8th International Conference on
Combinatorics, Cryptography, Computer Science and Computation
November. 15-16, 2023

Comparative Study of Novel and Existing Fuzzy Clustering Algorithms for Customer Segmentation Based on a New RFM Model

Mahsa Hamidi[1]
Ferdowsi University of Mashhad, Iran

Omid Solaymni Fard
Ferdowsi University of Mashhad, Iran

Abstract

Customer segmentation has been a hot topic for decades, and the competition among businesses makes it more challenging.The main objective of this paper is to compare the performance of a novel fuzzy clustering algorithm with three other existing fuzzy-based clustering algorithms on a new RFM-based model for customer segmentation. The RFM model is a technique that uses four parameters, namely recency, frequency, monetary value, and duration, to analyze customer behavior and assign them to different clusters. The paper also aims to determine the optimal number of clusters for each algorithm by using internal evaluation measures, such as the silhouette coefficient,Davies Bouldin and Calinski Harabasz. You can find the data set and codes in my GitHub.

Keywords: Fuzzy clustering, customer segmentation, Validation index, Intutionistic fuzzy c-means, Type2 fuzzy c-means.

1 Introduction

The interaction between society's needs and business firms' offerings is the basis of business. Customers are the center of attention for every industry; industries always try to satisfy their customers' demands. Whether a company is big or small, it has to face competition. Many competitors fail to survive. We think that one of the most frequent reasons for failure is "companies choosing to ignore their customers"[1]. It is much more expensive to acquire new customers than to keep the ones you already have. Therefore, The main issue for businesses is how to increase the sales of their existing customers. Analyzing the purchase data of a platform to understand how users make choices in the real world is a key challenge for improving the business performance. Customer segmentation is a simple way of dividing customers into groups based on different criteria and marketing to them accordingly. As a result, each customer segment requires a different strategy or approach. Knowing the characteristics of customers is essential for e-commerce success and creating marketing strategies that are tailored to different customer groups.[2] To achieve this goal, we need data that is segmented by customers, so that we can target them for marketing. In this study, the researcher used the biggest e-commerce dataset in Pakistan to help new and existing businesses in Pakistan.[3]

[1]speaker

Recency: The number of months between the specified date and the last transaction date for each customer. Frequency: The total number of transactions made by each customer. Monetary: The sum of the transaction amounts for each customer. Time: The total number of days between consecutive transactions for each customer, converted to months.

The authors examined the RFMT dimensions of the customers in this article. They used the silhouette, Calinsky–Harabasz and Davies–Bouldin index to evaluate the clustering quality from cluster 0 to cluster 10. They selected the stable cluster by majority voting, which resulted in 03 for FCM, T2FCM, IFCM and T2IFCM. By targeting customers based on their needs and habits, and by offering different packages that are identified in the customer segmentation process, the seller can improve their profit from their strategies. The main idea of customer segmentation is to group customers who have similar demographic characteristics, behaviors, values, etc. There are a lot of data mining techniques for customer segmentation, among these methods clustering is used widely. Fuzzy Clustering is a type of clustering algorithm in machine learning that allows a data point to belong to more than one cluster with different degrees of membership.

2 preliminary

2.1 Algorithms

Fuzzy c-means (FCM)[4]: FCM is the first fuzzy clustering algorithm that bridged clustering to fuzzy cluster- ing. It works on the supposition that number of clusters in the data set are pre-known. The FCM algorithm starts with an initial guess for the cluster centers, which represent the mean location of each cluster. Optimal clusters are obtained by minimizing FCM objective function (Obj_{FCM}), subject to the constraint that sum of all the memberships of a data object to each cluster is one and the same is formulated in Eq.(2):

$$Obj_{FCM} = \sum_{j=1}^{n}\sum_{i=1}^{z} u_{ij}^{m} dist_{ij}^{2} \tag{1}$$

where n is the number of clusters; z is count of data objects in given dateset; u_{ij} is the fuzzy membership of d_i in c_j; d_i is ith data object in the data set D ,where $D = \{d_1, d_2, ..., d_z\}$; c_j is jth centroid in the set of centroid C, where C is represented as $\{c_1, c_2, ..., c_n\}$; m is the fuzziness index; $dist_{ij} = \sqrt{\sum_{i}^{dim}(d_{ik} - c_{jk})^2}$ is Euclidean distance between d_i and c_j , dim represents dimensionality of the data set – D.

