

Extracting Groups of Individuals Using Custom-Clustering

Santiago Esteva

VICOROB Institute, University of Girona, Spain

Albert Figueras

VICOROB Institute, University of Girona, Spain

Josep Lluís De la Rosa

VICOROB Institute, University of Girona, Spain

ABSTRACT

This paper presents a methodology to work on data or how to generate data to be analyzed by the Custom - Clustering algorithm proposed, to obtain groups of individuals or people's behavior on non-accounting opinions or interests, to be used to promote products, sales, advertising, etc. designed and oriented to a certain group of people who can be influenced by their interests. An example is presented discovering different patterns on a topic.

Keywords: Data Analysis, Cluster

INTRODUCTION

Artificial intelligence involves automated methods for using data and obtaining results. The interest in focusing the work presented is on understanding what AI can do and extracting comprehensible information from the data they are analyzing ambient intelligence [1] integrate local environment and the machines used to interpret as one single entity. AI is often used to reproduce behaviors, usually human, social [2] [3], learn from data and reproduce them using an algorithm on a machine. AI can arise to automatize driving and in robots [4][5]. Cluster analysis is used in economics [6] to know the movements of the economy, stock exchange, currency. And NN allows capturing behavior between input and output data with certain restrictions. If you also know the structure of the system, it allows you to sophisticate the algorithm and reproduce more behaviors, so it is interesting to capture information from the data as structure, groups, limits, etc.

In this work we use the cluster, is the process of assigning a homogeneous group of objects into subsets called *clusters*, and in dealing with multidimensional systems we must rely on mathematical tools that generate results, as in a simple three-dimensional representation the solution would be obvious. It is also not possible to study that in 3 dimensions by rotating the axis, because there are many cases in which the output data related with more than 3 axis.

To be able to repeat a behavior, the smarter it is to understand the behavior, more than to repeat it in comparison. Understanding behavior through data is more complicated than if you know about its behavior and look for structures that allow you to structure how the system works. For example, if we want to reproduce human behavior, in front : what is beauty, or the

impulses to buy, or ways of driving, etc. you will need to analyze the data to find patterns and diagrams and try to represent them.

There are neural network structures that adapt to data to reproduce behaviors and somehow model algorithms into data behavior and where a behavioral structure or hypothesis could be determined. Instead, a hypothesis can be used to shape an algorithm and try to fit the data, and to be able to reproduce the behavior.

In this article we will focus on finding data cores that represent patterns, that is, data groups that are distinguishable from each other, interpreting them, and assigning them to different behaviors.

First of all it is necessary to have the data of the process in good conditions that is to say without incongruities, with certain dispersion, which indicates high information contents. Incongruities on data can also arise due to the lack of axis representation, and using more dimensions can be distinguished, and indicate different situations. Once we have the data with enough variables, better with excess than in defect, they can be classified to find groups of information.

METHOD

The method used in this article is Custom-Clustering [10], which allows to find the correct number of data sets that identify existing patterns of behavior with available data. The algorithm works with two different parts, the first tries to find in a supervised way, the number of different groups, and the second part consists of assigning each group to the minimum axis that define it. At the end we assign to each group found some characteristics that define it.

Use cluster function provided from Matlab. [7]:

$T = \text{cluster}(Z, 'MaxClust', N)$ constructs a maximum of N clusters using distance as a criterion

Proceed to solve the problem is the following diagram:

- Boucle1:
- To decide the initial numbers of clusters N to be segmented the data.
- Validate the discovered groups, measuring the false results in the area of the cluster. (result in % OK, store the groups and results).
- Repeat augmenting $N+1$

With the optimal results and the obtained desired group apply following boucle for each group or cluster:

- Boucle2:
- Use -1 Axis , reduce one dimension,
- Value the results , if reduces the effectivity (% OK) maintain this axis
- Follow reducing axis

As a result is the group defined by the minimal axis.

The importance of determining the correct number of groups that the algorithm has to find is fundamental, as can be seen in the figure (1), which, if only two groups are sought, leaves undefined the medium, between them. After obtaining the groups they need to be assessed and the degree of data concentration within the group confirmed.

In the case of the use of colored data, where there are positive, negative or other results, the degree of belonging of each color within the group will be analyzed, thus obtaining dominant groups for each color.

