NEW MARKET SEGMENTATION METHODS USING ENHANCED (RFM), CLV, MODIFIED REGRESSION AND CLUSTERING METHODS

Fahed Yoseph¹ and Mohammad AlMalaily²

¹School of Business and Economics, Åbo Akademi University
²Faculty of Engineering, Mansoura University

ABSTRACT

A widely used approach for gaining insight into the heterogeneity of consumer’s buying behavior is market segmentation. Conventional market segmentation models often ignore the fact that consumers’ behavior may evolve over time. Therefore retailers consume limited resources attempting to service unprofitable consumers. This study looks into the integration between enhanced Recency, Frequency, Monetary (RFM) scores and Consumer Lifetime Value (CLV) matrix for a medium size retailer in the State of Kuwait. A modified regression algorithm investigates the consumer purchase trend gaining knowledge from a point-of-sales data warehouse. In addition, this study applies enhanced normal distribution formula to remove outliers, followed by soft clustering Fuzzy C-Means and hard clustering Expectation Maximization (EM) algorithms to the analysis of consumer buying behavior. Using cluster quality assessment shows EM algorithm scales much better than Fuzzy C-Means algorithm with its ability to assign good initial points in the smaller dataset.

KEYWORDS

Segmentation, Clustering, RFM model, Retailing

1. INTRODUCTION

Mass marketing strategies which are mainly based on marketing experts and sales manager’s opinions of the market [17]. For example, Gholamian [22] states that the retail industry is highly competitive, with the number of products often overwhelming. Consumers are faced with a variety of products, causing the demand to be higher and more complex. In light of this trend, modern marketing moves from mass-marketing (products-focus) to target-marketing (consumer-focus). In response, small and medium-sized retailers (SMR) segment their markets to be more strategic in their planning and design and implement successful marketing strategies and retention policies. In the fast-changing retail industry, there is a clear need for advanced methods to discover market segments from sales and other data, with market-segmentation empowering retailers to precisely reach consumers with specific needs and wants, by dividing the market into similar and identifiable segments, to focus on individuals with similar preferences, choices, needs and interests on a common platform [24], [27]. Segmentation evaluates consumers as a segment indirectly, rather than individually or directly. It enables retailers to make full use of their limited resources to serve consumers effectively as consumer sub-groups [10]. Proper mechanisms for treating point-of-sales (POS) events convert ever-increasing transaction data into knowledge [34]. To take on this challenge, this study introduces an enhanced normal distribution formula to
eliminate outliers in the dataset, followed by developing two of the most popular techniques of market segmentation: Recency Frequency, Monetary (RFM) scores with Consumer Lifetime Value (CLV). RFM, Birant [4], distinguishes important consumers and their purchase behavior by involving: the consumer’s most recent purchase (R), the frequency of purchases (F), and spent money (M). These variables and appropriate feature weights calculate an RFM-score that is a key figure of market segmentation. CLV is quite different in that it is a quantitative measurement on the number of sales the consumer is expected to commit to a retailer in total time [14]. CLV is paired to RFM to predict the future cash flows attributed to the consumer during his or her entire lifetime with the retailer before he or she churn [35].

### Table 1. Market Segmentation Bases

<table>
<thead>
<tr>
<th>Base Description</th>
<th>Goals Benefits</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic; Identifiable population features</td>
<td>The goal is to have a precise customer purchase profile and focuses on measurable criteria of consumers and their households.</td>
<td>[9]</td>
</tr>
<tr>
<td>Geographic; Location-related features</td>
<td>The goal is to map consumer wants and needs from nationality, region to another etc.</td>
<td>[23]</td>
</tr>
<tr>
<td>Behavioral; Product attitudes, customer relationship-related features</td>
<td>The goal is to identify behavioral variables such as occasions, benefits, user status, usage rate, buyer-readiness stage, loyalty status and attitude.</td>
<td>[9], [5]</td>
</tr>
<tr>
<td>Psycho-Graphic; Lifestyle-related features</td>
<td>The goal is to target specific groups such as more budget-conscious customers, i.e. smart shoppers who value a good deal.</td>
<td>[5]</td>
</tr>
</tbody>
</table>

![Customer Value Matrix](image)

Fig.1: The customer value matrix
2. IMPERIAL CASE STUDY

A Kuwaiti medium retail chain with 25 domestic branches, has grown a changing set of consumers. The consumer-base is now diverse and owing to the demographic changes inside Kuwait is no longer possible to define consumer profiles based on the known, previous, consumption patterns and using traditional marketing strategy.

