A Hypothetical Investigation on Result Analysis of K-Implies Grouping Technique

Thirumuru Kavya

Department of Computer Science, Sri Venkateswara University, Tirupati

***Abstract*—** *One of the significant assignments of mining is to bunch comparative articles or comparable information into group which is especially valuable for examination and expectation. K-implies grouping technique is a mainstream partition-based methodology for bunching information as it prompts great nature of results. Grouping can be utilized in different applications like bunching on the web retailers, SMS, and email spam assortment, human movement acknowledgment and considerably more. There are colossal investigates and applications in the space of bunching. There are different bunching calculations are accessible like the k-implies, k-medoids and so on. The k-implies is one of the generally utilized calculations of grouping. This paper centers around K-implies bunching calculation on enormous datasets and present k-implies calculation by breaking down the Super Market information, which requires k or a lesser number of passes to a dataset.*

I. INTRODUCTION

Data mining is the search for relationships and patterns within this data that could provide useful knowledge for effective decision-making. Many different data mining techniques exist such as classification, association rules and clustering are used by different organizations to increase their capability for making decision.

Data mining is the process of extracting valid, previously unknown, and ultimately comprehensible information from large databases and using it to make crucial business decisions. The extracted information can be used to form a prediction or classification model, or to identify relations between database records [4].

II. CLUSTERING

Clustering of data is a method of grouping data into particular patterns or a method for classifying the information mountain into meaningful stacks. The goal of the clustering method is to divide a dataset into multiple groups so that the resemblance within a group is greater than between the groups. The objective of the clustering is to find out from the given data the underlying intrinsic structure of each cluster. [2]

Clustering is the process of partitioning or grouping a given set of patterns into disjoint clusters. This is done such that patterns in the same cluster are alike and patterns belonging to two different clusters are different. Clustering has been a widely studied problem in a variety of application domains. Therefore, the application fields are very broad, including different types of document classification, data compression and vector quantization, music, movies, classification based on user purchase behavior, the construction of recommendation systems based on user interests, pattern recognition and so on. Clustering algorithms can be divided into multiple types based on partitioning, density, and model [3]. A clustering algorithm is a process of dividing a physical or abstract object into a collection of similar objects. A cluster is a collection of data objects; objects in the same cluster are like each other and different from objects in other clusters [2]. For a clustering task, we want to get the objects as close as possible within the clusters: first cluster tends to sample or data point. However, the randomness of sample center point selection tends to make cluster aggregation not converge. Cluster analysis is based on the similarity in clustering data sets, which is unsupervised learning. In the partition-based clustering algorithm, K-means algorithm has many advantages such as simple mathematical ideas, fast convergence, and easy implementation [3]. The k-means method has been shown to be effective in producing good clustering results for many practical applications. However, a direct algorithm of k-means method requires time proportional to the product of number of patterns and number of clusters per iteration. This is computationally very expensive especially for large datasets.

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 k^{th} 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 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.

