Classifying Death Causes with Hierarchical Clustering: The Colombian Case

Isabel Cristina Puerta Lopera and Víctor Daniel Gil Vera

Universidad Católica Luis Amigó, Transversal 51A # 67B 90. Medellín, Colombia

Abstract: Hierarchical clustering is a clustering technique used primarily in natural and social sciences. Unlike the technique of k-means clustering, this allows us to understand the data at different levels of granularity, making it possible the recursively partition of \( n \) data points in 2, 3, 4 or \( n \) clusters. This can be represented by a tree rooted where all the sheets correspond to the data points given. Each internal node represents a cluster that consists of the data points corresponding to its leaves descendants. The objective of this paper was to present an application case of hierarchical clustering algorithm for classifying the leading death causes in the Colombian departments. Thanks to this classification, we identified some similarity patterns in the appearance of diseases in the north and west center regions of the country, which facilitates the identification of strategies and measures that contribute to the drop in the appearance of the same.

Key words: Machine Learning · Hierarchical Clustering · Big Data · Colombia · Diseases · Death

INTRODUCTION

Colombia is one of the Latin America countries that records the largest crude death rate. Deaths derived of health problems, murders and others, are recorded every year in all departments of the country. The objective of this paper was to apply the “hierarchical clustering” algorithm, to classify the departments into subgroups according to the most frequent causes of death.

We used the database of the colombian National Administrative Statistics Department (DANE) [1], which brought together the information of the leading death causes in 32 colombian departments. For the analysis and data processing we used the statistical software SPSS 25, the link between groups and the dissimilarity function "Euclidean distance to the square".

The sub-group formed by Antioquia, Valle del Cauca and Bogota departments and the sub-group formed by Santander, Tolima, Cundinamarca and Atlantic departments, are the ones who recorded the highest rates of similarity in the causes of death, cardiovascular diseases was the most common cause. These departments should adopt strategies that contribute to the drop of this diseases type and contribute to lead healthy life habits; how the decrease in consumption of tobacco and alcohol, avoid over weight, make physical activity and eating a diet low in saturated fats.

Hierarchical Clustering: This is an alternative approach to the clustering of partitions to group objects according to their similarity. In contrast to the grouping of partitions, the hierarchical clustering does not require you to specify the number of clusters that will arise [2]. The hierarchical clustering is subdivided into two types:

Agglomerative Clustering: Each observation is considered initially as its own cluster. Then, the more similar clusters are merged on until there is only a single large cluster [3].

Divisions Clustering: Is the inverse of agglomerative clustering, begins with the root, that all objects are included in a group. Then, the most heterogeneous clusters are divided successively until all observations are in its own cluster [3].

The result of the hierarchical clustering is a tree representation of the objects, which is also known as dendrogram [4]. The dendrogram is a multilevel hierarchy where the groups at a level join to form the groups at the
following levels [5]. This makes it possible to decide on the level at which to cut the tree to generate appropriate groups of a data object [5].

Death Causes in Colombia: Cardiovascular, respiratory and cerebrovascular diseases, urinary tract infections and homicides are the main causes of death in Colombia. Table 1, presents the official data set of death in the year 2016.

### MATERIALS AND METHODS

In the data analysis we used the latest database of the DANE with the information of the leading death causes in Colombia. For the analysis and data processing we used the statistical software SPSS 25, the link between groups and the dissimilarity function "Euclidean distance to the square". The calculation of the Euclidean distance is an important factor in most machine learning methods; \text{k- nearest neighbors} [6], \text{k- means} [7, 8] and learning vector quantization [9]. The equation to calculate the Euclidean distance to the square is the following:

\[
p = \sum (X_{ij} - X_{kj})^2
\]

where: \(X_1, X_2, \ldots, X_p\), are the observed variables, \(X_{ij}\) is the value observed in the \(i\)-th case in the \(j\)-th variable.

The grouping works "ascendant", each object is considered initially as a cluster of single element. At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster [10]. This procedure is iterated until all points are members of a single large group. The inverse of the agglomerative clustering is the divisive clustering and operates in "descending". It begins with the root, in which all objects are included in a single group. At each step of the iteration, the more heterogeneous cluster is divided into two. The process is iterated until all of the objects are in your own cluster [11].

After preparing the steps for performing a hierarchical clustering, selects the linking function to group objects in the hierarchical cluster tree, according to the dissimilarity information. The objects that are closest to each other
associated thanks to the linking method. Finally, it determines where to cut the tree to create the data partition.

