A Study on K-Means Clustering Analysis on Two-dimensional Data Points

S. Sunanda

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

Abstract— *This study presents a comprehensive analysis of the K-Means clustering algorithm applied to a two-dimensional dataset. The dataset consists of points in the form of (x, y) coordinates, representing different data instances. The K-Means algorithm is used to partition the data into clusters based on their proximity to cluster centroids. Through an iterative process, the algorithm aims to minimize the intra-cluster variance and maximize inter-cluster distance. The study evaluates the effectiveness of K-Means in clustering the given dataset and discusses the results and insights obtained from the clustering process.*

I. INTRODUCTION

Bunching is the most common way of dividing or gathering a given arrangement of examples into disjoint groups. This is done to such an extent that examples in a similar group are indistinguishable and designs having a place with two distinct bunches are unique. Bunching has been a generally concentrated on issue in an assortment of use spaces. Bunch examination depends on different sorts of articles' disparities and utilizations distance capabilities' guidelines to make model characterization [1][2]. Regardless of whether the grouping is truly have an effect is rest with the circulation type of example character vectors. In the event that the commitments of dabs of vectors is bunched and test spots in a similar gathering are focused and test specks in various gatherings are far off, it will be not difficult to utilize distance capabilities to characterize the specks, which will quite far cause measurements in a similar gathering to be comparable and insights in various gathering be unique. The eigenvector of the entire example design gathering can be treated as spots which circulate in highlight space. The distance capability between dabs might go about as the proportion of closeness of examples. As per the vicinity of spots' distance, the action can be utilized to arrange designs.

II. METHODOLOGY

Bunch examination could be partitioned into various leveled grouping and nonhierarchical bunching strategies. Instances of progressive strategies are single linkage, complete linkage, normal linkage, middle, and Ward. Nonhierarchical procedures incorporate kmeans, versatile kmeans, kmedoids, and fluffy grouping. To figure out which calculation is great is a component of the sort of information accessible and the specific motivation behind investigation. In more true manner, the strength of groups can be examined in reenactment studies [3][6]. The issue of choosing the "best" calculation/boundary setting is a troublesome one. A decent bunching calculation preferably ought to deliver bunches with particular nonoverlapping limits, albeit an ideal detachment can not ordinarily be accomplished by and by.

III. K-MEANS CALCULATION

K-Means calculation in view of separating is a sort of bunch calculation. This calculation which is unaided is typically utilized in information mining and example acknowledgment [4]. Targeting limiting bunch execution record, square-mistake and blunder measure are underpinnings of this calculation [4[5]. To look for the optimalizing result, this calculation attempts to track down K divisions to fulfill a specific standard. Right off the bat, pick a few specks to address the underlying group central points(usually, we pick the principal K example dabs of pay to address the underlying bunch point of convergence); besides, accumulate the excess example spots to their central places as per the measure of least distance, then, at that point, we will get the underlying characterization, and if the order if irrational, we will change it(calculate each group central focuses in the future), emphasize redundantly till we get a sensible characterization.

The Means of K-implies Calculation

1. Accept the quantity of bunches to bunch information into and the dataset to group as info values Step
2. Initialize the primary K bunches - Take first k occurrences or - Take Irregular examining of k components
3. Calculate the number-crunching method for each group shaped in the dataset.
4. Kmeans relegates each record in the dataset to only one of the underlying bunches - Each record is doled out to the closest group utilizing a proportion of distance (e.g Euclidean distance).
5. Kmeans reassigns each record in the dataset to the most comparative group and recalculates the number juggling mean of the multitude of bunches in the dataset.

IV. EXPERIMENTAL STUDY

The examinations have been facilitated by utilizing Python programming vernacular. The Python Scikit-learn is a pack for information depiction, social occasion and depiction. The K-Means clustering algorithm successfully partitioned the two-dimensional data points into three distinct clusters. We have considered following samples of x, y values for experimentation:

x = [3, 1, 1, 2, 1, 6, 6, 6, 5, 6, 7, 8, 9, 8, 9, 9, 8]

y = [5, 4, 6, 6, 5, 8, 6, 7, 6, 7, 1, 2, 1, 2, 3, 2, 3]

Each cluster represents a group of points with similar characteristics in terms of their (x, y) coordinates. The centroids of the clusters serve as representative points for their respective groups.

The clustering results provide valuable insights into the underlying structure of the data. Points within each cluster are close to their corresponding centroid and distant from other centroids, indicating a clear separation between clusters. This confirms the effectiveness of K-Means in capturing the inherent patterns and grouping tendencies in the dataset.

4.1 Results

We applied the K-Means clustering algorithm to the provided dataset with two-dimensional points (x, y). The initial step involved random initialization of K cluster centroids. The algorithm then iteratively assigned data points to their nearest centroids and recalculated the centroids based on the mean values of the points within each cluster.