POS data from this medium-sized retailer from Kuwait has the benefit that it is usually generated and stored in a structured way and is relatively easy to aggregate to consumer-level. This characteristic of POS data makes it useful for analyses that require (more or less) complete data, such as Fuzzy C-Means and EM clustering algorithms.

Our objective is to design an advanced segmentation method, construct a system utilizing the method and test this system (the artefact) by analyzing the results produced by the system. The idea is to show how the market segmentation process can be improved with a hybrid approach that utilizes both regression and clustering as steps in the analysis of a POS data warehouse. Also to show that with an appropriate design more usable results can be produced. More specifically, to answer the following study questions:

1. Can a method with RFM, clv, normal distribution, regression and clustering steps help discovering hidden patterns about consumer purchase behavior and segment the SMR customer-base?
2. Can the proposed method help mapping the entire consumer's journey?

Sections that follow elaborate this by reviewing appropriate literature, in section 2, developing the modeling system the system in section 3, and analyzing the results in section 4. Conclusions follow in section 5 with answers to the aforementioned study questions.

2.1 MARKET SEGMENTATION

The US Small Business Administration has traditionally defined Small to Medium Size Retailers as businesses employing fewer than 500 employees [34]. According to the European Commission, the SMR industry forms the backbone of the economy and are the key players in the creation of new jobs and economic growth [19]. Historically, small size retailers have had the privilege of developing close and mutually beneficial relationships with their consumers, thus keeping existing consumers and reaching new markets is a major challenge for the retailer [16], [20]. These relationships were possible because consumer’s buying behavior did not change much, and the price was less of an issue due to less competition [6]. However, the recent economic and social changes have transformed the retail industry, particularly the relationship between the retailer and consumers has changed significantly. As a result, retailers have been forced to seek new marketing strategies to identify the profitable segment of consumers, to develop marketing mixes that appeal to those potential segments of consumers and to focus on providing value to the key segments of consumers [16], [20].

The proposed enhanced normal distribution formula which allows the client to fully eliminate outliers or keep percentage of outliers in the retrieved dataset. The formula is developed using variables that can be changed based on the client desire whether to keep outliers or remove them. As the outliers can be an indication of variance in the dataset or a mistake during data collection phase. Our approach is to eliminate outlier points by removing any points above the Mean. However, we are not in position to decide if they are important or not. To be more practical in this
study we introduce enhanced Normal Distribution formula with an independent variable (0 to 1) to allow the client to change and compare output with different percentage of outliers.

following values (1 = Normal Distribution, closest point to the mean with small percentage of outliers. Reducing the value lower than 1 will increase probability of outliers in the extracted data), calculating the mean and standard deviation in the below formula. Probability density function $Y1$ of normally distributed variables is given by

$$Y1 = f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } -\infty < x < \infty$$

The total area under the normal curve $Y1$ is equal to 1.

The Coefficient $\frac{1}{\sigma \sqrt{2\pi}}$ keeps the area under the curve $= 1$, but this will make the value of the mean point varies between 1 and 0. While the objective to keep the value always at the mean point equal to 1. Therefore, we remove the effect of coefficient by multiplying the equation (Y1) by the coefficient to the power of -1.
\[ Y_2 = (\sqrt{\frac{2\pi}{\sigma^2}} Y_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{\sqrt{2\pi}}{\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

Continuous for all values of X between -∞ and ∞ so that each conceivable interval of real numbers has a probability other than zero. -∞ ≤ V ≤ ∞.

