Non-Supervised Clustering

Mariajose Franco Orozco
mfrancoo@eafit.edu.co
EAFIT University
Mathematical Engineering
Medellín, Colombia

I. Introduction

Non-supervised clustering, also known as unsupervised clustering, is a machine-learning technique that consists of grouping data points into clusters based on their similarities. What makes it unsupervised is that there is no need for prior knowledge or labels as in supervised learning. This is because non-supervised clustering aims to identify patterns and structures within the dataset [1].

In the actual world, it is common to deal with huge volumes of data which include images, videos, text, and more. Organizing such data into rational groups is important in order to analyze these data. Clustering techniques are useful for completing this task of organizing data into different groups [2].

There exist many clustering algorithms, but in this paper, we will perform 4 of them: Mountain Clustering, Subtractive Clustering, K-means Clustering, and Fuzzy C-means Clustering. Clustering algorithms tend to use different similarity metrics, we will implement 4 metrics: Euclidean, Mahalanobis, Manhattan, and Cosine distance. These metrics determine the similarity between the data points, and the algorithms determine the clusters according to these similarities [1].

In recent years, there has been a growing interest in non-supervised clustering, because of this, it has a lot of applications fields, for example, data mining, pattern recognition, image analysis, bioinformatics, and more [3].

II. Theoretical Framework

This section will explain the theory of the concepts used in this project.

A. Metrics

4 metrics were considered: euclidean, mahalanobis, manhattan, and cosine distances. As it was mentioned above, these metrics are in charge of determining the similarity between data points.

1) Euclidean distance:

$$d(A,B) = \sqrt{\sum_{i=1}^{d}(A_i - B_i)^2} \tag{1}$$

where d is the dimension of the space.

2) Mahalanobis distance:

$$d(A,B) = \sqrt{(A-B)C^{-1}(A-B)^T} \tag{2}$$

where C is the covariance of the dataset, and C^{-1} is the inverse covariance.

3) Manhattan distance:

$$d(A,B) = \sum_{i=1}^{d}|A_i - B_i| \tag{3}$$

where d is the dimension of the space.

4) Cosine distance:

$$d(A,B) = 1 - \frac{\sum_{i=1}^{d} A_i B_i}{\sqrt{\sum_{i=1}^{d} A_i^2 \sum_{i=1}^{d} B_i^2}} \tag{4}$$

where d is the dimension of the space.

B. Clustering Algorithms

Also, 4 clustering algorithms were implemented as well: mountain, subtractive, k-means, and fuzzy c-means clustering algorithms.

1) Mountain Clustering:

2) Subtractive Clustering:

3) K-Means Clustering:

4) Fuzzy C-Means Clustering:

III. Methodology

IV. Experimentation

V. Results

VI. Conclusions

References

[1] M. G. Omran, A. P. Engelbrecht, and A. Salman, "An overview of clustering methods," *Intelligent Data Analysis*, vol. 11, no. 6, pp. 583–605, 2007.

[2] A. Ghosal, A. Nandy, A. K. Das, S. Goswami, and M. Panday, "A Short Review on Different Clustering Techniques and Their Applications," in *Advances in Intelligent Systems and Computing*, 2020.

[3] L. V. Bijuraj, "Clustering and its Applications," *Proceedings of National Conference on New Horizons in IT - NCNHIT 2013*, pp. 169–172, 2013.