Bagging algorithm for grouping improvement in the context of three-dimensional shapes

Inácio Nascimento[1], Raydonal Martinez[2], Getúlio Amaral[1]
[1] Universidade Federal de Pernambuco, Recife, Pernambuco, Brazil
[2] Universidade Federal da Bahia, Salvador, Bahia, Brazil

Abbreviated abstract: Methods of Statistical Shapes Analysis are used to deal with statistical information of geometric shapes of objects. In general, Cluster Analysis techniques allow classifying the information of objects into groups. In order to achieve a better grouping into classes of geometric objects, in this work, the K-means, Hill Climbing and CLARANS algorithms, aimed at forming groups of three-dimensional shapes, were adapted under an ensemble approach using the Bagging procedure.

Related publications:

– Assis *et al,* International Journal of Business Intelligence and Data Mining 18 (1), 30-48 (2021)
– Vinué *et al*, Advances in Data Analysis and Classification 10 (1), 103-132 (2016)

UNIVERSIDADE FEDERAL DE PERNAMBUCO

inacio.nascimento@ufpe.br, raydonal@de.ufpe.br,
gjaa@de.ufpe.br

4th Conference on Statistics and Data Science
Salvador, Brazil (online)
December 1-3, 2022

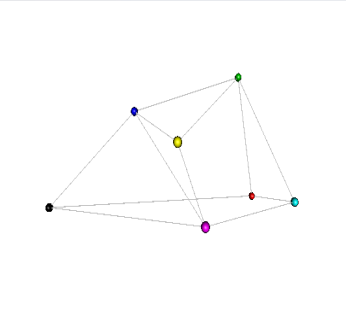
Motivation and Objectives

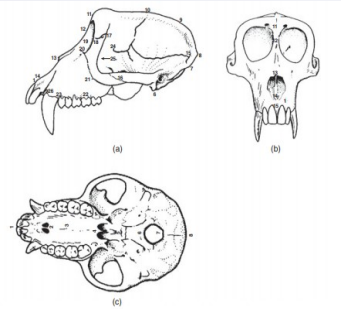
- In several situations it is necessary to classify data sets of shapes into clusters. Considering two-dimensional shapes (2D), there is the work by Amaral *et al.* (2010), who adapt the version of the K-means for an application of Oceanography. And the work by Assis *et al.* (2021) introduces the Bagging method executed in conjunction with the K-means algorithm.

- However, in the scenario of three dimensional (3D) shapes, there are not many works involving adaptations of clustering methods. One of the most significant works in this regard was developed by Vinué *et al.* (2016), and presents the adaptation of the K-means for clustering three-dimensional shapes.

- Motivated by all these works, this work aims at making use of the Bagging method, called BagClust1, proposed by Dudoit and Fridlyand (2003), to improve the quality of the grouping results of three dimensional shapes obtained by the K-means, Hill Climbing, and CLARANS, adapted to the context of 3D shapes.

3D representation of a monkey skull of the species Macaca fascicularis





UNIVERSIDADE FEDERAL DE PERNAMBUCO

inacio.nascimento@ufpe.br, raydonal@de.ufpe.br, gjaa@de.ufpe.br

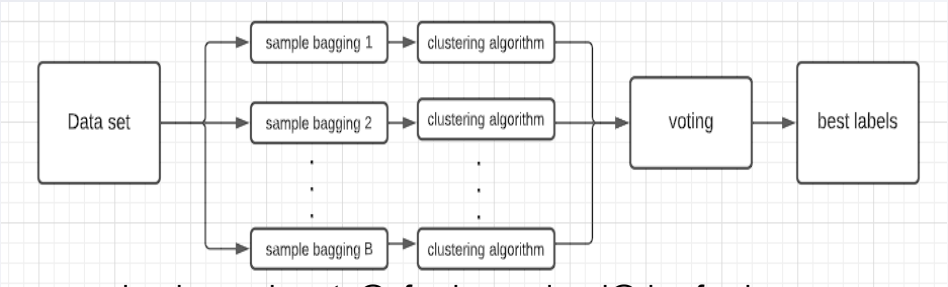
4th Conference on Statistics and Data Science
Salvador, Brazil (online)
December 1-3, 2022

Methodology

- In order to compare the performance of the K-means, Hill Climbing and CLARANS methods adapted to the context of three-dimensional shapes with their versions combined with the bagging method, experiments were performed on simulated data sets, besides the application on a real data set. The Riemannian distance was used on both in cases as an appropriate dissimilarity measure for the shape space.

- The analyzed algorithms were implemented in the R programming language (R Core Team, 2022) with the help of functions belonging to packages aimed at analyzing shapes. To compare the efficiency of the algorithms used in this article, we used the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985).

- The Adjusted Rand index is calculated for the best result obtained. There were B = 20 bootstrap replicates generated that were grouped together to form the labels that went through the voting process to generate the best partition. The objective is to show the efficiency of using the bagging method combined with clustering algorithms in comparison to their versions without it.



UNIVERSIDADE FEDERAL DE PERNAMBUCO

inacio.nascimento@ufpe.br, raydonal@de.ufpe.br, gjaa@de.ufpe.br

4th Conference on Statistics and Data Science
Salvador, Brazil (online)
December 1-3, 2022

Results and Conclusions

- Experiments were simulated in the space of landmarks based on the data simulated by Vinué *et al.* (2016). Two predefined geometric figures, a cube and a parallelepiped, were considered to represent the average configurations, with a number of anatomical landmarks of equal 8. Each simulated data set has N objects divided into two groups n1 and n2 for the sample size of N = 100, n1 = 60 and n2 = 40, as well as different dispersion scenarios: isotropy and anisotropy.

- We used as real data the Steroid data set. Each molecule has a binding affinity class: high, intermediate, or low. The data set consists of 31 steroid molecules. In this way, we can define each molecule in 3 distinct clusters according to the degree of attachment.

- Experimental results with different data sets showed the effectiveness of the approach using the bagging method together with the K-means, Hill Climbing and CLARANS methods, specifically in cases with high dispersion between landmarks.

Results for the Steroid Molecules data set of three-dimensional shapes

Method	Corrected Rand Index			
	Without Bagging	Comput. time	With Bagging	Comput. time
K-means	0,1613	25,25 s	0,2181	6,36 m
Hill Climbing	0,1784	3,22 m	0,1986	26,11 m
CLARANS	0,1165	31,37 s	0,2058	3,55 m