$$\sum_{j=1}^{n} u_{ij} = 1, \forall i = 1, 2, ..., z \tag{2}$$

Lagrangian multipliers is used to compute the optimal solution for Obj_{FCM} and the optimal solution is given by Eqs.(3) and (4):

$$c_j = \frac{\sum_{i=1}^{z} u_{ij}^{m} d_i}{\sum_{i=1}^{z} u_{ij}^{m}} \tag{3}$$

$$u_{ij} = \left(\sum_{k=1}^{n} \left(\frac{dist_{ij}^{2}}{dist_{ik}^{2}} \right)^{\frac{1}{(m-1)}} \right)^{-1} \tag{4}$$

Intuitionistic fuzzy-C-means (IFCM)[6]: There is an important measure in IFCM which is called hesitation degree. The objective function in IFCM is:

$$Obj_{IFCM} = \sum_{j=1}^{n}\sum_{i=1}^{z}(u_{ij}^{*})^{m}dist_{ij}^{2} + \sum_{j=1}^{n}\pi_{j}^{*}e^{1-\pi_{j}^{*}} \tag{5}$$

where $\pi_j^* = \frac{1}{z}\sum_{i=1}^{z}\pi_{ij}, i \in [1,z]$; $\pi_{ij} = 1 - u_{ij} - (1-u_{ij}^{\alpha})^{\frac{1}{\alpha}}, \alpha > 0$; $u_{ij}^* = u_{ij} + \pi_{ij}$.
where u_{ij} is fuzzy membership and α is a tuning parameter for algorithms using Intuitionistic fuzzy sets.u_{ij} and π_{ij} expresses fraction of belonging and not belonging of i_{th} data object (d_i) to j_{th} centroid (c_j) respectively. Moreover centroids updates as follows:

$$c_j = \frac{\sum_{i=1}^{z} u_{ij}^{*} d_i}{\sum_{i=1}^{z} u_{ij}^{*}} \tag{6}$$

Type-II fuzzy-c-means (T2FCM)[5]: T2FCM introduces the concept of Type-II membership (T2m), which is based on the idea that the contribution of data objects in updating a centroid should be proportional to their membership on that centroid. The objective function in T2FCM is represented by 1, and the centers are obtained using 7:

$$c_j = \frac{\sum_{i=1}^{z} a_{ij}^{m} d_i}{\sum_{i=1}^{z} a_{ij}^{m}} \tag{7}$$

where $a_{ij} = u_{ij} - \frac{(1-u_{ij})}{2}$. Now,we first introduce the latest fuzzy clustering algorithm and then discuss about validation index in fuzzy clustering to determine optimize number of centers.
Type-II intuitionistic fuzzy-c-means(T2IFCM)[7]: In T2IFCM algorithm, formulation of optimal clusters is based on minimization of objective function5, and all its parameters are the same as IFCM that is described in the previous page.The main feature of this algorithm is the introduction of Intuitionistic Type2 membership degree(IT2m), which is formulated as follows and is used in updating the centers in 9.

$$b_{ij}^{*} = a_{ij} + \pi_{ij} \tag{8}$$

where $a_{ij} = u_{ij} - \frac{(1-u_{ij})}{2}$.

$$c_j = \frac{\sum_{i=1}^{z} b_{ij}^{*m} d_i}{\sum_{i=1}^{z} b_{ij}^{*m}} \tag{9}$$

2.2 RFM (Recency, Frequency, and Monetary) Analysis

A common method for customer segmentation is the RFM (Recency, Frequency, and Monetary) analysis, but it does not include a crucial factor of time, i.e., T. In[8] authors explored the effect of adding time (T) to the RFMT model to better capture customer loyalty and customer behavior. This way, we can consider the long-term relationships with customers.

2.3 Determine the number of clusters

Since in the mentioned algorithms in 2.1, the number of clusters should be given as a default parameter to the algorithm, therefore, determining the appropriate number of clusters is one of the important things in such algorithms. For this purpose, a number of validation indices to determine the appropriate number of clusters in fuzzy clustering algorithms have been introduced by researchers. In this research work, we used three internal cluster validation measures: silhouette, the Calinsky–Harabasz index, and the Davies–Bouldin

index. By using multiple validation measures instead of one, we can achieve more accurate clustering of the data.

