12 April, 9:40 - 10:00, Room A2

Looking for atypical groups of distributions in the context of genomic data

Ana Helena Tavares[1], **Vera Afreixo**[1], **Paula Brito**[2]
[1] CIDMA, University of Aveiro, ahtavares@ua.pt
[2] FEP & LIAAD-INESC TEC, University of Porto

This work addresses the problem of detecting groups of observations (distributions) and flagging those that differ abnormally from the majority of the groups, termed as atypical groups. The proposed method combines a hierarchical classification technique, to identify groups of similar distributions, with a functional outlier detection method, to identify those groups that contain outliers. Groups with outlying observations are forwarded for sub clustering. Once the final partition is obtained, each cluster is represented by a class prototype, whose outlyingness is evaluated according to a functional approach. Clusters with atypical class labels are flagged as atypical groups. The method is applied for the detection of groups of atypical genomic words, based on their distances distributions.

Keywords: clustering, outlying distribution, atypical group

The identification of outliers can lead to the discovery of truly unexpected knowledge in several areas, e.g. electronic commerce, video surveillance and health care. A widely reported definition of outlier observation is the one proposed by Grubbs in 1969 and quoted in Barnett and Lewis [1]: *An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs.* This general definition of outlier is vague and becomes meaningful only under a given context or application.
In this work, we introduce the concept of *atypical group* and propose a procedure for its identification. We focus our work in the detection of such groups in data that can be represented by a distribution, in particular, in distances distributions between genomic words [3]. We are convinced that large heterogeneous datasets, where distinct patterns coexist, may exhibit one or more atypical groups, meaning groups of observations whose 'mean' pattern stands out from the majority of the 'mean' patterns.
If large heterogeneous datasets where distinct patterns coexist can validly be clustered, then the class prototypes may provide a meaningful description of similarities and differences in the data. By representing each group by a prototype, the inicial dataset is reduced to a given number of representative distributions. By applying a functional outlier procedure over the set of prototypes, it is possible to identify those groups whose prototype is flagged as outlier. Such group is then termed as an atypical group.

To identify distinct patterns in a set of distributions we have combined a hierarchical clustering method with a functional outlying detection method. The first creates a hierarchy of clusters according to a dissimilarity measure, while the second flags observations with atypical curves in the set of group members. In this second step, a measure of outlyingness is used that privileges the shape of the distributions and not only the magnitude of their values [2]. Groups in which atypical observations are identified, are forwarded for (sub)clustering, and the procedure is repeated until no outliers are identified. Once the final partition is obtained, each cluster is represented by a class prototype and its outlyingness is evaluated according to the same functional approach [2]. The key idea of our proposal is to use a functional outlyingness criterion as indicator of the cluster homogeneity and then use it again to identify the atypical class prototypes.
We are particularly interested in developing a method that recovers groups of genomic words with similar distribution patterns along the genome sequence and, in particular, those very small groups with a distribution pattern which is markedly different from the majority. We analyze the dataset of the inter-word distance distributions of words of length $k = 5$, which contains 1024 distributions. To form the clusters an agglomerative hierarchical method is applied, considering the Mallows L^1 distance and average linkage. To decide on the number of clusters to retain in each step of the procedure we resort to two validity indexes, the Calinski-Harabasz index and the Silhouette score.
The application of this new procedure allowed identifying three groups of distributions with homogeneous patterns and very different from the others. These groups are of small dimension, and the words belonging to the identified groups are rich in CG dinucleotides. The groups of genomic words identified may have a potential biological interest, since atypical distribution patterns may be related to words that have biological meaning.

Acknowledgements This work was partially supported by Fundação para a Ciência e a Tecnologia, within projects UID/MAT/04106/2019 (CIDMA) and UID/EEA/50014/2013 (INESC TEC), and by the European Regional Development Fund (ERDF) through the Operational Programme for Competitiveness and Internationalization - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961.

References

[1] Vic Barnett and Tomy Lewis. *Outliers in Statistical Data*. Wiley, 1994.

[2] Peter J Rousseeuw, Jakob Raymaekers, and Mia Hubert. A measure of directional outlyingness with applications to image data and video. *Journal of Computational and Graphical Statistics*, pages 1–15, 2018.

[3] Ana Helena Tavares, Vera Afreixo, João MOS Rodrigues, and Carlos AC Bastos. The symmetry of oligonucleotide distance distributions in the human genome. In *Proc. ICPRAM (2)*, pages 256–263, 2015.