Comparative Study of K-Means Clustering with Different Distance Metrics for Unbalanced Data Analysis: Evaluating Euclidean and Manhattan Distance

P. Bhuvaneswari

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

Abstract— *K-Means clustering is a widely used unsupervised learning technique for data grouping and pattern discovery. In this comparative study, we investigate the performance of K-Means clustering with two popular distance metrics, Euclidean distance and Manhattan distance, on unbalanced datasets. Unbalanced datasets are common in real-world scenarios, where certain classes have significantly fewer instances than others. The study aims to understand how different distance metrics impact the clustering results for unbalanced data and identify the distance metric that provides better performance. We utilize various evaluation metrics to assess the clustering performance and draw meaningful insights to support informed decision-making in data analysis.*

I. INTRODUCTION

Clustering analysis is a powerful technique used to group similar data points together based on their similarity or distance. Bunching is an unaided review. The fundamental point of bunching is to partition a dataset into a few unique subsets (known as 'groups') to such an extent that information into a specific subset having comparative properties while information in various subset showing various properties from information in another subset. Intends to say that bunching ought to fulfill the two properties [1][2][4]: 1. High Firm Property and 2. Low Coupling Property. Bunching calculations are predominantly isolated into two sorts in view of created group properties: various leveled and partitional. The progressive techniques, in everyday attempt to deteriorate the dataset of n objects into an order of gatherings. This progressive deterioration can be addressed by a tree structure chart called as a dendrogram whose root hub addresses the entire dataset and each leaf hub is a solitary object of the dataset. The bunching results can be gotten by cutting the dendrogram at various level. There are two general methodologies for the progressive strategy: agglomerative and troublesome [3][4]. Agglomerative methodology is a granular perspective beginning with n-leaf hubs, letting them as individual bunches, moving upwards towards the root for certain consolidating models. While disruptive various leveled grouping method is a big picture perspective, beginning from the root hub slowly dividing the information into various bunches downwards founded on the properties of the information.

II. K-MEANS CALCULATION

The K-Means is one of the renowned parcel grouping calculation. It takes the information boundary k, the quantity of groups, and parcels a bunch of n objects into k bunches so the subsequent intra-group similitude is high yet the between group closeness is low. The fundamental thought is to characterize k centroids, one for each group [5]. These centroids ought to be set in a shrewdness way due to various area causes various outcomes. Thus, the better decision is to put them however much as could reasonably be expected far away from one another. The subsequent stage is to take each guide having a place toward a given informational collection and partner it to the closest centroid. At the point when no point is forthcoming, the initial step is finished and an early groupage is finished. Right now we really want to re-compute k new centroids. After we have these k new centroids, another limiting must be finished between similar informational index focuses and the closest new centroid. A circle has been produced. Because of this circle we might see that the k centroids change their area bit by bit until no more changes are finished [6][7]. At the end of the day centroids move no more. At last, this calculation targets limiting a goal capability, for this situation a squared blunder capability.

The Proper Calculation is:

1. Select K focuses as introductory centroids.
2. Rehash.
3. Structure k bunches by relegating all focuses to the nearest centroid.
4. Recompute the centroid of each cluster.
5. Until the centroids don't change

III. DISTANCE ESTIMATIONS IN K-MEANS CALCULATIONS:

In K-Means calculation, we ascertain the distance between each place of the dataset to each centroid instated. In view of the qualities found, focuses are doled out to the centroid with least distance. Consequently, this distance estimation assumes the fundamental part in the grouping calculation. As we probably are aware, distance between two focuses can be registered with various strategies accessible, thus, our principal point is to get a legitimate strategy from the accessible ones. However, in picking such procedures, a few significant focuses to be noted, for example, the property of the information and the component of the dataset. In this trial, we take "Euclidean distance" and "Manhattan distance"- these distance estimation strategies for distance computations in the K-Means calculation [5][6]. Depiction about every procedure is referenced beneath:

3.1 Distance Measurements

To track down a highlight point distance among components and centroid, different distance measurements that assume a significant part in K-implies bunching are estimated to dole out these components to related groups (i.e.,centroids). Three distance measurements are carried out and examined as follows.

3.1.1 Euclidean Distance

Euclidean distance or Euclidean measurement is the natural and direct line between two components or the base distance between two articles [18], which is the most clear approach to addressing the distance between two focuses. On the off chance that focuses (x1, y1) and (x2, y2) are in 2-layered space, the Euclidean distance d between them is

$$d = \sqrt{(x2 - x1)^2 + (y2 - y1)^2} \quad (1)$$

3.1.2 Manhattan Distance

In the Manhattan distance capability [15], the distance between two focuses is the amount of the outright contrasts of their Cartesian directions. Basically it is the amount of the distinction between the x-directions and y-organizes. Consequently, the Manhattan distance d(x, y) can be characterized as

$$d(x, y) = \sum_{i=1}^{k} |xi - yi| \quad (2)$$

IV. EXPERIMENTAL RESULTS

The investigations have been coordinated by using Weka. The Weka is an open-source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problem. The unbalanced dataset used in this review was procured from the UCI data repository [8]. The dataset under study consists of 856 samples and 33 elements recorded. The statistical summary of the unbalanced dataset as shown in the figure-1.

