Contemporary Engineering Sciences, Vol. 10, 2017, no. 35, 1709 - 1717
HIKARI Ltd, www.m-hikari.com
https://doi.org/10.12988/ces.2017.711188

Classification of Levels and Monthly Flows for Hydraulic Affluent through Kmeans

Fernando Mesa

Department of Mathematics and GEDNOL
Universidad Tecnológica de Pereira
Pereira, Colombia

Pedro Pablo Cárdenas Alzate

Department of Mathematics and GEDNOL
Universidad Tecnológica de Pereira
Pereira, Colombia

Carlos Alberto Rodríguez Varela

Department of Mathematics
Universidad Tecnológica de Pereira
Pereira, Colombia

Abstract

This document presents the classification into three groups of the months of the year by means of the Kmeans algorithm or average distances. The algorithm allows grouping a set of observations in finite number of groups, the groups in which were classified the level and flow of a tributary were: Greater (wet), intermediate and lower (dry).

Keywords: Grouping, centroid, cluster, clustering, Kmeans

1 Introduction

There are many ways to use and manage databases depending on the amount of information, samples, types of samples, measured variables among others, for example, you can initially think of calculating variables of mean, median, standard deviation and arriving to a series of conclusions according to these statistical indicators. On the other hand, the data can be classified into homogeneous groups to visualize how much of these data have a similar behavior. The Kmeans method is a clustering algorithm that allows you to subdivide a set of data of the same nature or of a varied nature into K subgroups that you want, that is, if you have a set of m samples and each sample consists of n variables and it is required to classify or divide into subgroups the Kmeans method can be used to classify in K subgroups the data set representing m rows or samples with n variables. To achieve the aforementioned, it is necessary to perform a series of calculations that consist in determining the distances of each sample to a centroid by variables, this distance should be as small as possible because we want to group the largest number of samples in the same subgroup [1]. On the other hand, the water resource is an indispensable source for human existence and the tributaries undergo changes at different times of the year due to cooling and heating phenomena. These changes are reflected in variables such as the flow, level of the tributary, speed of the tributary, temperature, among others. Experts in climate change and hydrology say that the phenomenon of cooling is the main cause of climatic anomalies such as heavy rains and overflow of rivers in tropical areas or near the equator. In this paper we find the classification of the months of a year in three groups, the first group represents the driest months, the second group the intermediate group and the third group the wettest, the classification is determined by means of the Kmeans algorithm and the relationship between flows and levels of an arbitrary tributary [2].

2 Kmeans Algorithm

The problem to be addressed is the classification of the months of the year into three groups. In the first group you will find the months with lower flow and low river levels, the second subdivision groups the intermediate months where the levels are not exactly low or high according to the database provided by *IDEAM*, finally, the groups are grouped months of the year where the river levels are high and the flows increase in comparison to the first two subgroups. The data form rows and columns of a matrix, each element of the row represents a sample and each column represents a variable, for the case study of this paper, each column represents the flow or level of river in each month of the year and each row corresponds to a daily sample of the flow

or river level respectively. In turn, each column is a variable to be classified within the three subgroups by means of the Kmeans algorithm. Therefore we have [3,4]:

i) Define the number of clusters or regions where you want to group the data set.

ii) Initialize the centroids of the k groups or regions of the previous item.

iii) Calculate the inter-distances between each sample with each centroid: Calculation of the Euclidean distance between each sample or data and the centroids.

$$\min : \sum_{g\in\Omega k}^{k}\sum_{i\in\Omega g}^{n} \|xi - cg\| \tag{1}$$

$$d(i,g) = \|xi - cg\| \tag{2}$$

$$d(g) = \sqrt{\sum_{i=1}^{n}(xi - cg)^2} \tag{3}$$

iv) Assign each element or sample to the nearest centroid, that is, the minimum distance between each element and its proximal centroid with equations (1) and (3).

v) Recalculate the next centroid:

$$c_g^{t+1} = \frac{1}{ng}\sum_{i\in\Omega g} xi \tag{4}$$

where $i \in \partial g$. Each of the elements that belong to the centroid or set $g\Omega g$.

vi) Evaluate the elements or samples that changed cluster after resigning by distances with equation (2), executing steps 3, 4, 5.

vii) If there are no elements or samples that change cluster or group then the process stops. Otherwise, the optimality criterion of equation (1) is revised and steps from 3 to 6 are repeated or until the maximum number of iterations is met.