- **Davies–Bouldin:**The Davies–Bouldin index can be calculated by the following formula:

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

  where n is the number of clusters, c_i is the centroid of cluster i and σ_i is the average distance of all elements in cluster i, $d(c_i, c_j)$ is the distance between centroids c_i and c_j. The lower the Davies–Bouldin index, the better the clustering quality. Therefore, the best clustering algorithm based on this criterion is the one that produces the smallest Davies–Bouldin index for a collection of clusters.

- **Silhouette Score:** The formula for the silhouette score is:

$$S_{(i=2,\ldots,n)} = (S_i - S_i^{'})/\max(S, S_i^{"}) \tag{10}$$

  S_i = Average distance of items between i_{th}group/cluster.
  $S_i^{'}$ = Average distance between $i_{th} cluster with different groups/clusters.$
  $\max(S, S_i^{"})$ = Average distance between S_I and $S_i^{'}$.
  Objects with a high silhouette value are considered well clustered, objects with a low value may be outliers.

- **Calinski–Harabasz:** The CH Index (also known as Variance ratio criterion) is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). A higher CH index indicates that the clusters are dense and well-spaced.

2.4 Dataset

The largest Pakistani e-commerce dataset by Zeeshan-ul-Hassan[3] was used in this study, , which covers data from 1 July 2016, to 28 August 2018. The dataset has 21 fields and half a million transaction records. The fields we focus on are 'Status', 'created-at', 'price', 'MV', 'grand-total', 'category-name', 'payment-method', 'year', 'month', 'FY', and 'Customer ID'. The status field indicates whether the transaction is completed, incomplete, canceled, or refunded, etc., and we segment the data based on this field. The created-at field (the sale date) shows the date and time of each transaction. The price field shows the product price. The MV field is the monetary value or the actual amount paid for the product. The grand-total field is the total amount paid for a transaction. The category-name field shows the product category. The payment-method field shows how the product was paid for. The year field shows the year of the transaction. The month field shows the month of the transaction. The M-Y field shows the month and year of the transaction. The FY field shows the financial year of the transaction. The Customer ID field is a unique identifier for each customer. Null, negative, missing, and invalid literals are removed during data cleaning. We use the RFMT model to segment customers, so we need to convert the data from the dataset to the RFMT data format. First, we use the Customer ID as a unique identifier and a primary key. The column names are 'created at' for recency, 'increment id' for frequency, 'MV' for monetary and 'WorkingDate' for time. We calculate and rename the RFMT values for each customer from the dataset. The monetary (M) value is the sum of

all the expenses of a customer. The frequency (F) value is the number of purchases made by a customer. The recency (R) value is the time difference between the customer's last purchase and the reference date, 1 March 2020. We use months as the unit of time for recency and time. We also add purchase duration (T), the fourth variable, which measures the average time between consecutive purchases. If a customer's first and last purchase dates are t_1 and t_n, we can estimate the customer's purchase cycle (T) in months by dividing the months between t_1 and tn by the number of purchases. The formula to calculate T is:

$$T = t_n - t_1$$

The dataset had 584,524 shopping records from 115,081 distinct consumers.

3 Main results

We computed the Davies–Bouldin, Silhouette and Calinski–Harabasz index for four clustering algorithms (FCM, IFCM, T2FCM, and T2IFCM) in a RFMT model of customer segmentation problem on the dataset to compare their performance, which are included in the table.

Factors	**FCM**		**IFCM**		**T2FCM**		**T2IFCM**	
	Score	**Cluster**	**Score**	**Cluster**	**Score**	**Cluster**	**Score**	**Cluster**
Davies–Bouldin	1.0408	3	1.0320	8	1.291	4	1.0303	8
Silhouette score	0.3065	3	0.3083	3	0.3119	3	0.1769	8
Calinski–Harabasz	100.0024	3	102.2415	3	101.6793	3	84.9435	3
Algorithms, wise majority voting		3		3		3		8

Table 1: Clusters factors analysis scores of the corresponding cluster for different algorithms.

