An Imaginary Study on Analyzing Results Using the K-Implies Grouping Technique

B Kishor Kumar

Department of Computer Science Sri Venkateswara University, Tirupati

Abstract— *One of the primary tasks in data mining is to group similar articles or data into clusters, which is particularly useful for analysis and prediction. The K-means clustering technique is a popular partition-based approach for clustering data, known for producing high-quality results. Clustering finds applications in various fields such as clustering online retailers, SMS and email spam filtering, human activity recognition, and more. Extensive research and applications exist in the field of clustering, with various algorithms available like K-means, K-medoids, among others. K-means stands out as one of the most widely used clustering algorithms. This paper focuses on the K-means clustering algorithm applied to large datasets and presents its implementation by analyzing supermarket data, which requires K or fewer passes through the dataset.*

I. INTRODUCTION

Data mining involves exploring data to uncover relationships and patterns that can offer valuable insights for informed decision-making. Various techniques like classification, association rules, and clustering are employed by organizations to enhance their decision-making abilities. It entails extracting meaningful and previously undiscovered information from large databases, enabling crucial business decisions. This extracted information can be utilized to create prediction models, classification models, or to discern relationships between database records [4].

II. CLUSTERING

Data clustering entails grouping data into distinct patterns or meaningful stacks, serving as a method for classification amidst vast information sets. The aim is to partition a dataset into multiple groups where the similarity within each group surpasses that between groups. The objective is to reveal the intrinsic structure of each cluster within the data. Clustering involves segregating a set of patterns into separate clusters, ensuring similarity within clusters and dissimilarity between them. It finds extensive application across domains such as document classification, data compression, music and movie categorization, user behavior-based classification, recommendation system construction, and pattern recognition. Clustering algorithms vary based on partitioning, density, and model, each aiming to group similar objects together. The K-means algorithm, a partition-based clustering approach, is notable for its simplicity, rapid convergence, and straightforward implementation. While effective in practical applications, direct application of the K-means algorithm can be computationally expensive, especially for large datasets, due to its time complexity proportional to the product of the number of patterns and clusters per iteration.[2,3]

III. K-MEANS CLUSTERING

The K-means algorithm is a simple iterative clustering algorithm. Using the distance as the metric and given the K classes in the data set, calculate the distance mean, giving the initial centroid, with each class described by the centroid. For a given data set X containing n multidimensional data points and the category K to be divided, the Euclidean distance is selected as the similarity index and the clustering targets minimize the sum of the squares of the various types; that is, it minimizes [4][7].

$$D = \sum_{k=1}^{K} \sum_{i=1}^{n} ||(x_i - u_k)||^2 \quad (1)$$

where k represents K cluster centers, u_k represents the kth center, and x_i represents the i^{th} point in the data set. The solution to the centroid u_k is as follows:

$$u_k = \frac{1}{n} \sum_{i=1}^{n} x_i \quad (2)$$

The central idea of algorithm implementation is to randomly extract K sample points from the sample set as the center of the initial cluster: Divide each sample point into the cluster represented by the nearest center point; then the center point of all sample points in each cluster is the center point of the cluster. Repeat the above steps until the center point of the cluster is unchanged or reaches the set number of iterations. The algorithm results change with the choice of the center point, resulting in an instability of the results. The determination of the central point depends on the choice of the K value, which is the focus of the algorithm; it directly affects the clustering results, such as the local optimality or global optimality [1][5].

K-means algorithm is one of the partitioning-based clustering algorithms. The general objective is to obtain the fixed number of partitions/clusters that minimize the sum of squared Euclidean distances between objects and cluster centroids.

Let X= {xi| i=1,2,……..,n} be a data set with n objects, k is the number of clusters, mj is the centroid of cluster cj where j=1,2,……..,k. Then the algorithm finds the distance between a data object and a centroid by using the following Euclidean distance formula [1].

where X represents is the first data point, Y is the second data point, N is the number of characteristics or attributes in data mining terminology. Starting from an initial distribution of cluster centers in data space, each object is assigned to the cluster with closest center, after which each center itself is updated as the center of mass of all objects belonging to that particular cluster. The procedure is repeated until convergence.

