New Trends in Mathematical Sciences

http://dx.doi.org/10.20852/ntmsci.2022.455

On classification methods based on multiple correspondence analysis. Case study of distance education in Algeria using Python

Labdaoui Ahlam

Laboratory of Applied Mathematics and Modeling, Department of Mathematics, University Constantine 25000 Constantine, Algeria

Received: 25 December 2021, Accepted: 31 January 2022
Published online: 13 March 2022.

Abstract: To analyze survey questionnaire data we apply multiple correspondence analysis "MCA" as a method to help us convert data to cloud of points, but it is difficult to study it and get good results from it, so we have to do a classification to facilitate the study. Among the most useful classification methods, the CAH and the k-means. To compare them, we carried out a questionnaire on distance studies during the Corona-virus , which included the opinions of 304 university professors from most universities in Algeria. In our application, we used the python programming language.

Keywords: Survey, ACM, Classifications, Python.

1 Introduction

This increasingly complex world presents us with problems in the aid and production systems that humans may not be able to solve. This human shortage in this field have prompted humans to develop statistics and analyze data and extract statistical information, thanks to the graphic presentations provided that highlight the difficult relationships that need to be understood through direct data analysis [1].

2 Statistical methodology

2.1 Definition of the Multipie

Correspondence Analysis For a short ACM, is an extension of the correspondence factor analysis to summarize and visualize a data table containing more than two categories of vanabiles. It can also be considered as a generalization of principal component analysis when the variables to be analyzed are categorical rather than quantitative The ACM is generally used to analyze survey or survey data The ACM starts from a table of the compiet table(table Bart).

2.1.1 Data and evaluations

Multiple match analysis (MCA) is used to study a population of individuals described by qualitative Jvariable variables A qualitative (or nominal) variable is an application of the set of individuals in a finite set on which no structure is considered: for example, a set of three colors (white, red). The elements of this set are called terms of the variable and fon says for example that a blue individual has the blue modality The most common application of FACM is the processing of

[*] Corresponding author e-mail: ahlem_stat@live.fr

all survey responses (Each question is a variable whose modalities are the proposed answers from which each respondent can make a unique choice) We start by examining different ways to digitally transcribe all this data [2].

2.2 Classification algorithms

In addition to the methods for determining the main axes, classification methods are the second part of the geometric analysis of the data. Like all methods geometric data analysis, they aim to summarize the initial data; to do this, they produce homogeneous classes of objects so that objects of the same class resemble each other as much as possible and objects belonging to different classes stand out as possible. In other words, we look for classes, which, on the one hand, each form a coherent whole (compactness) and, on the other, are distinct from each other (reparability).

2.2.1 k-means

The $k-means$ clustering method is an unsupervised machine learning technique used to identify clusters of data objects in a data set. There are many types of clustering methods, but $k-means$ is one of the oldest and most affordable. These features make the implementation of $k-means$ in Python reasonably simple, even for novice programmers and data scientists.

The objective of the method is to partition the data into K groups and the value of K is set. The algorithm is relatively simple and it can be shown that at each stage of its execution, the value of $W(C)$ is decreased.

–We choose the number of K groups we want to obtain.
–The n observations are randomly partitioned into K groups.
–The coordinates of the centroids (the vector-mean) are calculated for each of the K groups, i.e
$\mu_k = \frac{1}{N_k}\sum_{i:C(i)=k} x_i k, \ \ k = 1,\ldots,K.$
–The distance between each observation and each of the K vector-means is calculated.
–Each of the n observations is assigned to the group with the closest mean vector.
–Repeat steps 3-5 until no observations are reassigned to a new group [3].

2.2.2 Ascending hierarchical classification

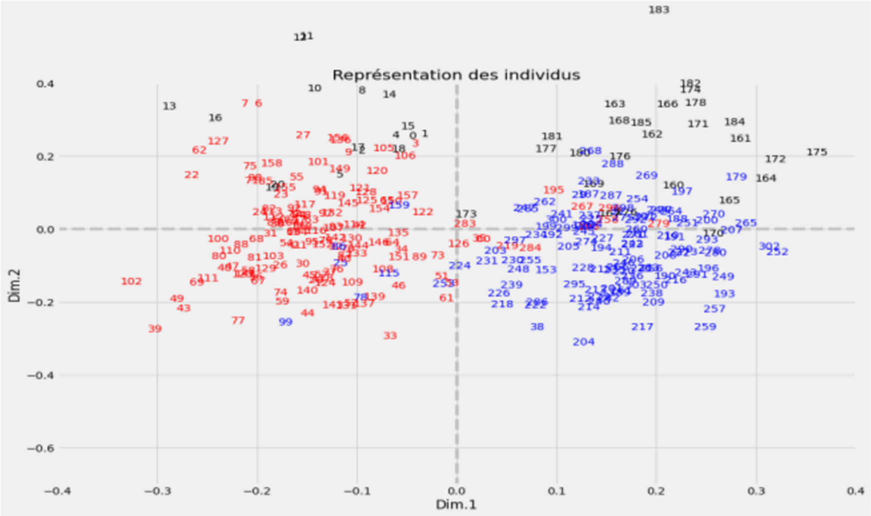
The HAC makes it possible to build an entire hierarchy of objects in ascending order. We begin by considering each individual as a class and try to merge two or more appropriate classes (depending on the similarity) to form a new class. The process is iterated until all individuals are in the same class. This classification generates a tree that can be cut at different levels to obtain a larger or smaller number of classes.
The hierarchical ascending classification algorithm is very simple. It is due to Lance and William (1967).