3.1 Majority Voting

One way to make a decision based on a group of classifiers is to use majority voting. This method has three types. Unanimous voting happens when all classifiers agree on the same prediction. Simple voting occurs when more than half of the classifiers make the same prediction. The most voted candidate by the classifiers is c-means = 3, IFCM = 3, T2FCM= 3, and T2IFCM =8. The factors that the clusters predict are (3, 3, 3, 8, 3, 3, 4, 3, 3, 8, 8, 3), and the frequency of each cluster value is

$$f_{cluster} = (NumberofOccurrencesofthecluster)$$

so we have $f_3 = 8$, $f_8 = 3$, $f_4 = 1$. Below are the various factors for cluster analysis. It might be hard

Cluster	Frequency of Occurrences
3	8
8	3
4	1

Table 2: Clusters and their frequency of occurrences.

to pick the right cluster because of the different components. Therefore, the model's cluster is chosen by a majority vote. The cluster number for each algorithm is given here.

$$Model_{algo} = Mode(Silhouette_{algo}, CH_{algo}, DB_{algo})$$

where algo is the algorithm, CH = Calinski–Harabasz, DB = Davies–Bouldin. They choose the optimum cluster, i.e., $f_3 = 8$ times, because of the majority voting.

3.2 Implementation of algorithms

Over 26 months, there are 115,081 consumers and PKRS.4195251105 purchases in the three clusters. The clusters C_0 of FCM, IFCM, T2FCM and T2IFCM have the following customer proportions: $0.16\%, 99.8\%, 2.5\%$ and 26.86%; cluster C_1 has a proportion of customers of $0.01\%, 0.13\%, 21.29\%$ and 68.20% respectively. Cluster C_2 has the proportion of customers of $99.8\%, 0.009\%, 76.15\%$ and 4.94%.
Recency-Frequency-Monetary(RFM), inter-purchase Time-Frequency-Monetary(TRM), and inter-purchase Time-Recency-Monetary(TRM) graphs are used to create a three-dimensional(3D) representation of data. Each algorithm in figure1 depicts the implementation of the mentioned fuzzy clustering algorithms with 3 clusters on the dataset in three of the four variables(RFMT).

3.3 External evaluation

In external evaluation, clustering results are evaluated based on data that was not used for clustering, such as known class labels and external benchmarks. Such benchmarks consist of a set of pre-classified items, and these sets are often created by (expert) humans. Thus, the benchmark sets can be thought of as a gold standard for evaluation.
A number of measures are adapted from variants used to evaluate classification tasks. In place of counting the number of times a class was correctly assigned to a single data point (known as true positives), such pair counting metrics assess whether each pair of data points that is truly in the same cluster is predicted to be in the same cluster.
In this case, we implemented K-means algorithm on the dataset(with 3 clusters) then it is considered as a ground truth to compute the efficiency of fuzzy clustering algorithms.

Algorithm	Rand score	Fowlkes Mallows score	V-measure
FCM	0.999	0.999	0.909
IFCM	0.998	0.999	0.762
T2FCM	0.663	0.813	0.001
T2IFCM	0.631	0.793	0.010

Table 3: Calculation of Rand score, Fowlkes Mallows score and V-measure as external indexes for each algorithms

References

[1] Rahul, S.; Laxmiputra, S.; Saraswati,J *Customer segmentation using rfm model and k-means clustering.* Int. J. Sci. Res. Sci. Technol. 2021, 8, 591–597.

[2] Jinfeng, Z.; Jinliang, W.; Bugao, X.,. *Customer segmentation by web content mining* J. Retail. Consum. Serv. 2021, 61, 102588.

[3] Zeeshan-ul-Hassan, U.. *Pakistan E-Commerce Largest Dataset Pakistan. Chisel. 2021. Available online*ccessed on 1 December 2021)