Algorithm:

K-means (D, K, C)

1. Arbitrarily choose k objects from D as the initial cluster centers.
2. Repeat
3. Reassign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster.
4. Update the cluster means, i.e., calculate the mean value of the objects for each cluster.
5. Until no change.

IV. EXPERIMENTAL RESULT

The implementation of proposed algorithm is using WEKA. We have evaluated our algorithm on Super Market data which was taken from UCI data repository [6], this Super market dataset which consists of 4627 records and 217 attributes of transactions. The statistical information for the entire dataset was shown in the figure-1

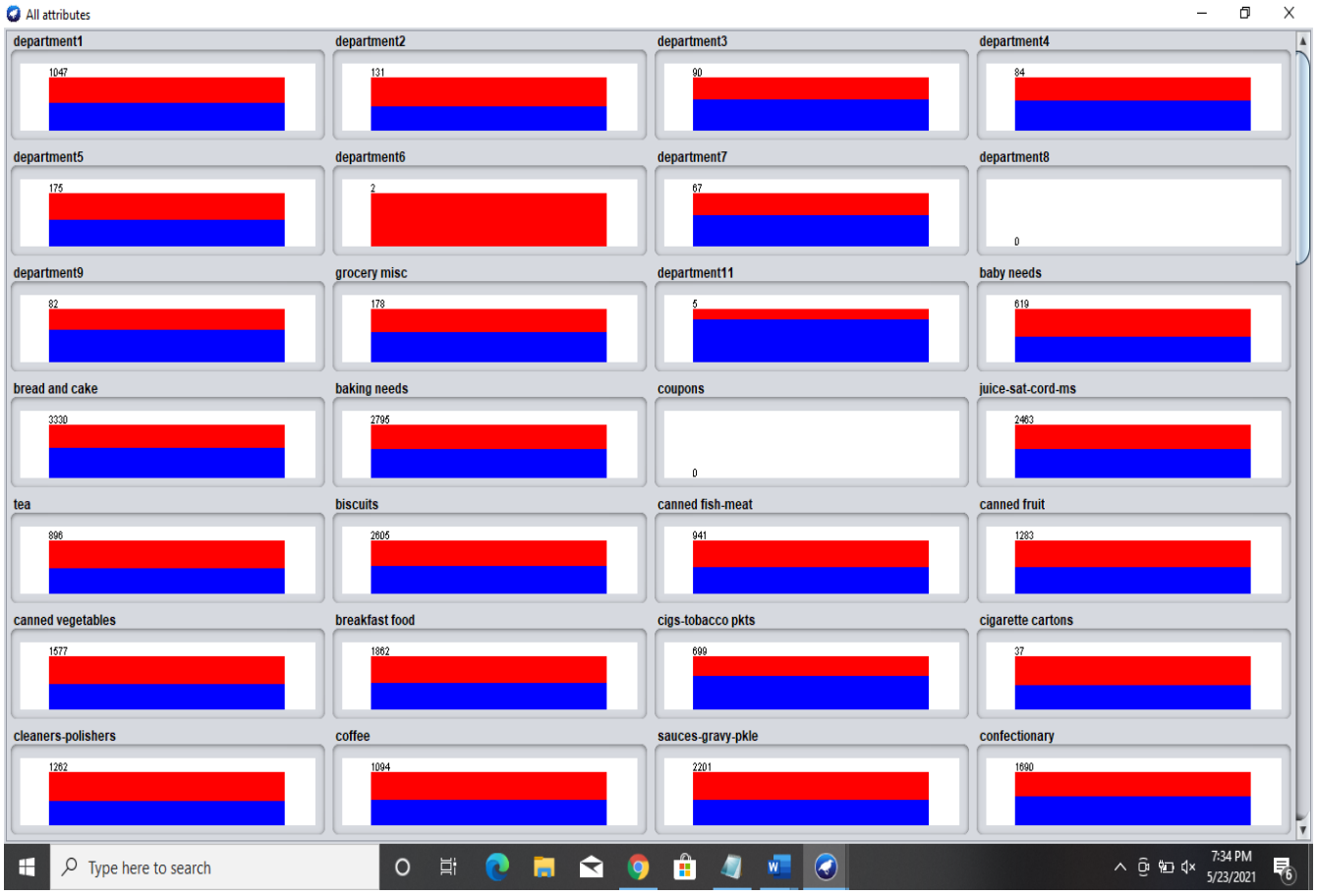

Figure-1: Statistical summary of the Dataset

The clustering results of the super market data are shown in the figures-2 and figure-3.