Initialization: Construction of the distance table, regardless of the formula used to build it because the HAC algorithm is independent of the metric used. Thus, between each couple of point (x, y) of M, we have a value d(x,y). The initial score is the finest ρ_0 of M.
Grouping: Browse the distance table to determine the nearest element torque (x*, y*) $d(x^*,y^*) \leq min_{x,y\in M} d(x,y)$. We combine the two elements in the same class $A = x^* \cup y^*$ the other classes remain unchanged. We get a new score ρ_i less. e than the previous one [4].

3 Application

Before the Covid-19 outbreak, we had no idea about distance education and had never studied remotely at our university, but after this crisis and continue the study, the intervention of the Ministry of Higher Education and Scientific Research imposed distance education on Algerian universities, and we chose the subject of distance learning because we wanted to deal with a subject in which we live as students. As we developed a questionnaire, asking questions about distance learning during the Covid-19 period. In this article, we chose to question only the professors from most of the wilaya of algeria, so we sent more than 1000 professors of the algerian university via their e-mail, we found the interest of the professors and we received 304 replies, which is sufficient to study the statistics and disseminate them to all professors of the algerian university. this topic was treated with mca and two different methods of hac and k-means classification that were used to find the best of them in order to select it as the best classifier that could be used in our statistical studies. Of course, statistical study needs a program, and we did not user or matlab, but instead we used python because it was not used at all in our university. We wanted a new program where we have to know it and learn it for the first time in this job. The version uses is python 3.9, the distribution used is anaconda with the following packages: numpy, pandas, matplotlib, fanalysis.

K-Means:
Hierarchical ascending classification



4 Discussion

Through our study, we noticed that the teachers' opinions about distance learning were somewhat similar; the students' results were similar to the results of previous years or lower, and even these were better because of the easier exams. One of the most important difficulties that teachers face in distance education is the weakness of internet, and the lack of means, especially for students, which was an obstacle to communicate with them, and would not have prevented the ability to evaluate them. In addition to all this, they encountered difficulties in preparing and presenting the courses, especially for those who have no previous experience on this teaching technique. Teachers' opinions on distance education are still somewhat positive, but several improvements must be made from improving the quality of internet and providing internet offers for teachers and students and it is better to provide free applications on Google Play Store containing all courses for each specialty, and very necessary to train students and teachers on this new technique. Nevertheless, the success of this mode of teaching requires the availability of adequate technological equipment and a high-speed internet connection. Nevertheless, these conditions are not met in our country; these shortcomings were among the most important reasons that led the rest of the professors to have a negative opinion about distance learning.

5 Conclusion

The goal of our work is to create a framework for comparing statistical classifications on a real data set (survey results: remote study) using Python, where automatically identifying similar data sets in a large data set is an important part of data mining. Automatic classification seeks to group data into clusters so that the data are more similar to each other within the same group than between groups. Since the concepts of similarity and grouping can be explained in several ways and among the existing methods, we have highlighted HAC and K-means to find out the most important differences between them and which is the best:

- k-means: It is considered the most widely used in big data analysis and increasing the number of k-groups increases its efficiency in classification, easy to understand and use, it collects similar results and displays the results quickly.
- Hierarchical ascending classification: This classification is suitable only for small mode. It provides richer information about the similarity structure of data. It is easy to extract several sections with different levels of "resolution".

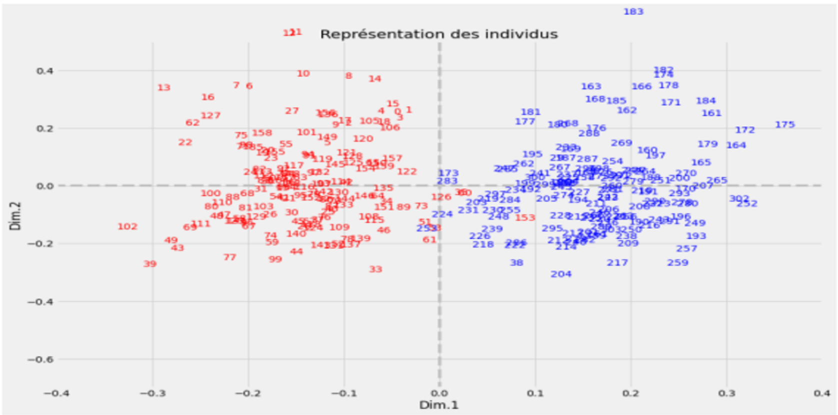
According to our application of these two classification methods on our data that used, the k-means classification gave satisfactory results compared to the HAC results, but not always because perhaps with other data, we can see that the HAC gives clearer results. In short, we say that for each specific data its own classification.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors have contributed to all parts of the article. All authors read and approved the final manuscript.



References

[1] Jacob Kogan, Introduction to Clustering Large and High-Dimensional Data, Cambridge University Press, Cambridge, 2007.

[2] MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. Dans Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, pages 281-297, 1967.

[3] MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. Dans Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, pages 281-297, 1967

[4] Python Data Analytics, Data Analysis and Science Using, Pandas, matplotlib, and the Python Programming Language,2012.