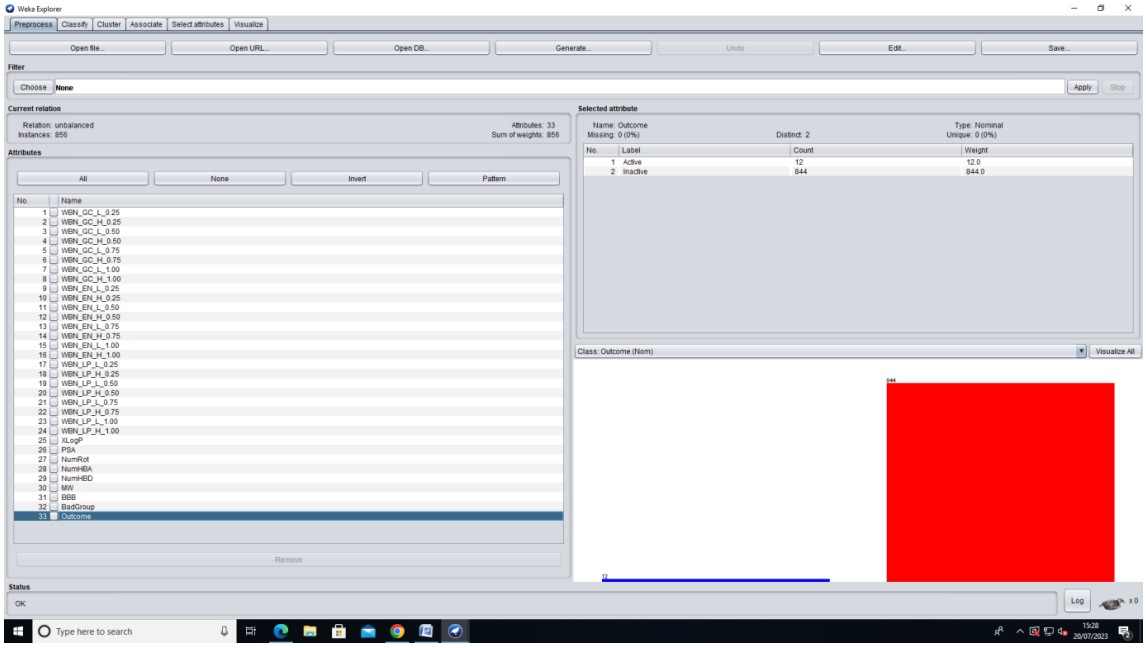

Figure-1: Statistical Summary of the dataset

4.1 Results

Upon analyzing the unbalanced dataset using K-Means clustering with both Euclidean and Manhattan distance metrics, the results demonstrate that Euclidean distance outperforms Manhattan distance in terms of clustering performance. The Euclidean distance metric achieved a higher clustering accuracy, better cluster separation, and improved clustering consistency compared to Manhattan distance. This indicates that Euclidean distance is more effective in capturing the underlying patterns and grouping data points accurately, even in the presence of class imbalances. The experimental results screen shots are shown from figure-2 to figure-3.

V. DISCUSSION

The findings of this comparative study reveal the significance of distance metrics in K-Means clustering, particularly when dealing with unbalanced data. Unbalanced datasets pose challenges for clustering algorithms, as they can lead to biased results and affect the quality of clusters formed. The superiority of Euclidean distance in this context can be attributed to its ability to calculate the straight-line distance between points, capturing the true similarity between data instances.