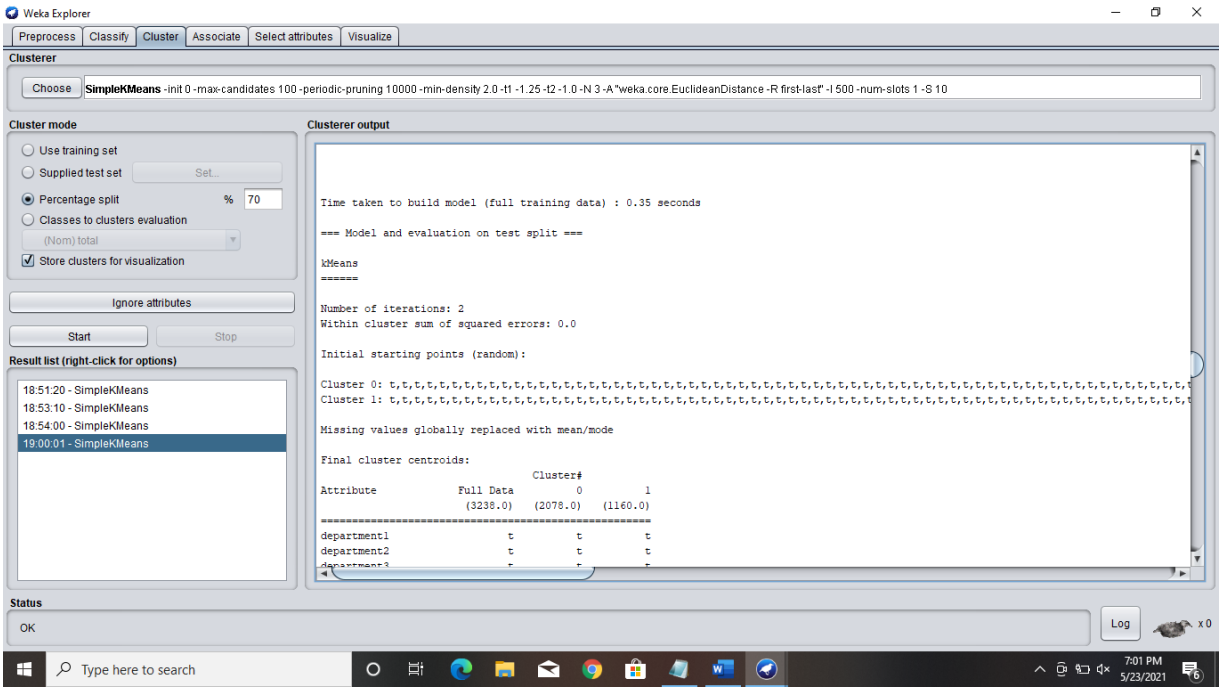

Figure-2: Figure-2: Results of K-means Algorithm

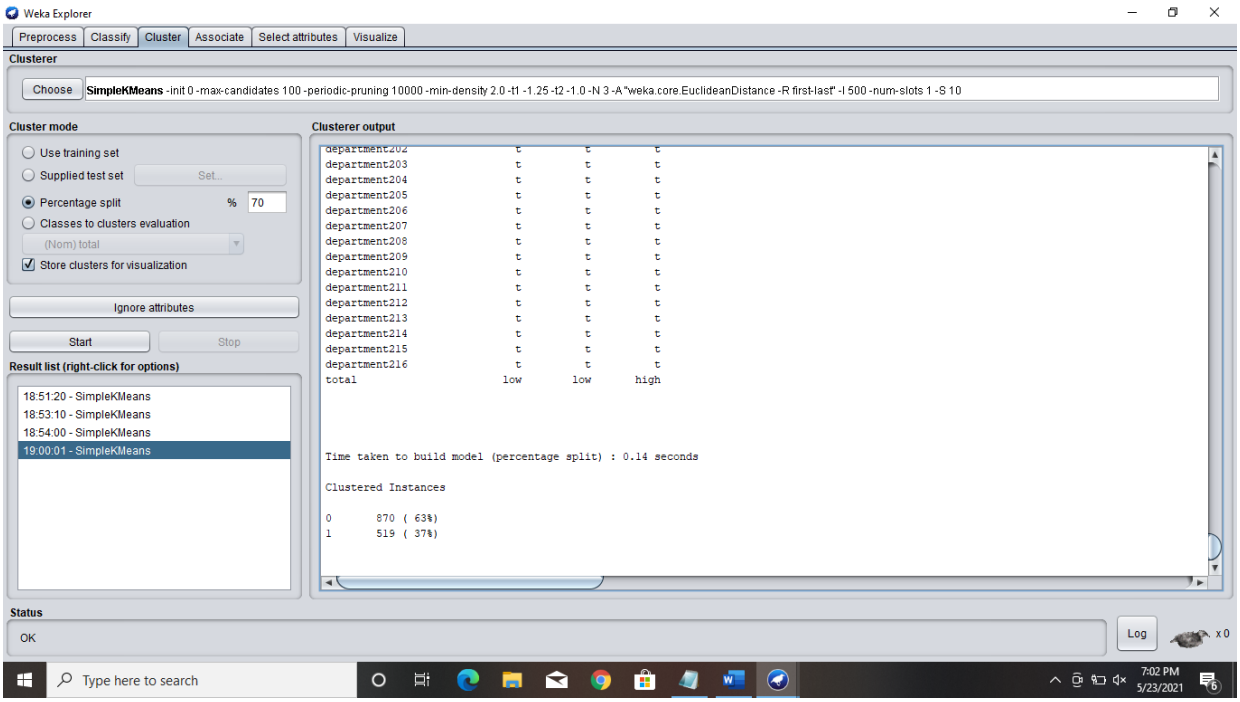

Figure-2: Results of K-means Algorithm

V. CONCLUSIONS

In this paper, we have examined a basic k-implies grouping calculation and breaks down the Super market dataset by utilizing the k-implies bunching calculation. In this paper we have utilized a subjective way to deal with break down the grouping calculation and we have utilized Euclidean distance for estimating the distance between objects. We showed the k-implies bunching calculation on a Super market dataset which comprises of 4627 records of exchanges. According to the hypothetical investigation and results got from the examination in this exploration, k-implies calculation conveys better proficiency for

bunching immense measures of information. A downside of k-implies bunching calculation is that it utilizes fixed number of groups. Setting proper beginning number of bunches is consistently a difficult undertaking.

REFERENCES

