High-dimensional Unsupervised and/or Supervised problems: a distance-based depth prototypes fuzzy approach

Itziar Irigoien[1], *Susana Ferreiro*[2], *Basilio Sierra*[3], *Concepción Arenas*[4]

[1]itziar.irigoien@ehu.es, Department of Computation and Artificial Intelligence, University of the Basque Country (UPV/EHU)

[2]susana.ferreiro@tekniker.es, Intelligent Information Systems Unit, Tekniker

[3]b.sierra@ehu.eus, Department of Computation and Artificial Intelligence, University of the Basque Country (UPV/EHU)

[4]carenas@ub.edu, Statistics Section of the Department of Genetics, Microbiology and Statistics, University of Barcelona (UB)

Supervised and unsupervised classifications are crucial in many areas such as biomedicine, or industry, among others, where different types of data sets and high-dimensional data (number of features p larger than the number of observations n, $p >> n$) are common. A very important issue is that some units are more typical of the group they belong to than others. Following the ideas of Hastari et al. (2020) We propose a new fuzzy supervised classification approach based on the construction of prototypes from an objective function that incorporates the information from the class labels, as well as a distance-based depth function to fuzzify the partition. As it is a fuzzy methodology, the objective function contains a term related to the entropy of the memberships. Obtaining the prototypes, label prototypes, and weighted memberships relative to each class is carried out by an iterative scheme. The method has hyperparameters that need to be tuned and a grid search is used. Then, when the approach is supervised, the selection of the hyperparameter is made according to an adequate metric (accuracy rate, for instance) reached on a k-fold cross-validation setting. When the approach is non-supervised, there is a lack of an external validation variable and, as an internal validation approach, a permutation approach related to the Gap statistics leads to good results. The procedure is distance-based, so it can be used in data sets of a very varied nature, particularly without restrictions with high-dimensional data and with data sets where the Euclidean distance is not suitable, but other distances are. Notably, it may be a very good choice for functional data. Furthermore, the method selects the prototypes among the deepest units preventing the prototype from not belonging to the sample, as can happen in the case of centroids. Its performance on synthetic data sets along with real data showed good rates of correct classification and it is competitive with other methods. The proposed method provides an interesting alternative to other fuzzy clustering purposes.

Keywords: Fuzzy classification, Prototypes, Depth function.

P. Ashtari, F. N. Haredasht, H. Beigy (2020). Supervised fuzzy partitioning, Pattern Recognition, 97, Article 107013.