

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

Table 1: Main death causes in Colombia

	Acute Myocardial Infarction	Other chronic obstructive pulmonary diseases	Pneumonia, a body not specified	Aggression with shot other firearms and not specified	Malignant tumor of the bronchi and lung	Malignant tumor of the stomach	Diabetes mellitus, not specified	Other cerebrovascular diseases	Essential Hypertension	Other disorders of the urinary system	Murders
Antioquia	4.255	2.103	1.233	1.213	1.192	737	589	582	508	455	31.390
Atlantico	2.221	597	516	378	377	261	234	218	217	179	11.669
Bogota	4.624	2.056	928	901	716	678	642	587	557	556	31.749
Bolivar	888	395	285	256	237	179	173	148	144	6778	7.621
Boyacá	579	195	191	190	158	135	117	108	106	15	970
Caldas	1.141	459	153	128	127	126	122	116	104	1807	5.973
Caqueta	87	81	53	41	36	36	34	32	30	20	208
Cauca	740	389	225	213	146	119	112	112	112	86	5.778
Cesar	517	238	185	108	98	96	89	89	80	61	4.261
Cordoba	904	375	314	204	200	137	132	131	126	126	6.434
Cundinamarca	2.374	983	415	333	285	255	215	206	184	183	12.786
Choco	122	103	50	42	40	32	32	30	27	25	1.370
Huila	840	331	219	167	130	125	111	102	101	94	5.795
La Guajira	202	155	99	61	54	52	44	41	39	36	2.154
Magdalena	792	255	236	133	132	124	115	97	88	71	5.065
Meta	679	211	177	142	105	99	97	90	84	83	4.753
Nariño	977	409	277	242	201	170	118	118	107	102	6.800
N. de Santander	937	442	418	300	203	185	180	132	130	109	7.045
Quindio	644	299	208	123	119	104	99	71	54	54	3.848
Risaralda	918	450	212	179	169	152	149	148	107	91	5.957
Santander	1.731	513	429	296	281	224	191	177	176	168	10.512
Sucre	670	193	172	125	102	94	90	85	61	59	3.615
Tolima	1.843	573	269	203	201	177	149	146	144	135	8.402
V. del Cauca	4.102	1.995	1.202	846	667	551	517	498	438	357	25.797
Arauca	99	57	53	25	25	24	23	21	21	20	1.028
Casanare	139	43	42	40	35	29	29	28	28	26	1.381
Putumayo	147	71	61	26	24	23	22	21	20	19	1.113
San Andres	51	16	14	12	10	8	8	8	6	6	269
Amazonas	10	8	7	7	6	5	5	4	4	4	187
Guainia	6	6	6	6	4	4	4	4	3	3	122
Guaviare	30	13	9	8	7	6	6	5	5	5	279
Vaupes	6	5	4	4	4	4	4	3	3	3	109
Vichada	25	12	11	7	7	6	5	5	5	4	189

Source: [1]

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; *k- nearest neighbors* [6], *k- means* [7, 8] and learning vector quantization [9]. The equation to calculate the Euclidean distance to the square is the following:

$$p \sum_{j=1} (X_{ij} - X_{sj})^2$$

where: X_1, X_2, \dots, 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 are

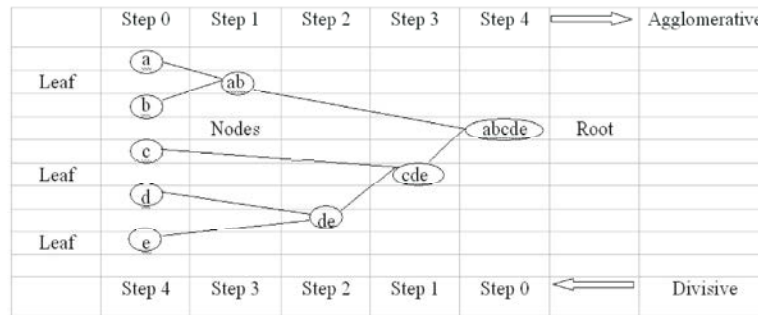


Fig. 1: Clustering process

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 Bolívar, 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]. Fig. 3, presents the political division of the colombian departments:

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 Antioquía, 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

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
33	76, 7	10	23, 3	43	100, 0

Source: author elaboration

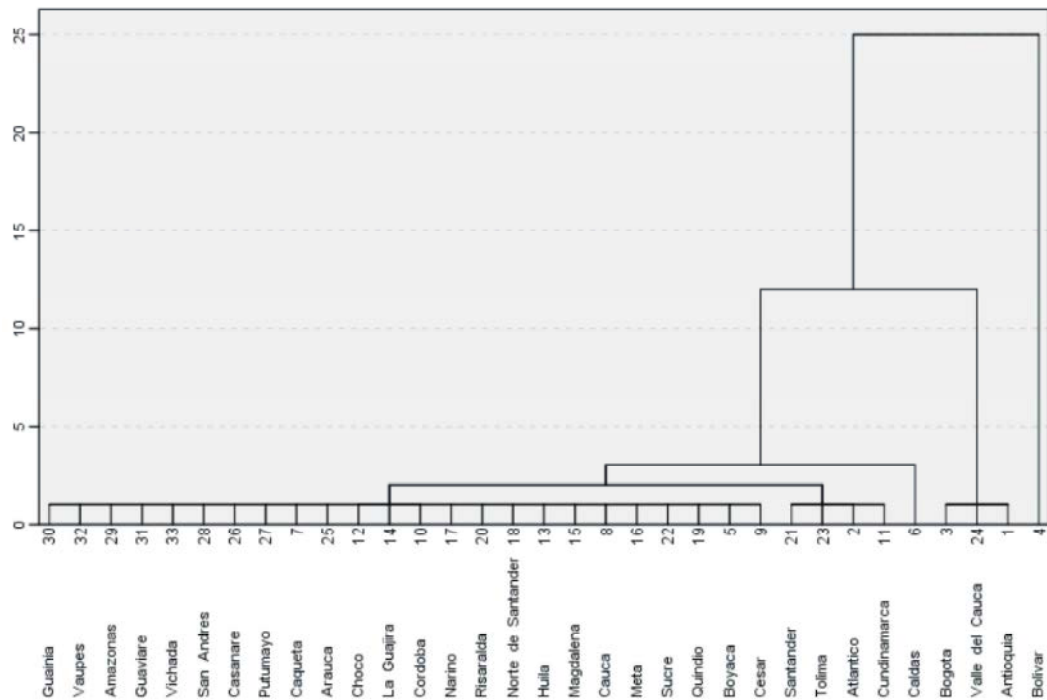


Fig. 2: Dendrogram presenting the graphical representation of the results. Source: author elaboration



Fig. 3: Colombian departments

Table 3: Conglomeration history

Stage	Combined Cluster			First appearance in the stage cluster		
	Cluster 1	Cluster 2	Coefficient	Cluster 1	Cluster 2	Next Stage
1	30	32	10, 0	0	0	3
2	31	33	33, 0	0	0	4
3	29	30	41, 0	0	1	4
4	29	31	500, 8	3	2	6
5	7	25	1564, 0	0	0	9
6	28	29	1581, 2	0	4	19
7	26	27	1773, 0	0	0	10
8	16	22	1932, 0	0	0	13
9	7	12	2624, 0	5	0	10
10	7	26	3451, 8	9	7	15
11	13	15	10269, 0	0	0	17
12	10	17	11690, 0	0	0	14
13	16	19	13285, 0	8	0	18
14	10	20	18443, 0	12	0	20
15	7	14	19942, 2	10	0	19
16	5	9	21511, 0	0	0	18
17	8	13	22265, 5	0	11	21
18	5	16	25197, 8	16	13	21
19	7	28	29301, 2	15	6	28
20	10	18	41043, 6	14	0	23
21	5	8	58812, 4	18	17	23
22	21	23	63840, 0	0	0	25
23	5	10	146185, 3	21	20	28
24	2	11	194741, 0	0	0	25
25	2	21	430510, 0	24	22	29
26	3	24	450144, 0	0	0	27
27	1	3	540960, 0	0	26	31
28	5	7	718053, 0	23	19	29
29	2	5	3419162, 0	25	28	30
30	2	6	3887246, 2	29	0	31
31	1	2	20676137, 4	27	30	32
32	1	4	46184144, 3	31	0	0