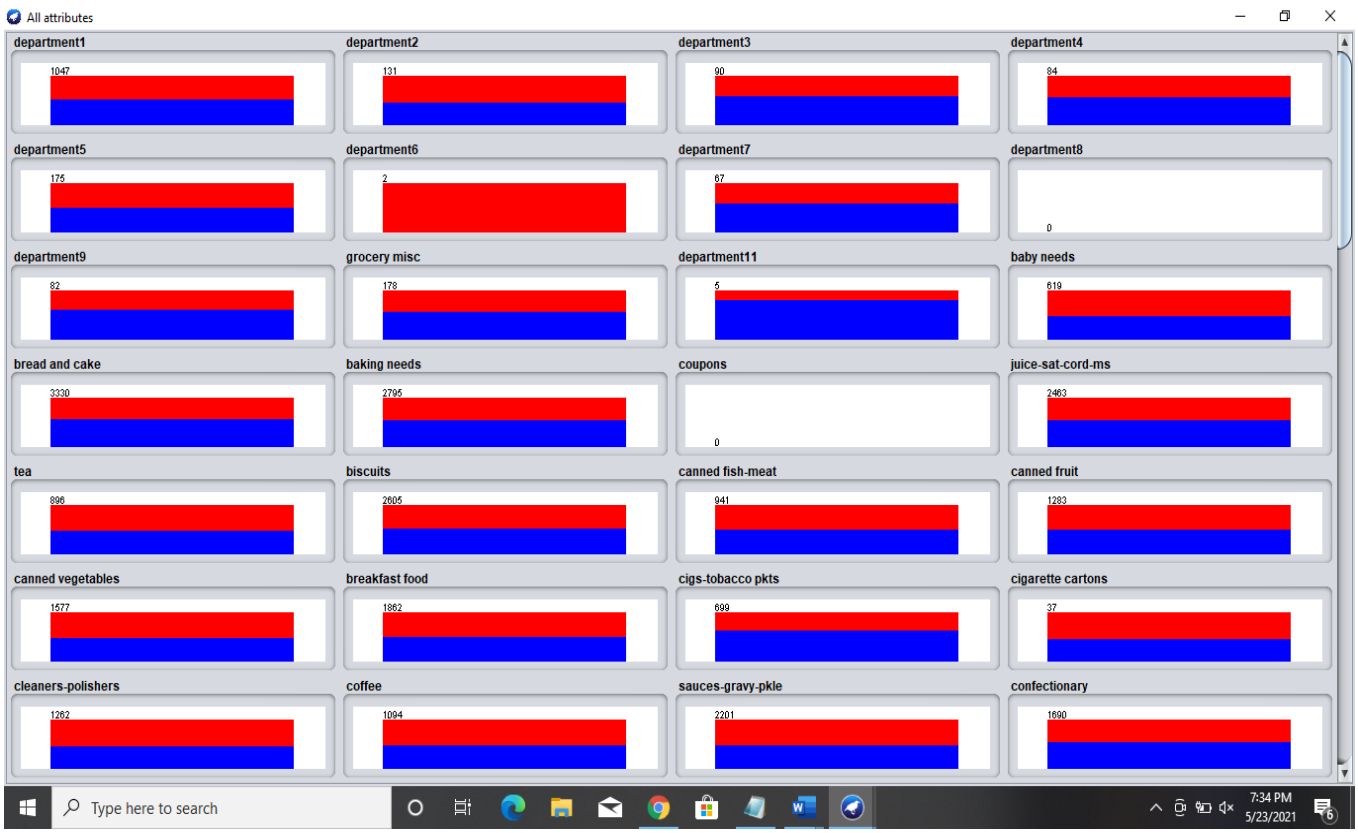

FIGURE 1: Statistical summary of the Dataset

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

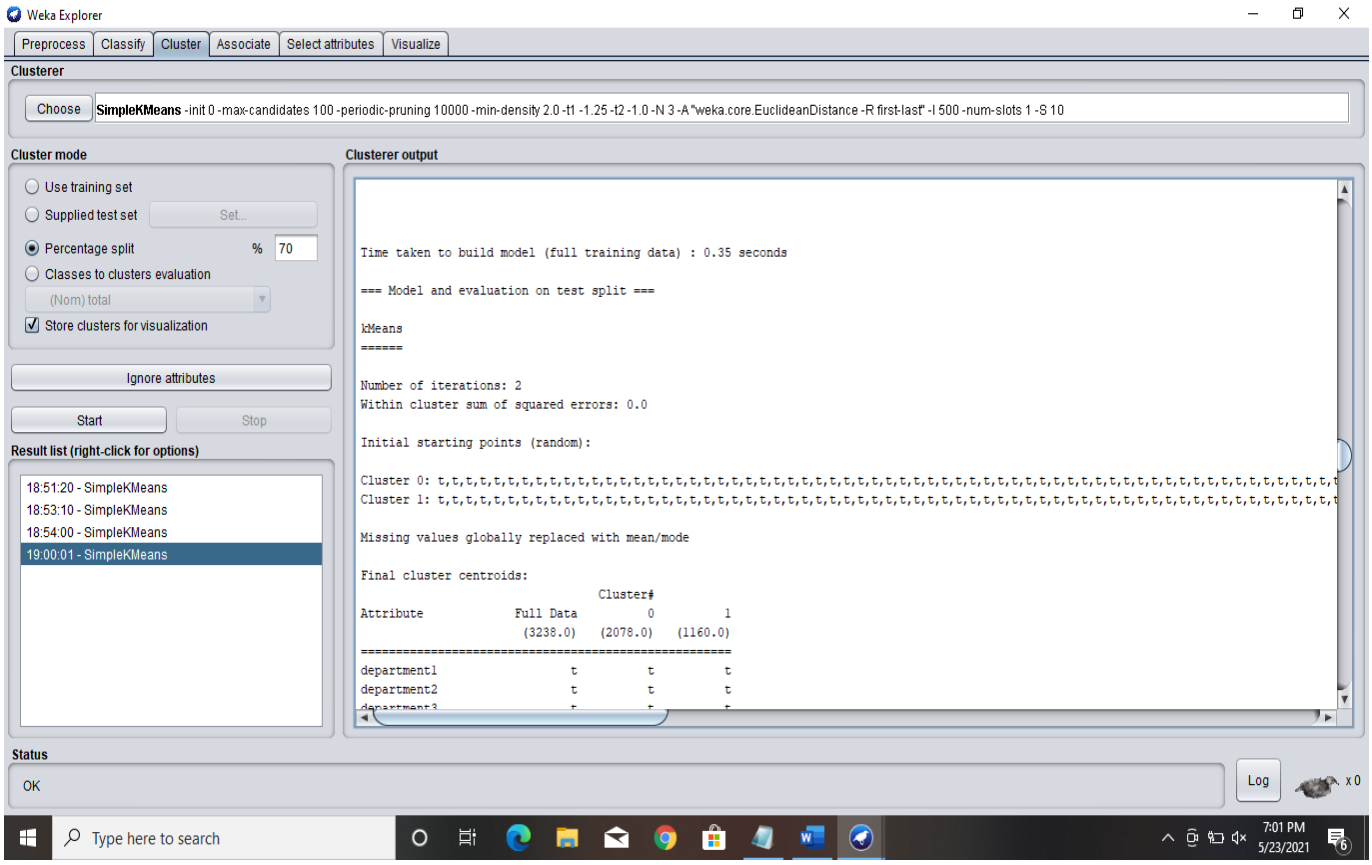

