IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CUSTOMER SEGMENTATION USING HIERARCHICAL AGGLOMERATIVE AND K-MEANS CLUSTERING ALGORITHMS

[1]K N Brahmaji Rao, [2]M Tirumala Prasad, [3]M Renee, [4]Katyayani Krishnan, [5]Jayasri

[1]Associate Professor, Raghu Institute Of Technology, Visakhapatnam, Ap, India.

[2,3,4,5]Student, Raghu Institute Of Technology, Visakhapatnam, Ap, India.

Abstract: In recent years, every shopping place focuses on customer relationship management(CRM) to provide better services to the customer as compared to their competitors. To maintain a better relationship with the customer and helps to increase profit and satisfaction. It is important for marketers to identify the potential customers in the market by getting the customer data to gain profitable insight. It is the best way to find the appropriate customer segmentation by using clustering algorithms. The possibility of a hybrid combination of clustering algorithms like k-means and agglomerative individual models has also been discussed.

Keywords: **Customer segmentation; Clustering; Agglomerative Hierarchical; K-Means; Machine learning Algorithms.**

I. INTRODUCTION

Data mining and analysis help to derive knowledge from historical data and thus form the context for predicting future outcomes. In data mining, agglomerative clustering algorithms are widely used due to their flexibility and conceptual simplicity. It works by grouping data objects into a cluster tree. The purpose of customer segmentation is to divide the user base into smaller groups that can be targeted with specialized content and offers

Customer segmentation allows the company to address specific customer groups most effectively. For decades, investors have relied on customer segmentation models built on basic demographic factors like age, income, education, gender, and more. In fact, they are poor predictors of consumer behaviour.

II. RELATED WORK

Key concepts such as CRM, customer segmentation, and the usability of customer segmentation are reviewed and discussed in this section. The significance of these concepts to businesses and organizations is also highlighted.

2.1 Customer Relationship Management

Customer relationship management is an important business method for developing and ensuring stable, long-term customer relationships. Modern marketing methods encourage the use of CRM as part of an organization's business strategy to improve customer service satisfaction [1],[2]. CRM allows business enterprises to analyze customer value as well as target higher-value customers. It also helps business organizations develop long-term, high-quality business relationships that increase loyalty and profitability. Accurately assessing customer profitability and targeting high-value customers are key success factors for CRM [3],[4],[5]. Since CRM is a customer-centric strategy, it is important for companies to know the characteristics and behaviors of their customer base. Insights into this customer data can then become useful when used in information technology (IT) solutions that deliver valuable results for better target customers. profitable customers [6],[7]. CRM plays an important role in targeting customers, once they are identified using essential segmentation strategies. According to [8] and [9], CRM strategy is a closed-loop structure with four dimensions: customer identification, customer acquisition, customer loyalty, and customer development. Thus, this structured customer identification clearly implies that grouping or segmenting customers based on their behavior and customer segmentation characteristics emerges as a central function of CRM.

2.2 Customer Segmentation

As the market expands, the speed of competition among business entities increases rapidly. Therefore, these business enterprises increase spending on their marketing strategies to gain a competitive advantage [10], [9]. Against this backdrop, the importance of using information technology (IT) solutions for marketing campaigns has emerged as an important step in a modern approach to business. Customer segmentation is a popular technique for dividing customers into internally and externally homogeneous groups to create diverse marketing strategies that target each group based on their characteristics. Broadly speaking, it is defined as the process by which the consumers of a business enterprise are divided into groups based on their preferences, characteristics, and purchasing behavior [5]. By researching and analyzing large volumes of customer data collected, companies can improve their marketing decisions based on customer preferences. Maximum profit can be generated for any business entity if resources are used wisely to nurture the most valuable and loyal customer group when segmentation and customer aggregation enabled the allocation of customers to these groups. A total number of customers can be divided and grouped into groups based on purchasing behavior, frequency, demographics, etc. So, instead of studying each customer individually, companies can group similar customers together to better understand their needs.

2.3 Use of Customer Segmentation

Targeting customer satisfaction and marketing analysis are correlated between the best fitting algorithms like k-means and agglomerative. First, customers in the selected market are segmented into different groups based on their characteristics It is important for marketers to identify the potential customers in the market by getting the customer data to gain profitable insight. The further section delivers customer segmentation by using clustering algorithms and the various machine learning algorithms.

III. METHODOLOGY AND IMPLEMENTATION

3.1 Clustering

Clustering is the process of dividing a population or data points into several groups specified the info points within the same group are more like than other data points within the same group and are different from the info points within the groups. It's basically a group of objects supported by their similarities and differences. differing kinds of clustering algorithms are available for efficiency.

3.1.1 K-Means Clustering

K-means clustering algorithm is one of the division-based clustering algorithms. It applies an iterative heuristic process to subdivide data objects and update cluster centers. the essential idea of the algorithm is, to assume a group with element objects and therefore the number of clusters to get [2]. within the first round, a sample item is randomly selected because of the initial cluster center [6], and therefore the distance between the opposite sample items and therefore the analyzed center point, the corresponding clusters are divided by distance. In each subsequent round, the iteration of the above steps is performed continuously, and also the average value of the element objects obtained at this point is taken because the center of the following round of clustering until the condition is reached. Events are at the guts of grouping. only longer changes within the iteration are satisfied. When we use K-Mean we've got to specify the number of clusters. To do that, we use the Elbow Method to find the optimal number of clusters.

3.1.1.1 The Elbow Method

Calculate the Within Cluster Sum of Squared Errors (WSS) for various values of k, and choose the k that WSS first starts to diminish. within the plot of WSS versus k, this is often visible as an elbow.

3.1.2 Hierarchical Clustering

Hierarchical clustering could be a method of cluster analysis that builds a hierarchy of information points as they get into a cluster or out of it [10]. Strategies for this algorithm generally constitute two categories:

3.1.2.1 Agglomerative - This clustering algorithm doesn't require us to specify the number of clusters prior to. Bottom-up algorithms treat each bit of knowledge as an initial singleton cluster, so successively assemble pairs of clusters until all clusters are merged into one cluster containing all the information. There are several options to merge consecutive data points:

- Minimal or singly linked clustering: compute all pairwise differences between elements in cluster 1 and elements in cluster 2 and consider the littlest of those differences because of the binding criterion. It tends to provide long "loose" beams.
- Mean or Mean Link Grouping: Calculate all pairwise differences between group 1 and group 2 elements and treat the mean of those differences because of the distance between the 2 groups. May vary within the compactness of the clusters it generates.
- Binding clustering centroid: calculates the difference between the middle of cluster 1 (one mean vector of length p, one element for every variable) and also the center of cluster 2.
- Ward's method of least variance: minimize total variance with inclusion. At each step, the pair of clusters with the tiniest distance between clusters is merged. Tends to supply more compact assemblies.

3.1.2.2 Divisive - Also called the top-down approach. This algorithm also doesn't require pre-specify the number of clusters. Top-down clustering requires a technique for splitting a cluster that contains the entire data and proceeds by splitting clusters recursively until individual data are split into singleton clusters. The divisive algorithm is additionally more accurate. Agglomerative clustering makes decisions by considering the local patterns or neighbor points without initially taking into consideration the worldwide distribution of knowledge. These early decisions can't be undone. whereas divisive clustering takes into consideration the worldwide distribution of information when making top-level partitioning decisions.

3.2 Dendrogram

A dendrogram could be a diagram representing a tree. This diagrammatic representation is usually employed in different contexts, in hierarchical clustering, it illustrates the arrangement of the clusters produced by the corresponding analyses.

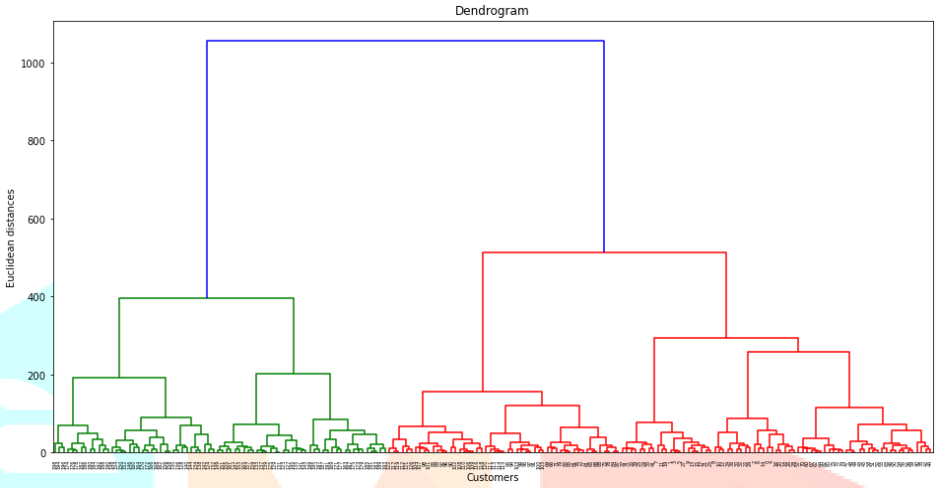

Fig 3.2.1 Dendrogram results of quantitative cluster analysis. A different color represent each cluster.

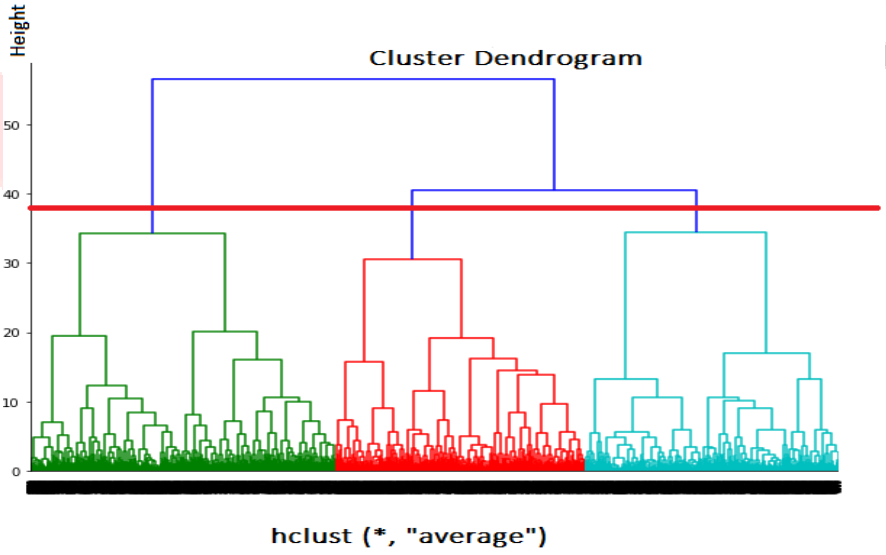

Fig 3.2.5 we cut the dendrogram at the height that will give us an optimal number of clusters say three, as shown in the figure.

3.3 K-Means clustering algorithm

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to its closest centroid, which will form the predefined

Step-4: Calculate the variance and place a new centroid in each cluster.

Step-5: Repeat the third steps, which means reassigning each data point to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.

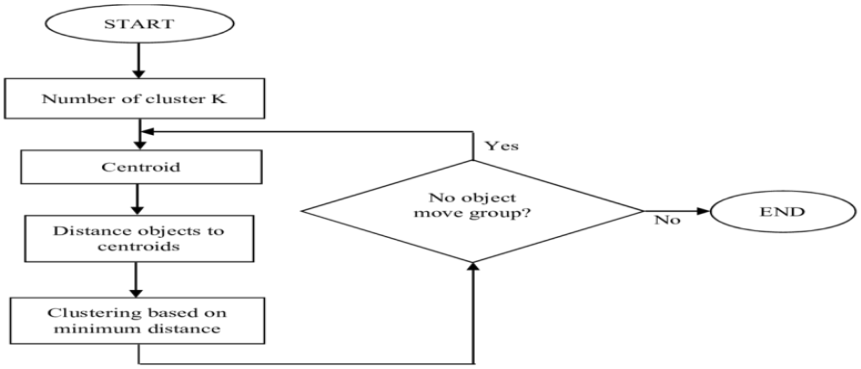

Fig 3.3.1 K-Means algorithm flowchart

3.4 Hierarchical clustering Algorithm

Step 1: Input measured features.

Step 2: Compute the distance matrix.

Step 3: Set each point as a cluster.

Step 4: If the number of clusters is equal to 1, then go to Step 8.

Step 5: Merge the closest clusters.

Step 6: Update the distance matrix.

Step 7: Repeat steps 4 to 5.

Step 8: Terminate.

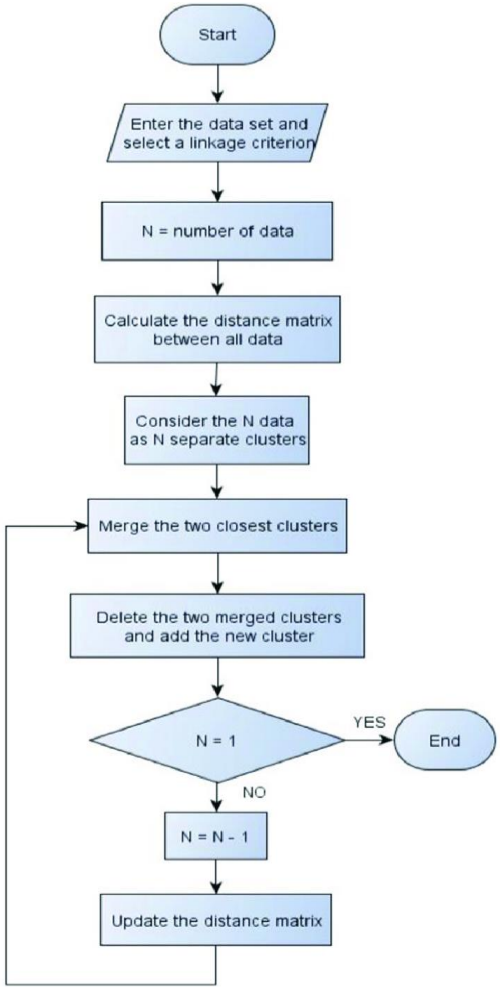

Fig.3.4.1 Flowchart for implementation of agglomerative hierarchical clustering

IV. RESULTS and DISCUSSIONS

4.1 Distance Matrix

The Hierarchical Clustering Distance Matrix is a matrix that contains the distances, taken pairwise, of a set of points. N is the number of points, nodes, or vertices.

4.2 Comparison Between K-Means & Hierarchical Clustering

It is necessary to compare the different clustering techniques discussed in order to identify which clustering algorithm should be used in which situation.

Each clustering algorithm has its own advantages also as a disadvantages in relation to the specific situation. K-Means is the most generally used clustering algorithm for customer segmentation. The K-Means require an initial number of clusters which is difficult to predict and might affect the clustering result. Hierarchical clustering doesn't require an initial number of cluster conditions.

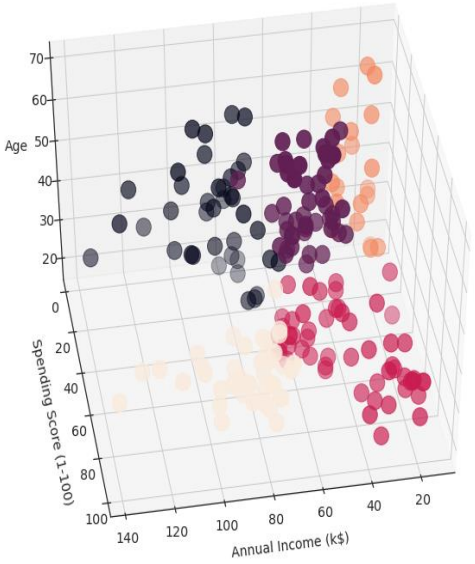

Fig 4.2.1 Clusters of customers using K-Means for small data

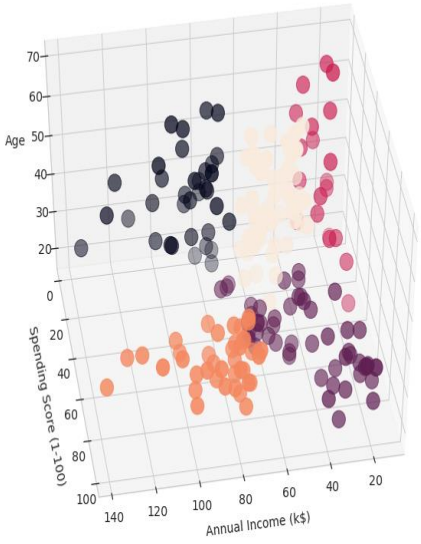

Fig 4.2.2 Clusters of customers using agglomerative clustering for small data.

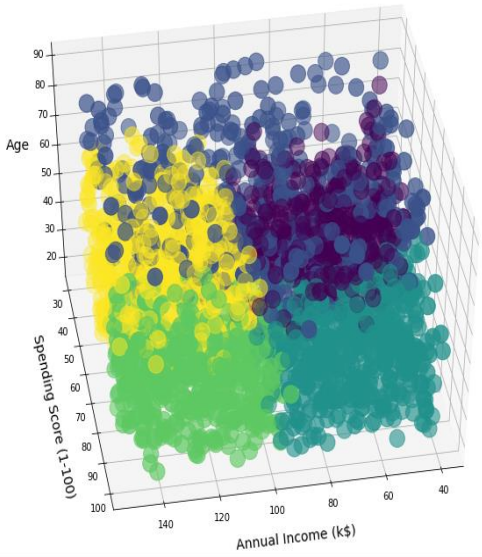

Fig 4.2.3 Clusters of customers using K-Means for large data.

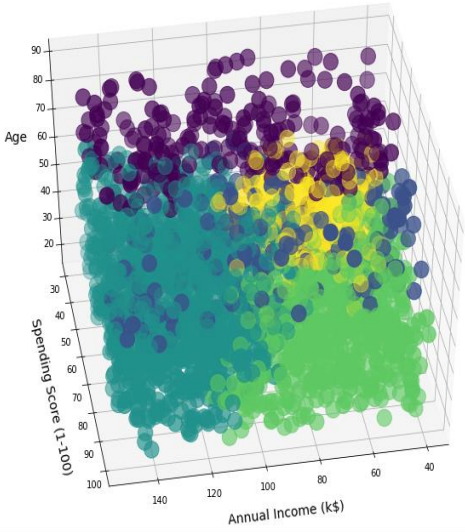

Fig 4.2.4 Clusters of customers using agglomerative clustering for large data.

Hierarchical clustering gives pretty similar results to K-means while handling small to medium data in fig 4.2.2 and 4.2.3. Here we've got used the ‘average’ linkage because of the proximity metric. the sole disadvantage in agglomerative clustering is that run time would be more compared to k-means and also the selection of using ‘num clusters’ would be pretty intuitive as hostile k-means which uses the elbow method.

4.3. The Validation of Hierarchical Clustering and K-means Clustering

The evaluation of the clustering outputs is extremely important for info modeling. The validation may be supported either by external criteria (evaluate the results with regard to a pre-specified structure), or internal criteria (evaluate the results with regard to information associated with the information alone). There are different measures for the inner validation of clustering.

Algorithm	Silhouette The coefficient for small to medium data set	Silhouette The coefficient for large data set
Hierarchal Clustering	0.417	0.217
K-means Clustering	0.410	0.279

Table 4.3.1 validation metrics for the Hierarchical and K-means clustering

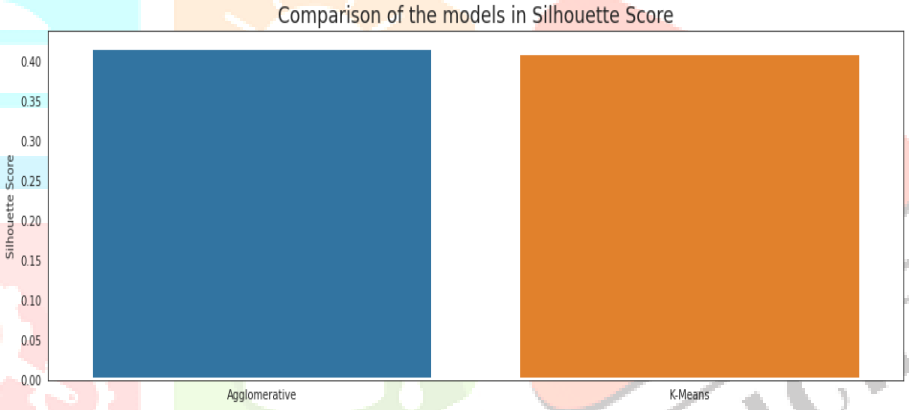

Fig 4.3.1 Silhouette Scores of K-Means and Agglomerative models for small to medium data set.

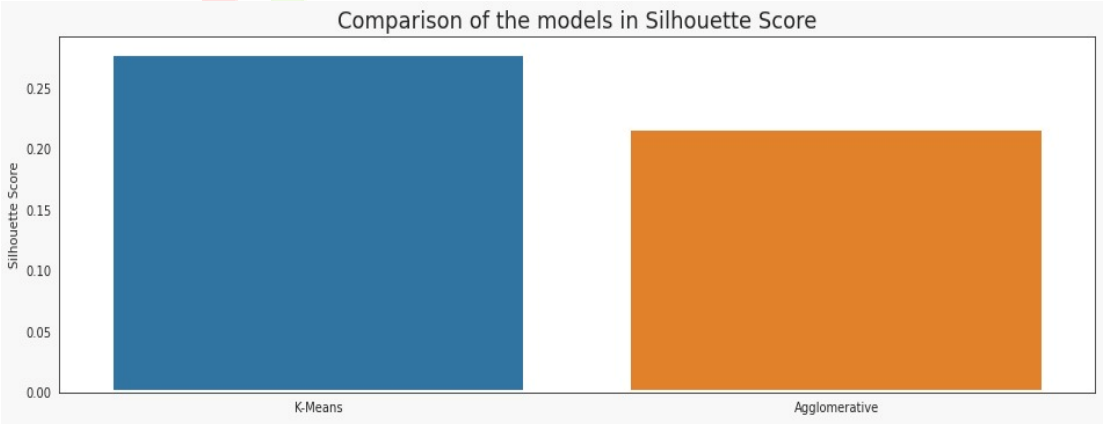

Fig 4.3.2 Silhouette Scores of K-Means and Agglomerative models for large data set.

From the above Fig 4.3.1, Fig 4.3.2, and Table 4.3.1, it is clear that the silhouette scores of both the algorithms are much similar and they have very small differences when the data size is small to medium. This proves that the agglomerative clustering algorithm can work as efficiently as K-means for customer segmentation in small to medium-scale shopping malls, businesses, marketplaces, etc. On the other hand, K-means clustering has to be considered most effective than agglomerative clustering for customer segmentation in large shopping malls and e-commerce sites as K-means can handle large and dynamic data more efficiently than agglomerative clustering.

4.4 Drawbacks of K-mean clustering for small to medium datasets

- Hard to predict K-Value
- Being dependent on initial values.
- Clustering outliers.
- Scaling with the number of dimensions.

V. CONCLUSION

The strong trends of service marketing in recent years have presented data analysts with new tasks and challenges, requiring more sophisticated and advanced analytical methods, which also yield results. significant results in future strategic planning.

This study uses a hierarchical clustering algorithm on small shopping data set and a k-means algorithm on large shopping data set to perform customer segmentation. Based on the results obtained, analysts can test more profitable tailored marketing strategies.

The proposed system includes a set of methods for handling all steps, from data pre-processing to results in visualization. The downside of this method is that it is quite slow and hardware dependent. Therefore, we recommend that when we use a very large dataset it is better to prefer a cloud environment.

5.1 Scope of Future Research

It was also discovered that in addition to discovering all available information about the data, it was necessary to try several different clustering algorithms. The different properties of the data are perhaps best exploited by tailoring them to different types of clustering. In the case of data used for this project, the K-means team seems to indicate the best fit for large data and agglomerative for small to medium data.

However, we can also study more complex models for future work. Possible advanced clustering methods include Fuzzy C-means clustering, density-based clustering, and distribution pattern-based clustering. Further exploration of K-mean association and hierarchical clustering can be applied with this additional work. A Hybrid model can be developed with the combination of both K-means and Agglomerative clustering algorithms which can work effectively for all types of data sets.

REFERENCES

[1] Masciari E, Mazzeo GM, Zaniolo C. “A new, fast and accurate algorithm for hierarchical clustering on Euclidean distances” Springer-Verlag Berlin Heidelberg 2013, 111-114, LNAI 7819.

[2] Karuna Ghai, Jaspreet Singh. Clustering Algorithms in Gene Expression: Data Analysis. 2021,,, 1-4. https://doi.org/10.1109/ICRITO51393.2021.9596549

[3] Ge Wang, Pengbo Pu, Tingyan Shen. An efficient gene bigdata analysis using machine learning algorithms. Multimedia Tools and Applications 2020, 79 (15-16) , 9847-9870. https://doi.org/10.1007/s11042-019-08358-7

[4] Zhang C, Zhang H, Wang J (2018) Personalized restaurant recommendation method combining group correlations and customer preferences. Inf Sci 454–455:128–143. https://doi.org/10.1016/j.ins.2018.04.061

[5] P. Brito, et al., “Customer segmentation in a large database of an online customized fashion business”, Robotics and Computer Integrated Manufacturing, vol. 36, pp. 93-100, 2015.

[6] P.Badase, G. Deshbhratar and A. Bhagat, “Classification and analysis of clustering algorithms for large datasets”, in International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, 2015, pp. 1- 5.

[7] C. Xiong, et al., “An Improved K-means Text Clustering Algorithm by Optimizing Initial Cluster Centres”, in 7th International Conference on Cloud Computing and Big Data

(CCBD), Macau, 2016, pp. 265-268.

[8] O. A. Abbas, “Comparisons between data clustering algorithms”, International Arab Journal of Information Technology, vol. 5, no. 3, pp. 320-325, 2008.J. Lee and S. Park, “Intelligent profitable customers segmentation system based on business intelligence tools”, Expert

Systems with Applications, vol. 29, no. 1, pp. 145-152, 2005.

[9] D. A. Kandeil, A. A. Saad and S. M. Youssef, "A two-phase clustering analysis for B2B customer segmentation", in International Conference on Intelligent Networking and Collaborative Systems, Salerno, 2014, pp. 221-228.

[10] R. Swift, Accelerating Customer Relationships: Using CRM and Relationship Technologies, 1st ed. Upper Saddle River, N.J.: Prentice Hall PTR, 2000.