[1] H. Witten and E. Frank, “Data mining: practical machine learning tools and techniques with Java implementations”, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., (2006)
[2] H. Xiong; J. Wu; J. Chen, "K-Means Clustering Versus Validation Measures: A Data-Distribution Perspective, " IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol.39, no.2, pp.318, 331, April 2009.
[3] H. Xiuchang, SU Wei, “An Improved K-means Clustering Algorithm”, Journal of Networks, VOL. 9, NO. 1, January 2014.
[4] J. Han & Kamber, M. (2012). Data Mining: Concepts and Techniques. 3rd.ed. Boston: Morgan Kaufmann Publishers
[5] N. Michael, “Artificial Intelligence - A Guide to Intelligent Systems”, 2nd edition, Addison Wesley, 2005.
[6] UCI machine learning repository. http://archive.ics.uci.edu/ml/.
[7] W. Yintong; L. Wanlong; G. Rujia, "An improved k-means clustering algorithm, " World Automation Congress (WAC), 2012, vol., no., pp.1, 3, 24-28 June 2012.
[8] Yen-Chung Liu, Yen-Liang Chen,” Customer Clustering Based on customer Purchasing Sequence Data”, Int. Journal of Engineering Research and Application, ISSN: 2248-9622, Vol. 7, Issue 1, (Part -1) January 2017, pp.49-58.