3 Samples of flow and level of tributary

The affluent from which the flow and level measurements were obtained is located in the department of Caquetá, municipality of Florencia, Orteguaza tributary [5,6]. Latitude: 0133 N, longitude: 7531 W, elevation: 0520 m.s.n.m. The samples are purified and organized in Annex I of this document. The measures of flow and levels of tributary are source of information to determine that times or months of the year are rainier than others and vice versa. With the classification into groups of months: Humid (Rainy), intermediate or dry (not very rainy) it is possible to estimate at what times of the year the reservoirs that supply the hydroelectric power stations are at an upper level or at a minimum level to alert the competent authorities about of rationing of water and electric power, this seeing from the worst scenario. On the other hand, when classifying the seasons of the year and knowing the wettest months, the amount of electrical energy to be generated could be assessed. This type of classifications can not only be applied to the electricity generation sector but also to the prevention and health promotion sector of the departments with high risk of proliferation of respiratory diseases, that is, there are times of the year where insects that generate respiratory diseases in remote areas of the national territory. By means of the classification with the Kmeans method it is possible to mitigate the health effects of the rainy seasons and improve the quality of life of the inhabitants of the remote areas of the national territory affected by climatic phenomena [7].

4 Application and results

$$
D= \begin{bmatrix}
1.7277 & 6.3676 & 0.0386 \\
1.7277 & 6.3676 & 0.0386 \\
1.9832 & 6.9130 & 0.0740 \\
1.4774 & 5.7552 & 0.0579 \\
0.9333 & 4.7874 & 0.3201 \\
3.0674 & 0 & 6.0347 \\
0.3058 & 2.4274 & 2.5540 \\
0.3058 & 4.3191 & 0.8938 \\
1.1036 & 5.6923 & 0.0934 \\
1.4409 & 6.1207 & 0.0351 \\
1.3531 & 5.7983 & 0.0591 \\
1.8240 & 6.8777 & 0.0458
\end{bmatrix}
$$

Each element of each column represents the minimum distance between each of the centroids with each monthly flow sample.

$$
\text{sumd}= \begin{bmatrix} 6.1161 & 0 & 7.7664 \end{bmatrix}
$$

Each element of the row of *sumd* represents the minimum distance of equation (1) between each sample and each centroid.

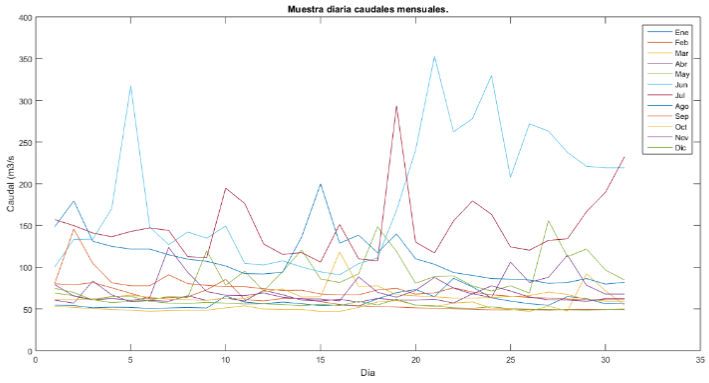

Figure 1: Samples of daily flows $\frac{m^3}{s}$.

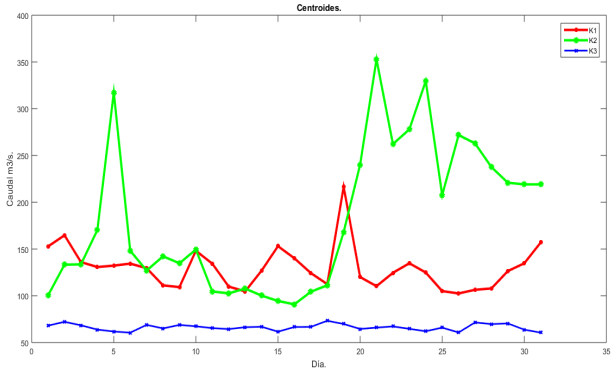

Figure 2: Centroids. Group 1 red line, group 2 green line, group 3 blue line.

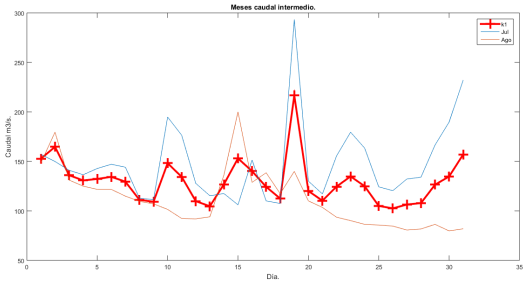

Figure 3: Centroid 1, group 1 of months with intermediate flow.

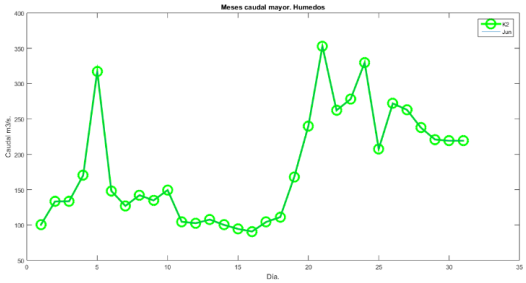

Figure 4: Centroid 2, group 2 of month with greatest flow.

$$
\begin{array}{lccc}
\text{D1}= & [1.7277 & 6.3676 & 0.0386 \\
 & 1.6963 & 6.7769 & 0.0527 \\
 & 1.9832 & 6.9130 & 0.0740 \\
 & 1.4774 & 5.7552 & 0.0579 \\
 & 0.9333 & 4.7874 & 0.3201 \\
 & 3.0674 & 0 & 6.0347 \\
 & 0.3058 & 2.4274 & 2.5540 \\
 & 0.3058 & 4.3191 & 0.8938 \\
 & 1.1036 & 5.6923 & 0.0934 \\
 & 1.4409 & 6.1207 & 0.0351 \\
 & 1.3531 & 5.7983 & 0.0591 \\
 & 1.8240 & 6.8777 & 0.0458]
\end{array}
$$

Each element of each column represents the minimum distance between each of the centroids with each monthly level sample.