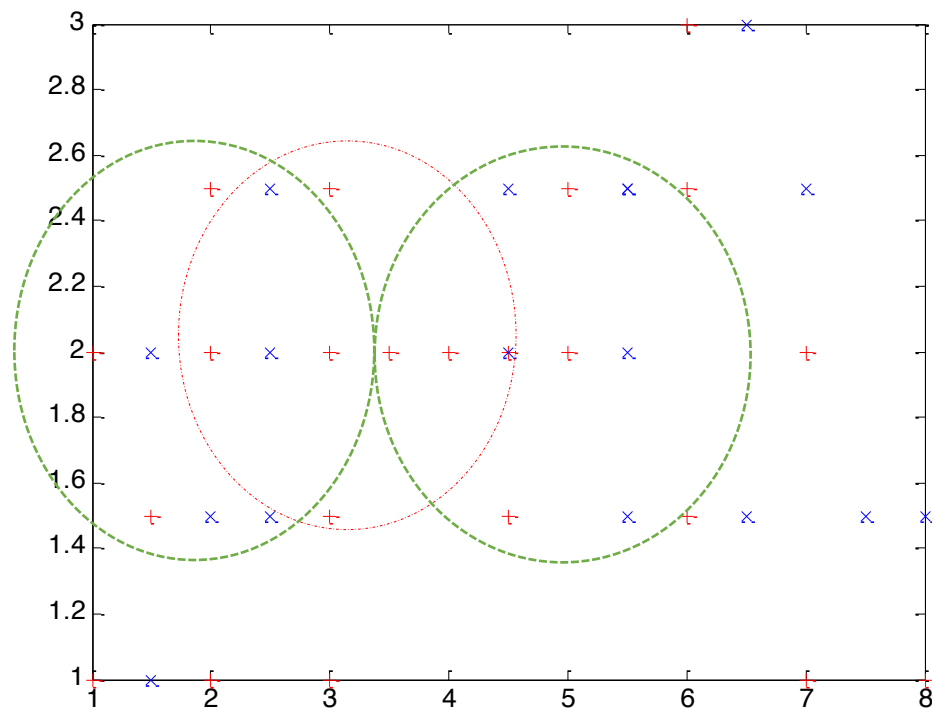


Figure 1. Different forms to group the red symbols depending on N

For example, if you want to know in which population groups a medical treatment has a good response, you will take data on patients, as age, previous illnesses, medication, blood type, etc., you can expand the data to is provided, but not all of them provide information, and it is not known if it only responds correctly to one or more groups of individuals. The proposed algorithm is responsible for finding the groups and defining their characteristics that define it, and from there the result of a new individual can be ensured before the treatment studied.

The algorithm works with data, when numerical data is not available, it will be necessary to give value to the physical or subjective characteristics, for example sensation of security, which will be able to be represented of very little to high, and middle, and to give value for ex. 1 to 5.

RESULTS

To understand how it works and not disclose personal or business data from pharmacological studies, is analyzed the canons of beauty, which will depend on who values it, cultural influences, fashion, etc., these parameters are not the proposed to analyzed otherwise the characteristics that make beauty. Based on the analysis of hundreds of award-winning faces in beauty contests, is characterized a large number of facial parameters, such as eye color, shape, eyebrows, nose, etc., since the excess of axes does not harm to define beauty. The important thing is to know if there is only one or more beauty profiles, and how they are defined.

In this example, they have analyzed white and Caucasian faces, and with a similar hairstyle, to separate it from the hair, which are the ones that make a big difference. From the first part of the algorithm arise two different groups, represented in a multispace of a large number of axes, and in the second part the axes are reduced with the indispensable ones, which are:

In the first group: face, lips, nose. Second group: face, eyebrows, nose. The relationships that are given into data for first group located are Long face, long eyebrows, long nose. And the other group Round face, small lips and round nose. Consequently, some of the cases that meet these criteria are in figure (2)



Figure 2, two types of faces determined by the Custom-Clustering algorithm.

CONCLUSION

This work defines the method to be able to classify data and obtain valuable information, for example extract profile of type of buyer, patient, customs, etc. Data where there is a great dispersion and ignorance that triggers this behavior. And a reduction of information is achieved, leaving only the valuable one, which defines the behavior, allowing the expert to interpret and take advantage of it. A simple and understandable example presented shows the possibilities of the method.

References

- [1] Charles Robinson D, Sanders D and Mazharsolook E. Ambient intelligence for optimal manufacturing and energy efficiency. *Assembly Automation*. 35/3 (2015) 234-248.
- [2] Alquist, J.S. and Breunig, C. (2011) Model-Based Clustering and Typologies in the Social Sciences. *Political Analysis*, 20, 92-112.
- [3] Aldenderfer, M.S. and Blashfield, R.K. (1984) *Cluster Analysis*. Quantitative Applications in the Social Science, Sage University Paper Series.
- [4] Figueras A, Esteva S, De La Rosa J.LL and Cufi X. Robot team in the improvement of the neighborhood. *IEEE International Smart Cities Conference (ISC2 2018)* Kansas
- [5] Figueras A, Esteva S, De La Rosa J.LL and Cufi X. Applying AI to the motion control in Robots System. *IAV 2019 10th IFAC Symposium on Intelligent Autonomous Vehicles*. 2019 Gdansk
- [6] Burns, R. and Burns, R. (2008) *Cluster Analysis*. In: *Business Research Methods and Statistics Using SPSS*. Sage Publications.
- [7] Matlab Data Analysis toolbox