014). An advancement in clustering via nonparametric density estimation. Statistics and Computing, 24(5), 753-767.