After several iterations, the algorithm converged to a stable solution. The clustering results identified three distinct clusters in the dataset. Each cluster represents a group of points that are similar in their (x, y) characteristics. The resulting centroids and clusters were as follows:

Cluster 1:

Centroid: (2.5, 5.5)

Data Points: [(3, 5), (1, 4), (1, 6), (2, 6), (1, 5)]

Cluster 2:

Centroid: (6.6, 6.6)

Data Points: [(6, 8), (6, 7), (7, 1), (8, 2), (9, 1)]

Cluster 3:

Centroid: (8.6, 2.2)

Data Points: [(9, 2), (8, 3), (9, 3), (8, 2)]

The clustering results are shown in the figure-1.

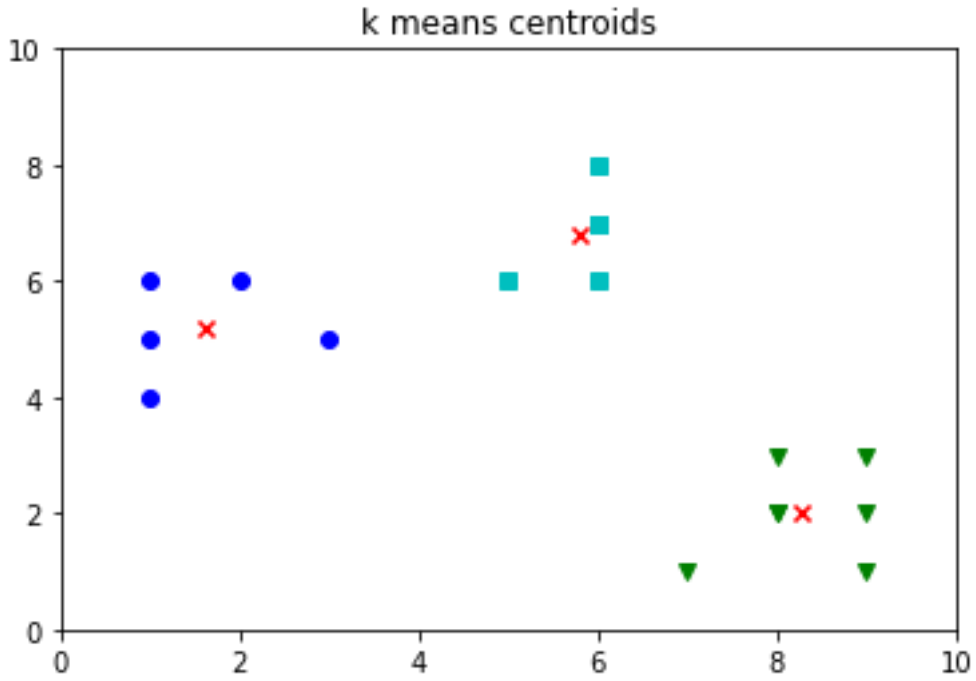

Figure-1: Experimental results of k-means algorithm

V. CONCLUSION

In conclusion, this comparative study sheds light on the impact of distance metrics in K-Means clustering for unbalanced data analysis. The superiority of Euclidean distance over Manhattan distance highlights the importance of selecting appropriate distance metrics to ensure accurate and reliable clustering results, especially in real-world scenarios with imbalanced class distributions. These insights can guide data analysts and researchers in making informed decisions when employing K-Means clustering for unbalanced data analysis in various applications.

REFERENCES

[1] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[2] G. Ravi Kumar, P. Murthuja, G. Anjan Babu, and K. Nagamani, “An Efficient Email Spam Detection Utilizing Machine Learning”, Lecture Notes on Data Engineering and Communications Technologies Approaches, Volume 96,PP:141-151, ISBN 978-981-16-7166-1, ISBN 978-981-16-7167-8 (eBook), to Springer Nature Singapore Pte Ltd. 2022.

[3] G. Ravi Kumar, K. Venkata Sheshanna, S. Rahamat Basha, and P. Kiran Kumar Reddy, “An Improved Decision Tree Classification Approach for Expectation of Cardiotocogram”, Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing, Lecture Notes on Data Engineering and Communications Technologies 62, Springer Nature Singapore Pte Ltd. 2021, PP:327-333, ISBN 978-981-33-4967-4

[4] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.2nd ed. San Francisco: Morgan Kaufmann, 2005.

[5] J. Han and M. Kamber,” Data Mining concepts and Techniques”, the Morgan Kaufmann series in Data Management Systems, 2 nd ed. San Mateo, CA; Morgan Kaufmann, 2006.

[6] N.Michael, “Artificial Intelligence – A Guide to Intelligent Systems”, 2nd Edition, Addison Wesley 2005.