FIGURE 2: Results of K-means Algorithm

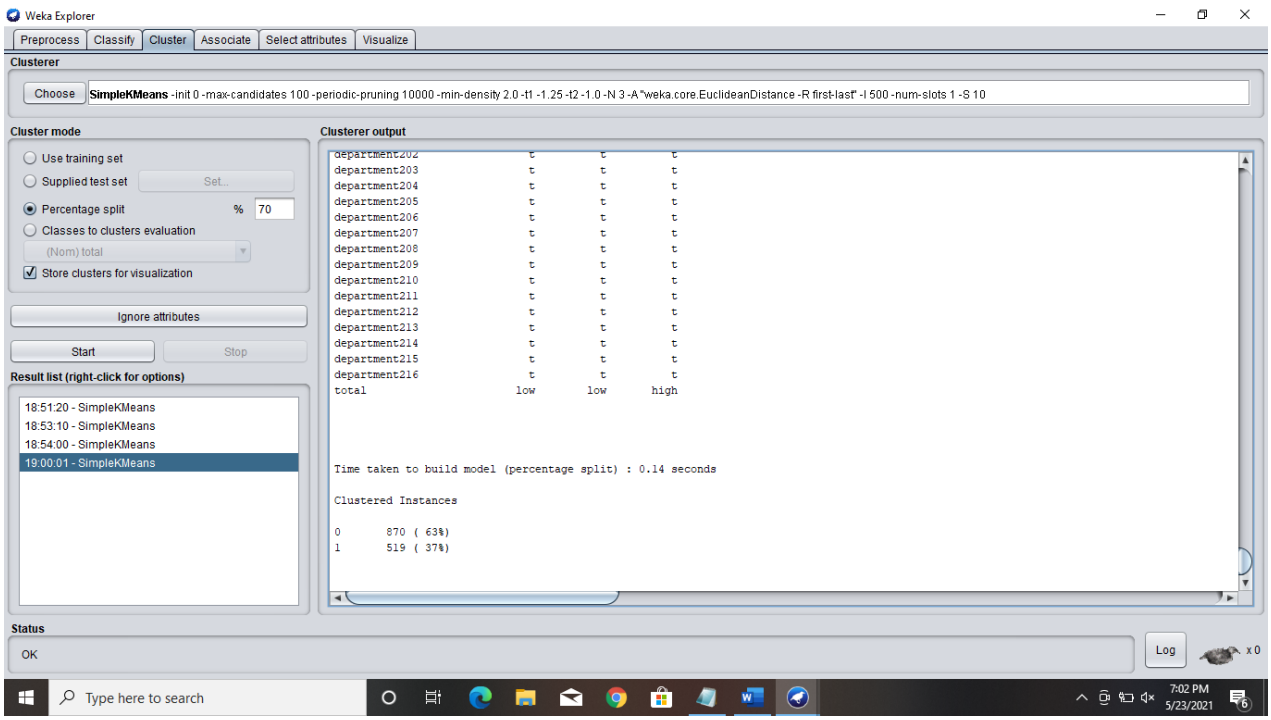

FIGURE 3: Results of K-means Algorithm

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.