RESULTS AND DISCUSSION

Tables 2 and 3, present the results of the data analysis, in which the employment the “Euclidean distance to the square” and the method of linking “Average link between groups”.

In Table 2, it can be seen that 76.7% of the sample data were valid, 23.3% were classified as missing values. The total of the sample was composed by 43 data.

In Fig. 2, it is noted that the first level was formed by a main cluster which were grouped together the departments of Bolivar, Antioquia, Valle del Cauca, Bogotá, Caldas, Cundinamarca and Atlántico. The main death cause of these departments were cardiovascular diseases, specifically acute myocardial infarction. Not only in Colombia, the coronary heart disease is the leading cause of mortality in the United States. In 2017, an estimated 695, 000 Americans will have a new acute myocardial infarction (AMI) and another 325, 000 will have a recurrent event [12].

On the second level, it formed a cluster with the departments of Valle del Cauca, Bogotá, Caldas, Cundinamarca, Atlántico, Tolima, Santander and Cesar. The main death cause of these departments were diseases of the urinary system. The urinary system is a common target site for toxicity of drugs and environmental chemicals [13]. The kidney is particularly susceptible because of the high blood flow to this organ relative to its mass and the unique property of renal tubular epithelium in concentrating urine and its constituents including drugs and chemicals [13]. At the third level a cluster formed by the departments of Caldas, Cundinamarca, Atlántico, Tolima, Santander, Cesar, Boyacá, Quindío, Sucre, Meta and Cauca. The main death cause of these departments were lung diseases. Older persons frequently report respiratory risk factors and symptoms and have a high prevalence of lung disease, most commonly obstructive airway disease, interstitial lung disease and lung cancer [14]. Notably, coexisting age-related nonrespiratory risk factors are also prevalent and may misidentify or modify respiratory diagnoses and their clinical course [14].

In the fourth level it formed a subgroup with the departments of Tolima, Santander, Cesar, Boyacá, Quindío, Sucre, Meta, Cauca, Magdalena, Huila, Norte de Santander, Risaralda, Nariño, Córdoba and Guajira. The main death cause of these departments were specifically pneumonia. This disease is one of the leading causes of hospital admission, morbidity and mortality among elderly patients and one of the leading causes of mortality and hospitalization among adults [15].

In the fifth and final level three subgroups were formed; the first formed by the departments of Antioquia, Valle del Cauca and Bogotá, the second by Cundinamarca, Atlántico, Tolima and Santander and the third by Cesar, Boyacá, Quindío, Sucre, Meta, Cauca, Magdalena, Huila, Norte de Santander, Risaralda, Nariño, Córdoba, La Guajira, Chocó, Arauca, Caquetá, Putumayo, Casanare, San Andrés, Vichada, Guaviare, Amazonas, Vaupés and Guanía. The main death causes of these departments subgroups were; aggression with firearms, malignant tumor of stomach, diabetes mellitus, hypertension and cerebrovascular diseases. Diabetes mellitus increases the risk of acute myocardial infarction, which can result in cardiogenic shock [16]. Cerebrovascular diseases, characterized by striking morbidity and mortality, have become the most common life-threatening diseases [17]. The existing drugs of cerebrovascular diseases target one or a few of pathogenic factors, the efficacy of which is limited because of the complexity of this disease [17].
Table 2: Summary of cases processing

<table>
<thead>
<tr>
<th>Cases</th>
<th>Valid</th>
<th>Missing</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Percent</td>
<td>Percent</td>
<td>Percent</td>
</tr>
<tr>
<td>Valid</td>
<td>33</td>
<td>10</td>
<td>43</td>
</tr>
<tr>
<td>Percent</td>
<td>76.7</td>
<td>23.3</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Source: author elaboration

Fig. 2: Dendogram presenting the graphical representation of the results. Source: author elaboration

Fig. 3: Colombian departments
Hierarchical clustering allows to identify objects based on the similarities that are presented. In this paper, it allowed to group the departments that have the highest similarity in the death causes. From the classification, identified some patterns of grouping, the leading causes of death in most departments that are located to the north of the country differ from those that are located in the center and south of the country. This type of clustering has the advantage that you can use any distance measure (euclidean, euclidean distance to the square, Manhattan, Mahalanobis, Maximum, Cosine similarity), unlike other clustering techniques, allowing you to perform best clusters.

REFERENCES