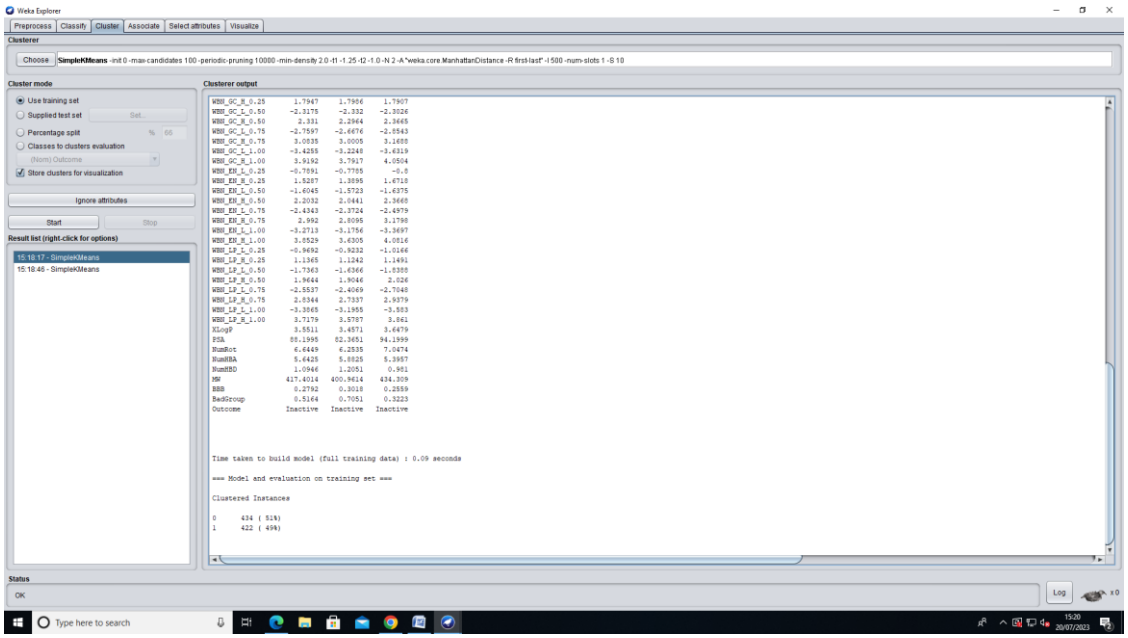

Figure-2: Experimental Results of Euclidean Distance

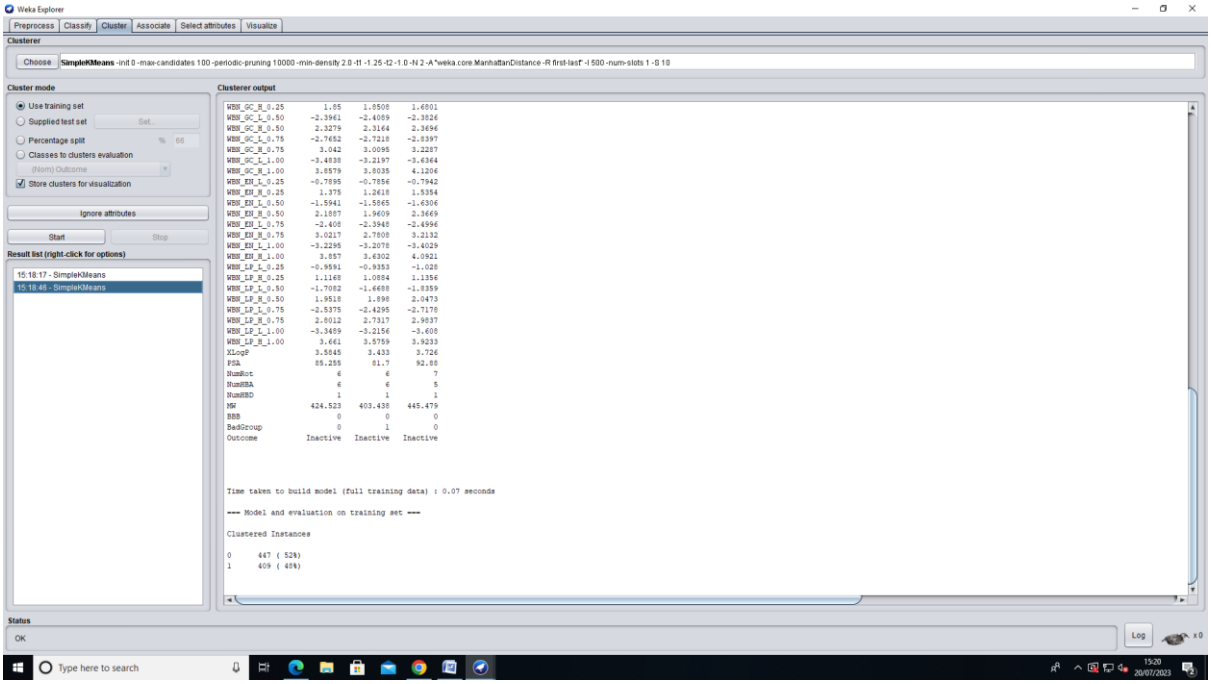

Figure-3: Experimental Results of Manhattan Distance

Manhattan distance, on the other hand, calculates the sum of absolute differences along each dimension, making it sensitive to outliers and potentially leading to less accurate cluster formation, especially in unbalanced datasets. This characteristic might contribute to its relatively poorer performance compared to Euclidean distance.

The choice of the distance metric plays a critical role in K-Means clustering and should be carefully considered based on the characteristics of the dataset and the specific objectives of the analysis. For unbalanced data, where minority classes are of particular interest, Euclidean distance emerges as the more suitable choice due to its robustness in handling skewed class distributions and ability to produce better-defined clusters.

VI. 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

[7] M. V. Lakshmaiah, Dr. G. Ravi Kumar and Dr. G. Pakardin, “Frame work for Finding Association Rules in Bid Data by using Hadoop Map/Reduce Tool”, International Journal of Advance and Innovative Research, Volume 2, Issue 1(I), PP:6-9, ISSN 2394 -7780, January-March 2015.

[8] UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets.