ISSN: 0974-5823 Vol. 7 No. 2 February, 2022

International Journal of Mechanical Engineering

An Enhanced Hierarchical Model for Customer Segmentation in Customer Relationship Management with Demographic, Recency, Frequency and Monetary Values

Dr.B.Arivazhagan[1], Dr.G.Vijaiprabhu[2]

[1]*Assistant Professor, Department of Computer Science, Erode Arts and Science College, Erode-638009, Tamilnadu, India*

[2]*Assistant Professor, Department of Computer Science, Erode Arts and Science College, Erode-638009, Tamilnadu, India.*

ABSTRACT – The relationship between customers and business is often correlated and complex to understand. The relationship is based on buying and selling of products or making use of services provided. This communication leads to large dataset and it cannot be understand in a single view. The dataset is analyzed in certain point of view a variety of demographic, psychographic characteristics and multitude of buying behaviors, risk patterns and level of profitability with the variant customers might be known. Data mining ease this process with its numerous techniques to reveal the hidden interesting pattern among customers. Segmentation is the act of splitting a database into distinct sections or groups with two approaches market-driven and data-driven. Segmentation use data mining cluster analysis in order to find homogeneous groups of customer and they are mutually exclusive in which no one can be a member of other segment. One of the most common profiling methods to segment the customer is to use Recency, Frequency and Monetary (RFM) values. These three values or characteristics can be used alone or in combination with other characteristics to assist Customer Relationship management (CRM) efforts. This work combines these values with demographic data (DRFM) to segment the customers using enhanced Hierarchical Agglomerative clustering. The enhanced method is analyzed with two variant sectors namely Banking and Telecom as these sectors are highly influenced by valuable customers. The work is implemented in the data mining tool Rapid Miner with appropriate measures.

Keywords: *Hierarchical Agglomerative clustering, Flatten Clustering, Quick sort, Euclidean distance, Inner product Similarity.*

1. Introduction

Customer profile analysis [1] is an excellent method to know about customer value to reach the benefit of improved target and product development. The use of targeting models is very common in marketing and it starts with defining the goal. This model generally falls into two categories such as predictive model that represent future activity based on target values like balance amount and descriptive model which creates rules to group customers into descriptive categories. Customer relationship management forms a rule in the form of procedures, goals, and guidelines that an organization follows to improve their customer base. The entire relationship includes direct sales and service-related processes, analysis of customer trends and behaviors and predicting loyal customer and high profitable customer. Customer relationship management (CRM) [2] is a routine process related with a company to handle its service with their customers. It focuses on improving customer service, business relationships, profitability, keeping retention and increasing sales growth. Customer Satisfaction is very vital in business sector as it has the huge impact in economic growth. It has the ability to increase customer loyalty and usage behavior by having timely, accurate processing of customer orders and requests with ongoing management of customer accounts. This process can be done with [3] customer segmentation.

A. Data Mining in CRM

A Good Customer [4] Relationship Management (CRM) gives single image to all the customers and makes way to interact with the enterpriser at any time. The CRM should anticipate the customer needs and address them proactively. In such case Data Mining plays an efficient role in every facet of CRM. With the applications of data mining, the enterpriser turn their numerous records of customer database into a sort of coherent picture of customers. Data Mining presents a good decision support system with informed decisions. It needs complex procedures to have analysis in vast quantities of data. Now-a-days, the incredible advances in computing power, accepting big data, giving output in milliseconds made data mining to have analysis on large scale.

B. Process in data mining with CRM

Fig 1, depicts CRM with [5] data mining techniques in order to have analysis regarding customers.

The six foremost activities in data mining are:

1. *Classification:* It examines the features of a presented object and assigns it to a predefined class. So, it generally comprise with a class code. It updates each record by filling with class.

Example: Classification of credit card holders as low, medium and high risk, assigning patients with having and not having disease.

2. *Estimation:* Estimation is like classification but it deals with continuous valued attribute by handling dataset with income, height, weight. This technique is used to predict the response of the customer.

   Example: Estimating the family's household income, estimating the value of the price of the real estate.

3. *Prediction:* All prediction is termed as classification or estimation. The difference is predictive tasks will predict future behavior or estimated future value based on the previous results.

   Example: Predicting the customer whether he will leave from the organization in the next following periods, predicting which subscribers will order a new service.

4. *Association rule*s: It determines which things go together. Retail chains can use this model to derive rule for getting knowledge about the items often purchased together.

   Example: In CRM, this rule is used to identify cross-selling opportunities, to design attractive packages, to frame products with services to attract customers.

5. *Clustering*: Segment the [6] diverse groups into a number of similar groups. It does not depend on pre-defined class like classification. The records are grouped together based on similarity.

   Example: Group the customer based on their buying habits, grouping the customers based on their credit score.

6. *Description and Visualization:* Human can understand thousand times better through visualization than simple description. Data mining in its final step provides description and visualization to explain what is going on in a complicated database.

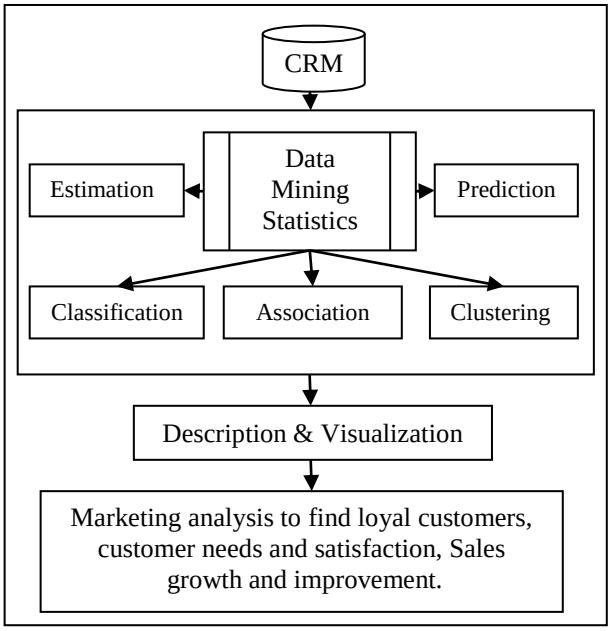

Fig 1. CRM with Data mining

CRM use data mining techniques to focus on:

- Targeting profitable customers.
- Connect customers from all channels to a single platform.
- Improving customer service and efficiency.
- Make customized products and services.
- Trace Customer satisfaction with products.

2. Literature Review

Anika Singh *et al* [7] proposed a model to segment customers with K-Means and Hierarchical clustering techniques. Market data is taken for analysis to know the expectation of customers and group the customers based on their preferences. For clustering, distance metric is the key component as each metric gives different clusters. Data are standardized within scale in prior. K-Means require number of clusters as its primary input while hierarchical doesn't need. The hierarchical model in this paper is cut into clusters after the dendrogram is produced. With the elbow method the number of cluster is selected as three for hierarchical method. Both clustering uses Euclidean distance as the metric. But K-Means produce different clusters for each time as it fully depends on the centroid.

Shreya Tripathi *et al* [8] explored the importance of customer segmentation in CRM through Centroid based clustering and Hierarchical based clustering. Target marketing is completely depends on grouping customers based on their purchasing behaviors

and needs. In Clustering, data points are clubbed together in which same cluster has more similarities. They are internally homogeneous and externally heterogeneous. Centroid based clustering K-Means (with five clusters) and Hierarchical clustering use Euclidean distance to form clusters. Number of clusters is not prefixed in Hierarchical and it is produced huge number of clusters which is not able to visualize. Hence a cut-off line is used in the dendrogram for five clusters to make visualization better. The quality of hierarchical clustering improves when compared to K-Means clustering with the increased number of K Values.

Phan Duy Hung *et al* [9] implemented hierarchical clustering algorithm for the segmentation of customers on credit card data sets to determine the appropriate marketing strategies. The purpose of customer segmentation is to divide the user on to specific group to target and to provide offers. The credit card dataset summarizes the usage behavior of active credit card holders within 6 months and it includes 18 features where credit limits, purchase and payments were used for segmentation. Initially, missing values and outliers are handled to get a structured dataset. Agglomerative hierarchical clustering is performed to get sub groups of customers. To cut the dendrogram in to meaningful clusters, three types of methods namely Elbow method, Silhouette method and Gap Statistic method is used and based on this method three clusters are preferred. These three clusters defines the variant customers based on the selected attributes.

Fahed Yoseph *et al* [10] applies soft clustering fuzzy C-Means (FCM) and hard clustering Expectation Maximization (EM) algorithms to group consumers based on similar purchase. This paper concentrates to maximize consumer Lifetime Value (LTV) through RFM. Retail Dataset is taken for analysis. RFM segmentation allows retailers and marketers to target specific groups of consumers. FCM combines partitioning and hierarchical clustering procedures while EM has probability distribution over the features. The work is done in four phases starts from preprocessing, then changing the dataset to RFM to have LTV. Clustering is applied in third phase and in final phase accuracy is calculated. The results shows EM has high inter cluster distance than FCM.

Sabitha *et al* [11] build LRFM customer relationship model to calculate customer's loyalty in retail marketing. Ta-Feng dataset which is a super market dataset is used for analysis. LRFM model gives input attributes to yield quantitative value for K-means clustering. K-Means is used to cluster customers to identify high profile customers in order to promote marketing strategies. For improving the efficiency of the clustering algorithm, initial cluster centroids values are modified and assigned using Random, Hartigan and Binary search methods. Finally the validity of clustering is analyzed with R-Squared index. From the analysis it is noted that binary search centroid initialization outperforms other initialization method in K-Means clustering.

3. Methodology

Linkage based Hierarchical Clustering (HAC)

It is one of the method in cluster analysis that produces [12] hierarchy of clusters with the objects. There are two types namely Agglomerative and Divisive.

1. Agglomerative: A Bottom –Up approach in which initially each observation begins from its own cluster and merged based on distance function iteratively until all observation belongs to same cluster.

2. Divisive: A Top-Down approach begins with one cluster and split into number of small cluster down the hierarchy.

The results of hierarchical clustering are usually viewed in the form dendrogram. It is a representation of clusters in tree format. [13] Generally hierarchical clustering has the time complexity of $O(n^3)$ and memory of $\Omega(n^2)$. So it has a poor running time for even medium dataset. This work concentrates on Agglomerative clustering.

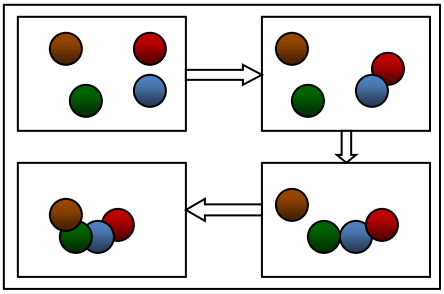

Fig 2. Level of Hierarchical cluster in Agglomerative

Fig 2, shows hierarchical agglomerative in which the objects are in distinct cluster initially, then it gets merged based on distance function and finally all the objects are merged to get single cluster.

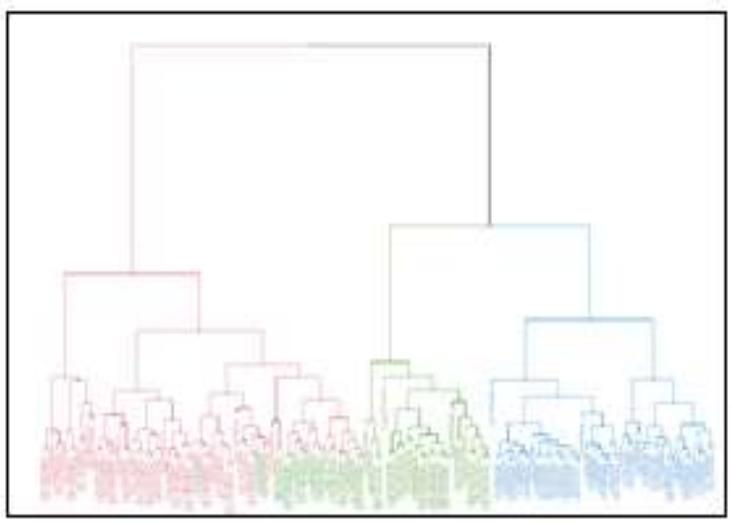

Fig 3. Dendrogram of Agglomerative Hierarchical clustering

The dendrogram of hierarchical agglomerative clustering is shown in Fig 3. There are so many clusters formed since the hierarchical clustering doesn't need the number of clusters to be prefixed.

B. Linkage Criteria

It determines the [14] distance between sets of observations. Three types of linkage methods are in agglomerative clustering are Single linkage, Complete Linkage and Average Linkage.

1. Single Link – Two clusters are merged if the minimum distance between the points in the clusters are less than or equal to the threshold distance.

The Linkage between two observations (B, C) is calculated as follows:

$$\text{Single Linkage} = \min\{d(b,c)\} \quad (1)$$

2. Complete Link - Two clusters are merged if the maximum distance between the points in the clusters are less than or equal to the threshold distance.

$$\text{Complete Linkage} = \max\{d(b,c)\} \quad (2)$$

3. Average Link - Two clusters are merged if the average distance between the points in the clusters is below the threshold distance.

$$\text{Average Linkage} = \frac{1}{|B|.|C|}\sum_{b\in B}\sum_{c\in C} d(b,c) \quad (3)$$

Where,

'b, c' are the objects in the cluster B,C and d (b, c) is the distance between the objects (b, c).

C. Distance function in Hierarchical clustering - Euclidean Distance (HAC-ED)

The existing work use Euclidean distance function to merge similar clusters.

$$\text{Euclidean distance} = \text{SQRT}\left(\sum_{i=1}^{n}(a_i - b_i)^2\right) \quad (4)$$

1. Steps in Agglomerative Clustering

1. Proximity of individual points is [15] calculated and considers all the data points as individual clusters.
2. Based on the Euclidean distance function similar clusters are merged together on the basis of single or complete or average linkage to form a single cluster.
3. All the clusters are merged together repeatedly and form a single cluster and in each iteration proximity is calculated.

2. Ways to visualize optimal clusters from Dendrogram

- Using Elbow method, Average Silhoutte method and Gap Statistic method.
- Using Thumb rule to decide number of clusters from dendrogram.
- Using cut-off line on seeing the maximum distance or height of dendrogram.

D. Drawbacks in existing customer segmentation and Agglomerative Clustering

- For segmentation, only RFM values are considered to group the data for customer segmentation. Recency refers to the recent transaction, Frequency refers to the number of transaction and Monetary refers to total amount of transaction.

- Agglomerative Hierarchical clustering takes high processing time.
- Euclidean distance is not suitable for all types of customer sector and it sensitive to outliers and takes high time for calculation.
- Hierarchical clustering doesn't require number of clusters to be prefixed and produces large number of clusters for the dataset with more distinct entries.
- Evaluation metric is not directly used to qualify the clusters as Hierarchical clustering produce numerous clusters.

Proposed work and method of Agglomerative Clustering

A. Hierarchical Agglomerative Flatten Clustering with Quick sort and Inner product similarity (HAFC-QI)

The proposed work use DFRM (Demographic, Frequency, Recency, Monetary) Values for customer segmentation. Other than basic information customer's location, age, marital status, job, education are also considered for the analysis of customers in order to observe their behavioral regularities and to develop customer strategies and improve their campaign response speed. This additional information leads to better strategy to maintain customer relation and further develop personalized marketing strategy. These four values are the key components to calculate Life Time Value (LTV) of a customer.

In Agglomerative Hierarchical clustering, the proposed work uses Inner product space similarity to form clusters instead of Euclidean distance. Also Quick Sort is used to minimize the Time Complexity of Hierarchical Clustering. Finally Flat clustering model is combined to visualize meaningful and compact clusters based on the distance function. Two types of evaluation metric Sum of Squares and processing time are used to assess the performance of proposed Hierarchical with Flat Clustering model.

B. Proposed Flow of Work

The proposed flow of work with DFRM values for customer segmentation is shown on Fig.4. The work is implemented with quicksort and Inner product space similarity function while the existing is implemented with euclidean distance function. Three clusters are formed and the performance is validated for existing and proposed one.

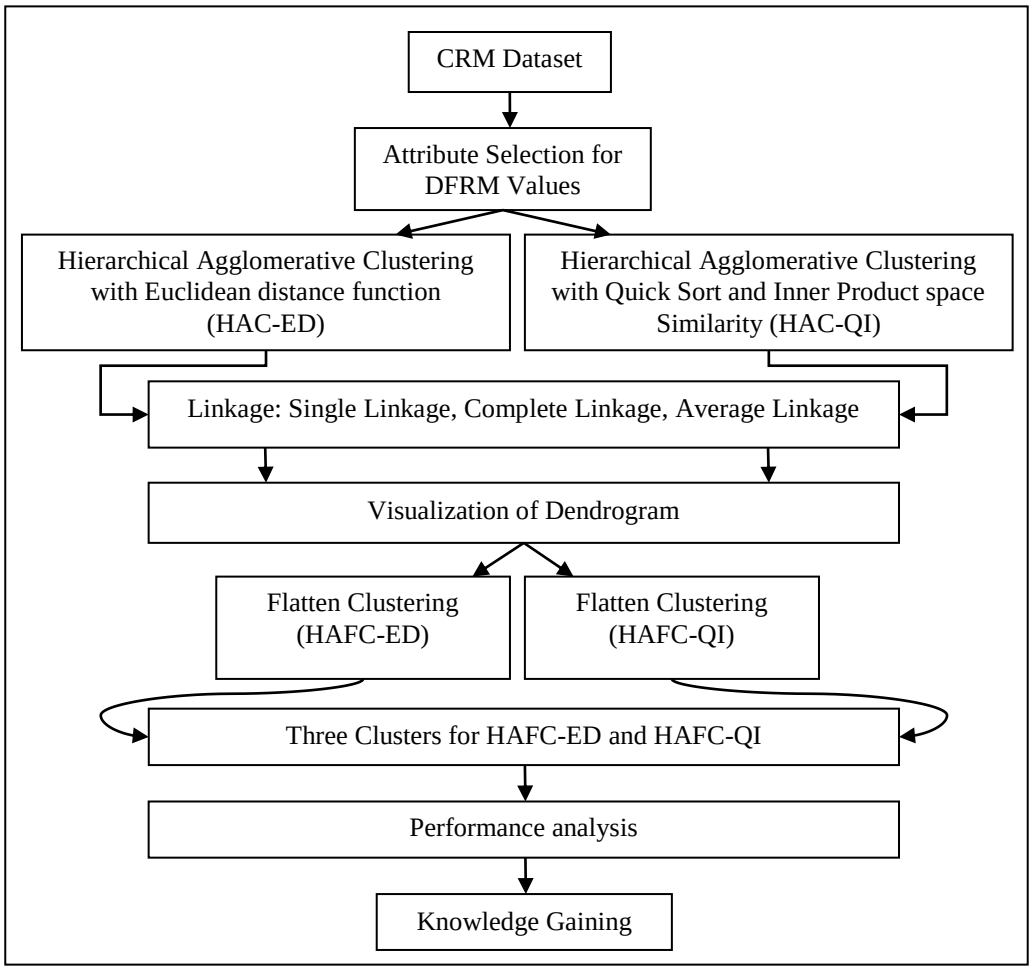

Fig. 4 Proposed Flow of work

Quick Sort

It is a divide and conquer algorithm with a pivot element. Other than the pivot, initially the array is divided into halves. These sub arrays are sorted repeatedly by placing the least element before the pivot and largest element after the pivot. It is an in-place algorithm hence it doesn't needs more additional memory. The steps includes choosing an element as pivot, partitioning other elements based on pivot and recursion which repeatedly do the partition step.

Inner product space Similarity

It changes the vector space with an additional structure by taking product (inner product) of each pair of vectors with a scalar quantity. This product denotes the length of a vector or the angle between the vectors. It is a generalization of Euclidean space by taking scalar dimension.

For the Euclidean space given as $<(x_1, x_2, .., x_n)>$, $<(y_1, y_2, .., y_n))>$

$$\text{Inner Product} = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n \quad (5)$$

Generally it satisfies the properties such as Linearity, Symmetric property and Positive Definite Property.

Flatten Clustering Model

It is viewed as an unstructured cluster model used with Hierarchical to cut the tree or clusters into pre-determined clusters based on the distance of the nodes till the desired number of cluster is reached. It expands the nodes in the order distance between the nodes. The Flatten clustering takes the hierarchical cluster model as the example set as an input and returns the flat cluster model with desired number of clusters. It creates a flat set of clusters from hierarchical model without any explicit structure. This final model can be evaluated with metrics to have the detail about the formed clusters.

Procedure Hierarchical Agglomerative with Flatten Clustering (HAFC-QI)

Step 1: Start with DFRM Values.

Step 2: Sort the dataset with Quick sort.

Step 3: Implement Hierarchical clustering with Inner product similarity distance function as in equation (5) for single, complete and average linkage as per equation (1) – (3).

Step 4: Apply Flatten clustering from Hierarchical Model with predetermined number of clusters.

Step 5: Apply evaluation metric to assess cluster model.

Step 6: Visualize dendrogram from hierarchical cluster and clusters from flatten clustering.

Advantages

Proposed work HAFC-QI overcomes the existing procedure HAFC-ED by including the following improvements.

- Customer is segmented with additional information with demographic information (DFRM).
- Time Complexity of Hierarchical Clustering is reduced with sorted dataset by Quick sort algorithm.
- Inner product similarity is used instead of Euclidean distance.
- Flatten clustering is applied with the hierarchical model to get compact view of meaningful clusters.
- Evaluation metric is used with the output of Flatten clustering to know how well the objects or instances are distributed over the clusters.

4. Result and Discussion

A. Dataset Description

The proposed work is analyzed with two variant sectors namely Banking and Telecom Sector. These two datasets are taken from Kaggle repository.

Banking Sector – It includes 3500 instances and eighteen attributes such as Customer ID, Age, Job, Marital Status, Level of Education, Default Credit score, Balance, Housing Loan, Personal Loan, Deposit details and Regarding contact details (type of communication, last contact day, last month of communication, last contact duration, contact performed during compaign, Number of days passed after client final contact, number of contacts performed before compaign, outcome of the previous compaign.

Attribute Selection: From these attributes, DRFM values are extracted by selecting the appropriate attributes such age, job, marital status, Level of Education which has the details of Demographic information and attributes such as last contact day, last month of communication, last contact duration, Number of days passed after client final contact, number of contacts performed before compaign which gives the detail of Recency, Frequency and Balance with Monetory information.

Telecom Sector – It includes 3910 instances and seventeen attributes such as Customer ID, Partner, Dependent, Tenure, having Phone service, Multiple lines, Internet service, Online service, Online backup, Device Protection, Technical support, Streaming TV, Streaming Movie, Contract, Paperless billing, Payment mode, Monthly charge and Total Charge.

Attribute Selection - From these attributes, Partner, dependent gives information about Demographic, Contract and Tenure provides details regarding Recency, Frequency and Monthly, total charges gives details about Monetory values regarding customers.

B. Evaluation Measures for Hierarchical Clustering

1. Sum of squares – It measures the deviation or variation of the mean value of the data points. It will be zero while all the data objects remain same within the cluster. Higher value of sum-of-squares implies large degree of variability and a lower value implies less variability from the mean value of the data points.

2. Processing time – Total time to build hierarchy model and flatten clustering to produce clusters.

C. Results

The performance for Hierarchical Flatten clustering with Quick sort and Inner product similarity (HAFC-QI) is compared with Hierarchical Flatten clustering with Euclidean distance function (HAFC-ED).

1) Sum of Squares

Clusters are formed for three types of linkage single, Complete and Average with existing HAFC-ED and proposed HAFC-QI methods. The results are shown in Table 1.

Table 1. Performance analysis for Sum of Squares

	Bank			Telecom		
	Single	Complete	Average	Single	Complete	Average
HAFC-ED	0.954	0.844	0.902	0.721	0.654	0.701
HAFC-QI	0.596	0.412	0.512	0.551	0.402	0.484

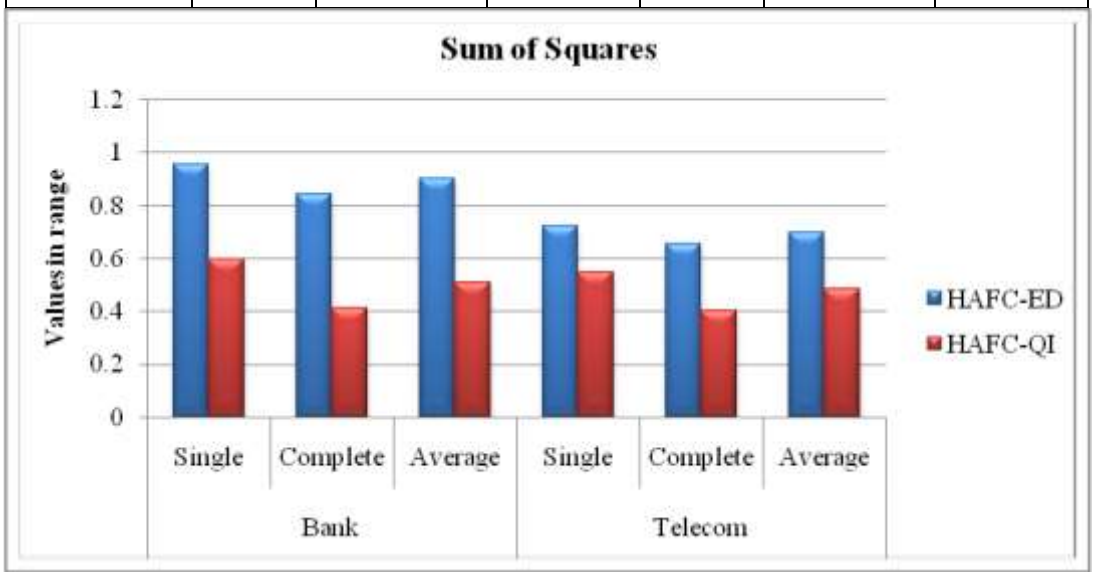

Fig 5. Sum of Squares for Banking and Telecom dataset

From Fig 5, it is concluded that hierarchical method HAFC-QI produce better clusters with less sum of squares. Among them, HAFC-QI with complete linkage hierarchical method has least sum of squares which ensures cluster quality.

2) Processing Time (in sec.)

Time to create three clusters is listed in Table 2 for three types of linkage in Hierarchical clustering.

Table 2. Performance analysis for Processing time

	Bank			Telecom		
	Single	Complete	Average	Single	Complete	Average
HAFC-ED	2.36	2.09	2.23	3.32	3.10	3.21
HAFC-QI	1.55	1.42	2.00	3.12	2.51	3.02

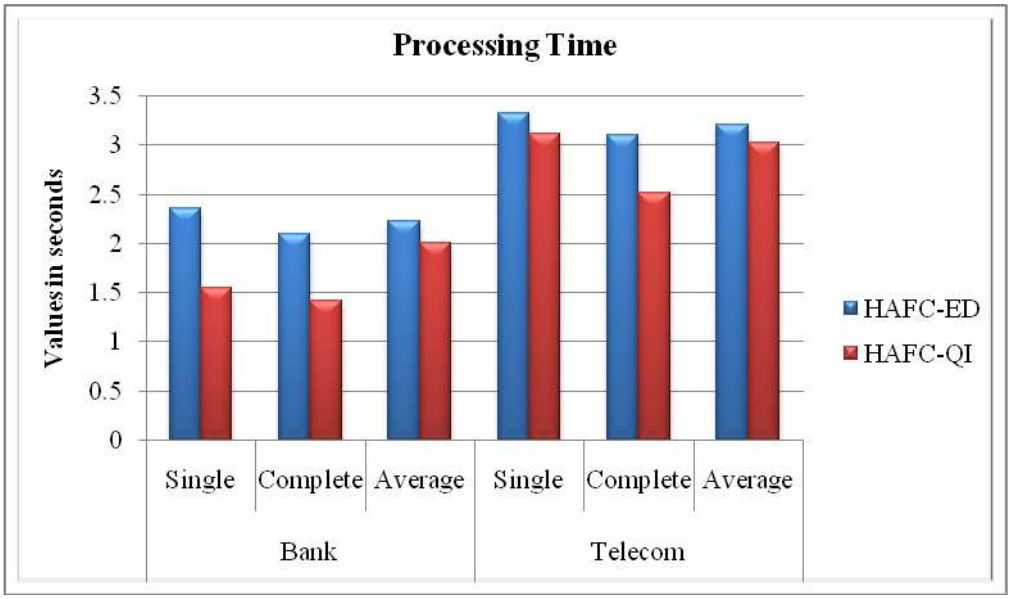

Fig 6. Processing time analysis

In Fig 6, processing time is analyzed for existing and proposed method for single, complete and average link method in Hierarchical clustering. The results confirmed that complete link method in HAFC-QI takes less processing time for both sectors.

D. Cluster Result Analysis

From Table 1, 2, it is concluded that complete linkage in Hierarchical clustering gives better results for both sector Banking and Telecom. So the clustered results are analyzed with complete linkage method for HAFC-QI in Table 3.

Table 3. Clusters for Banking and Telecom for HAFC-QI with Complete Linkage

Sectors	Total Instances	Cluster 0	Cluster 1	Cluster 2
Banking	3500	226	2816	458
Telecom	3910	2219	370	1321

From the result analysis the description of the clusters with complete link method is:

Banking Sector with complete linkage

The proposed HAFC-QI method creates three clusters in which Cluster 0 has 226 instances, Cluster 1 has 2816 instances and Cluster 2 has 458 instances.

Cluster 0 – Customers having primary level education working as technicians, blue-collared and other services, with singled, married or divorced marital status, having no loans with minimum balance belongs to this cluster.

Cluster 1- Customers having secondary level education, working as self employed, retired, students with marital status singled, married, having any one loan with moderate balance belongs this cluster.

Cluster 2 – Customers having tertiary level education, in management, administration job or entrepreneur with singled, married or divorced marital status, having two types of loan with high balance belongs to this cluster.

Telecom Sector with complete linkage

The proposed HAFC-QI method creates three clusters in which Cluster 0 has 2219 instances, Cluster 1 has 370 instances and Cluster 2 has 1321 instances.

Cluster 0 – Customers paying total charges above 5000, monthly charge around 150, total downloaded services more than 6 and the customers are tenured for more than fifty years combined with partner, having dependents belongs to this cluster.

Cluster 1 – Customers paying total charges above 3000-5000, monthly charge around 100, total downloaded services minimum 4 and the customers are tenured for minimum thirty years combined with partner, having dependents belongs to this cluster.

Cluster 2 **–** Customers paying total charges below 3000, monthly charge around 50, total downloaded services minimum 2 and the customers are tenured for below ten years, having no partner and dependents belongs to this cluster.

5. Conclusion and Future Work

Customer segmentation with Demographic, Recency, Frequency and Monetary (DFRM) values have huge impact in customer relationship management in improving the business. Hierarchical Agglomerative is a bottom up clustering model that produces numerous clusters and tough to visualize and extract the pattern of the clusters also it takes high processing time. By considering these two issues, the proposed procedure HAFC-QI implements Hierarchical agglomerative Flatten clustering with Inner product similarity measure and Quick sort algorithm. This is compared with existing Euclidean distance measure in Hierarchical Flatten clustering HAFC-ED. The procedure is applied for two variant sectors Banking and Telecom and from the results it is proved that the proposed procedure outperforms the existing in terms of sum of squares and processing time.

In future, the work can be extended by implementing the procedure for other type of customer oriented sector. Also, variant distance metric and sorting algorithm may be used to further improve the performance of Hierarchical clustering.

References

[1] Alex Berson, Stephen Smith, Kurt Thearling, "Building Data Mining Applications for CRM", Tata McGraw-Hill Publishing, Seventeenth reprint, 2011.

[2] Michael J.A.Berry, Gorden S. Linoff, "Mastering Data Mining- The Art and Science of Customer Relationship Management", John Wiley & Sons Publication, 2000.

[3] Brito .P, " Customer segmentation in a large database of an online customized fashion business", Robotics and computer integrated AMnufacturing, Vol. 36, pp. 93-100,2015.

[4] Olivia Parr Rud, Michael J.A.Berry, "Data Mining:Modeling Data for Marketing, Risk and CRM", Wiley Publications, 2003.

[5] Anjali Nair Rohan Khasgiwala Snigdha Mishra, " Improving Customer Relationship Management Using Data Mining", International Journal of Scientific & Engineering Research, Volume 7, Issue 2, February-2016.

[6] Sajana .T, Sheela Rani .C and Narayana .K, "A Survey on Clustering Techniques for Big Data Mining", Indian Journal of Science and Technology, vol. 9, no. 3, 2016.

[7] Anika Singh, Prasanna Balaji Venkataraman, "Customer segmentation through K-Means and Hierarchical Clustering Techniques", IJSRD - International Journal for Scientific Research & Development| Vol. 5, Issue 07, 2017.

[8] Shreya Tripathi, Aditya Bhardwaj, Poovammal, " Approaches to Clustering in Customer Segmentation" International journal of Engineering and Technology, Vol 7 (3.12), 2018.

[9] Phan Duy Hung, Nguyen Thi Thuy Lien, Nguyen Duc Ngoc, "Customer Segmentation Using Hierarchical Agglomerative Clustering", *ICISS,* Association for Computing Machinery (ACM), 2019.

[10] Fahed Yoseph, Nurul Hashimah Ahamed Hassain Malim, Mohammad Aimalaily, "New behavioral segmentation methods to understand consumers in retail industry", International Journal of Computer Science & Information Technology (IJCSIT) Vol 11, No 1, February 2019.

[11] Sabitha .R, Karthik .S, "Customer Segmentation Framework using Redefined Centroids", International Journal of Applied Engineering Research, ISSN 0973-4562, Vol. 10, No.29, 2015.

[12] Kaushik, M, Mathur, B, "Comparative Study of K-means & Hierarchical Clustering Techniques. International Journal of Software & Hardware Research in Engineering, 2(6), 93–98.

[13] Rani, Y., Rohil, H. "A Study of Hierarchical Clustering Algorithm. International Journal of Information and Computation Technology, 2013, 3(11), 1225–1232.

[14] Maimon .O and Rokach .L, "Clustering methods", in Data Mining and Knowledge Discovery Handbook. Boston: Springer US, 2005, pp. 321-352.

[15] Kettani, O., Ramdani, F., Tadili, B. "An Agglomerative Clustering Method for Large Data Sets. International Journal of Computer Applications, 2014, 92(14), 975–8887.