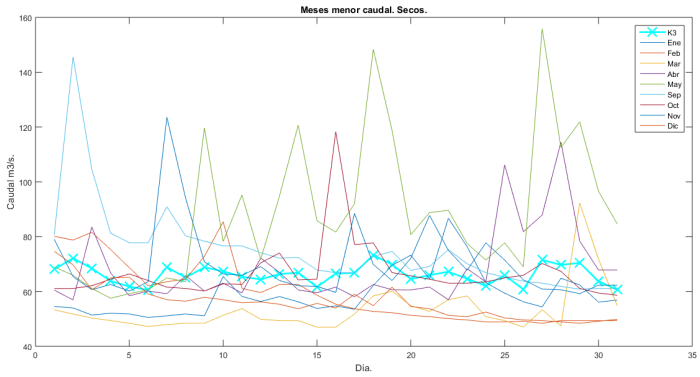

Figure 5: Centroid 3, group 3 of months with lower flow.

$$\text{sumd1}= [3120 \quad 0 \quad 3934]$$

Each element of the row of sumd1 represents the minimum distance of equation (1) between each sample and each centroid.

5 Conclusion

The kmeans algorithm allows grouping and assigning to each centroid a set of data close to it, since what is sought is to minimize the distance between each sample to a nearby centroid. In this document both the affluent and flow level samples were grouped into three groups respectively. The initial centroids were proposed as a constant line with the highest value, a constant line for an

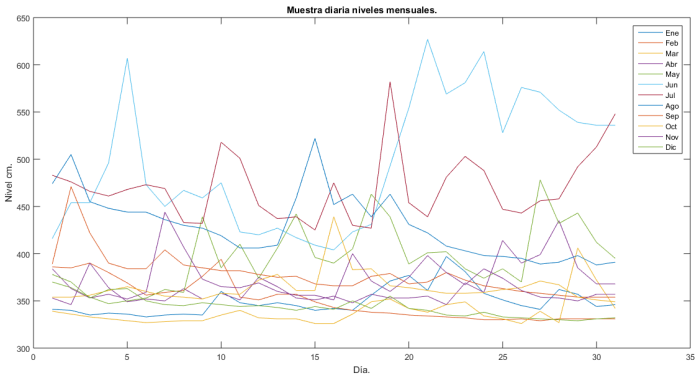

Figure 6: Daily level samples.

average value and a constant line with the minimum value of both the flow and level samples. In order to classify the months between wet, intermediate and dry months, it is necessary to analyze figures 4, 5 and 6, since together they give information to assign each sample to a nearby centroid. The method of Kmeans does not classify the samples by itself, it only indicates the indices to which the samples belong, the classification and interpretation of the data must be done according to the nature of the samples.

Acknowledgements. We would like to thank the referee for his valuable suggestions that improved the presentation of this paper and our gratitude to the Department of Mathematics of the Universidad Tecnológica de Pereira (Colombia) and the group GEDNOL.

References

[1] D. Arthur and S. Vassilvitskii, Kmeans++: The Advantages of Careful Seeding, *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, Vol. 7, (2007), 1027-1032.

[2] Pavel Berkhin, *Survey of Clustering Data Mining Techniques*, Technical Report, Accrue Software, San Jose, CA, 2002.

[3] Stuart P. Lloyd, Least squares quantization in pcm, *IEEE Transactions on Information Theory.,* **28** (1982), 129-137. https://doi.org/10.1109/tit.1982.1056489

[4] C. Cardona, J. Henao, Predicción de los precios de contratos de electricidad usando una red neuronal con arquitectura dinámica, *Innovar.,* **20** (2010).

[5] J. Hamilton, *Time Series Analysis*, *Princeton University Press.,* Vol. 2, 1994.

[6] Alan Dagoberto Arias-Hernndez, Ricardo Alberto Hincapi-Isaza, Ramn Alfonso Gallego-Rendn, Comparación de flujos de carga probabilśticos empleados en sistemas de distribución levemente enmallados, *Scientiae et Technica,* **19** (2014), 153-162.

[7] A. Doucet, N. Gordon, An introduction to sequential Monte Carlo methods, Chapter in *Sequential Monte Carlo Methods in Practice,* Springer, 2001, 3-14. https://doi.org/10.1007/978-1-4757-3437-9_1

Received: December 4, 2017; Published: December 24, 2017