Source: author elaboration

CONCLUSIONS

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

1. DANE, 2018. "Defunciones no fetales".National Administrative Statistics Department, Bogotá D.C., Colombia. Retrieved March 10, 2018 from: https://www.dane.gov.co/files/investigaciones/poblacion/2017/22-diciembre-2017/nofetales2016/causas_defuncion_2016def.xls.
2. Ball, G.H. and D.J. Hall, 1967. "A clustering technique for summarizing multivariate data," Syst. Res. Behav. Sci., 12(2): 153-155.
3. Gowda, K.C. and G. Krishna, 1978. "Agglomerative clustering using the concept of mutual nearest neighbourhood," Pattern Recognit., 10(2): 105-112.

4. Mondal, S.A., 2018. "An improved approximation algorithm for hierarchical clustering," *Pattern Recognit. Lett.*, 104: 23-28.
5. Mazumder, O., A.S. Kundu, P.K. Lenka and S. Bhaumik, 2016. "Ambulatory activity classification with dendrogram-based support vector machine: Application in lower-limb active exoskeleton," *Gait Posture*, 50: 53-59.
6. Cover, T. and P. Hart, 1967. "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, 13(1): 21-27.
7. MacQueen, J., 1967. "Some Methods for Classification and Analysis of Multivariate Observations," in the proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp: 281-297.
8. Rauf, A., S.M. Sheeba, S. Khusro and H. Javed, 2012. "Enhanced k-mean clustering algorithm to reduce number of iterations and time complexity," *Middle-East J. Sci. Res.*, 12(7): 959-963.
9. Kohonen, T., 2012. *Self-organization and associative memory*. Springer Science & Business Media, pp: 125.
10. Leela, V. and R. Manikandan, 2014. "Comparative Study of Clustering Techniques in Iris Data Sets 1," *World Appl. Sci. J.*, 29: 24-29.
11. Pavithra, M., 2017. "Cluster Ensemble Approach for Semi Supervised Clustering 1," *World Appl. Sci. J.*, 35(1): 75-81.
12. Castro-Dominguez, Y., K. Dharmarajan and R.L McNamara. 2018. "Predicting death after acute myocardial infarction, " *Trends in Cardiovascular Medicine*, 28(2): 102-109.
13. Echouffo-Tcheugui, J.B., D. Kolte, S. Khera, H.D Aronow, J.D. Abbott, D.L. Bhatt and G.C. Fonarow, 2018. "Diabetes Mellitus and Cardiogenic Shock Complicating Acute Myocardial Infarction, " *The American Journal of Medicine*, 131(7): 778-786.
14. Khan, K.N.M., G.C. Hard, X. Li and C.L. Alden, 2018. Chapter 11 - Urinary System. In M. A. Wallig, Eds., W. M. Haschek, C. G. Rousseaux, and B. Bolon, Academic Press, pp: 213-271.
15. Najafi, S. and C. Sandrock, 2017. "Hospitalized Patients with Acute Pneumonia," *Hospital Medicine Clinics*, 6(4): 456-469.
16. Vaz Fragoso, C.A., 2017. "Epidemiology of Lung Disease in Older Persons, " *Clinics in Geriatric Medicine*, 33(4): 491-501.
17. Wang, J., L. Zhang, B. Liu, Q. Wang, Y. Chen, Z. Wang and Y. Wang, 2018. "Systematic investigation of the Erigeron breviscapus mechanism for treating cerebrovascular disease, " *Journal of Ethnopharmacology*, 224: 429-440.