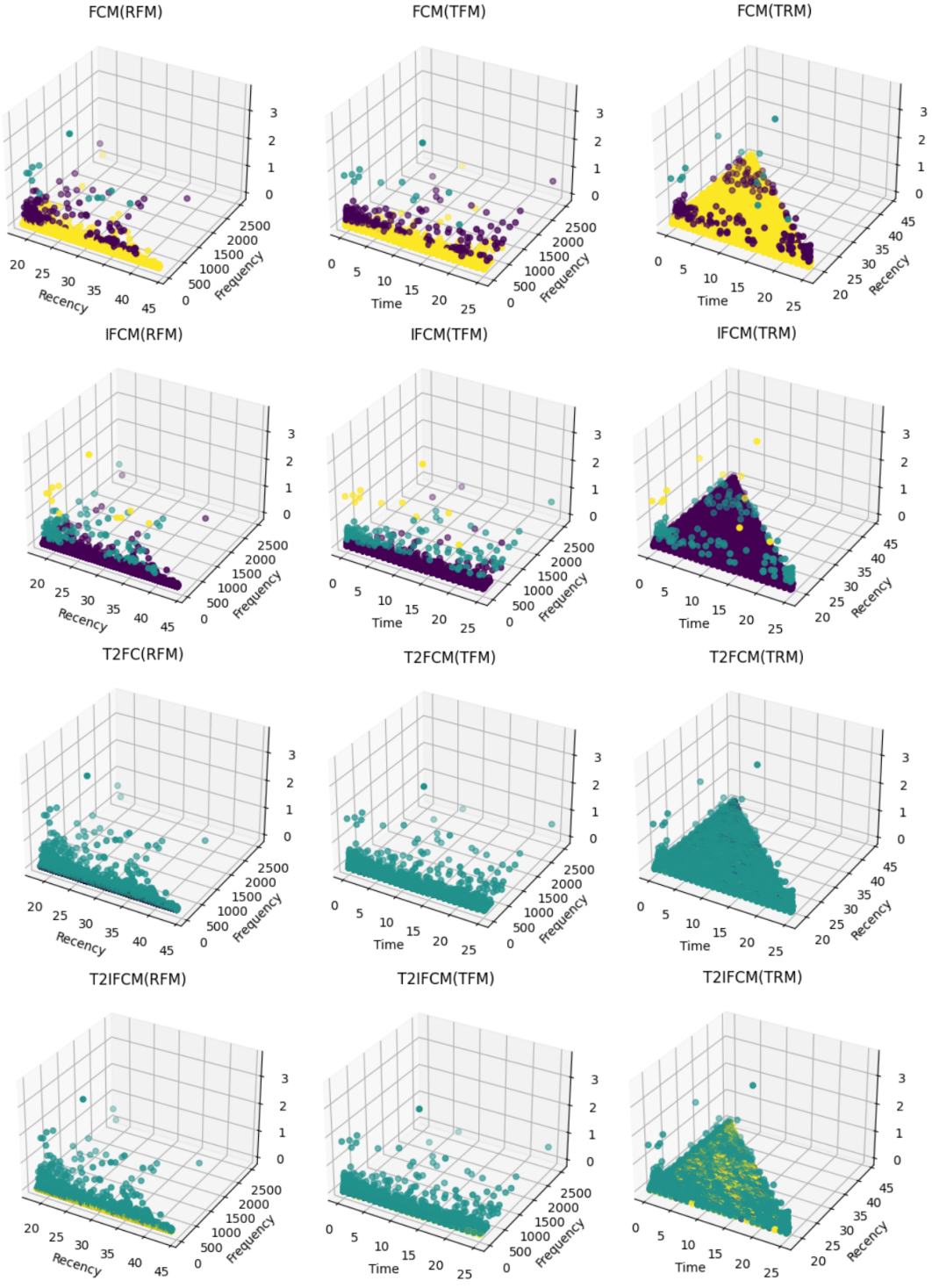

Figure 1: Customer segmentation in the three clusters of RFMT in different algorithms.

[4] Saumendra, and Janmenjoy Nayak, *Customer segmentation via data mining techniques: state-of-the-art review.*, Proceedings of ICCIDM 2022.

[5] Ali, Noureddine Ait. *The performances of iterative type-2 fuzzy C-mean on GPU for image segmentation*, 2022.

[6] Dogan, Onu,*Segmentation of retail consumers with soft clustering approach.*, 2020.

[7] Dahiya, Sonika, and Anjana Gosain. *A novel type-II intuitionistic fuzzy clustering algorithm for mammograms segmentation*, 2022.

[8] Ullah, Asmat, et al. *"Customer Analysis Using Machine Learning-Based Classification Algorithms for Effective Segmentation Using Recency, Frequency, Monetary, and Time."* Sensors 23.6 (2023): 3180.

e-mail: moonlikehamidi@gmail.com
e-mail: omidsfard@gmail.com