The final step is we further modified the equation by adding a variable (v Where 0 < v < ∞) to extend the Y2 horizontally so most of the values closer to the mean will be closer to 1. And the values far from the mean point will have a lower value:

\[ Y_3 = (Y_2)^v = \left( e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right)^v \]

![Fig 2: Shows the 3 Phases in the Normal Distribution](image)

The literature has traditionally defined RFM analysis as the standard approach to assess and understand consumer lifetime value, and it is quite popular, especially in the retail industry. According to Tsiptsis and Chorianopoulos [33], RFM involves the calculation and the examination of three variables – Recency, Frequency, and Monetary (RFM). Recency refers to the inverse of the most recent interval from the time when the latest consuming behavior happens to the present moment. Frequency is the number of events the consumer purchases in a period. Monetary is simply the amount of money consumed during the period. As the weighted average of its individual components, the RFM score and is calculated as

\[
\text{RFMPQ score scaled} = \frac{\text{RFMPQ score} - \text{min(RFMPQ score)}}{\text{max}(\text{RFMPQ score}) - \text{min}(\text{RFMPQ score})}
\]  

where \( rs = \) recency score and \( rw = \) recency weight, \( fs = \) frequency score and \( fw = \) frequency weight, \( ms = \) monetary score and \( mw = \) monetary weight, \( ps = \) average purchase power (per consumer) score and \( pw = \) average purchase power weight, \( qs = \) average purchase power (per product) score and \( qw = \) average purchase power weight.

One limitation of RFM analysis in market segmentation is that the features are assumed static, and they ignore behavioral changes. The recency parameter, however, indicates a momentary change, but it only shows one static, transient event and cannot properly capture long-term dynamic changes in consumer behavior. This is the reason for our proposal to apply a new dynamic variable (C) to show the quantity and sign of change in consumer purchase behavior. Dwyer [14] defines consumer lifetime value (CLV) as a quantitative measurement of the amount of sales the consumer is expected to spend with a retailer over their lifetime. Furthermore, Safari [32] considers CLV as the present value of all future profits obtained from a consumer over his or her lifetime relationship.
with the retailer. To better utilize CLV in every-day decision making, Marcus [29] introduced CLV matrix as a variant of the RFM analysis for small-business retailers. In CLV matrix, F, the frequency of purchase and M, the average purchase amount are used for the segmenting consumers. The easiness to understand quadrant identifiers is considered as its main advantage. In Marcus’ approach, the average values for the number of purchases and the average amount spent per consumer are calculated. After identifying these, each consumer is segmented to one of the four resulting categories (quadrants) based on whether consumers are above or below the axis averages (Fig. 1).

<table>
<thead>
<tr>
<th>Age</th>
<th>1-17</th>
<th>18-24</th>
<th>25-34</th>
<th>35-44</th>
<th>45-54</th>
<th>55+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

### 2.1.2 DATA MINING

Modern information and communication technology generates massive amounts of data to databases, data warehouses, and other repositories. Transforming the insights about (big) data into knowledge can help retailers to make better business decisions [9]. Tuffery [34] sees data mining as a powerful analytical tool for gaining insight into the retail industry. According to Azevedo [2], data mining provides the analyses on product sales, consumer buying habits, data and identify naturally occurring clusters of behavior, which then form the basis of segments. Ramageri and Desai [31] say that in the retail sector, data mining offers insightful measures, taking into account all the factors that affect the value of the consumer to the retailer over the entire course of consumer relationship [34].

Gunaseelan and Uma [21] stated that the main aim in data mining is to discover valuable patterns from a large collection of data for users. It can identify patterns, and apply data analysis and discovery algorithms to produce a data mining model. Models help in generating a model, a hypothesis about the data, that key executives can use to make better-informed decisions [40]. There are two primary data mining process goals, which are verification and discovery. Verification is verifying the user’s hypothesis about the data while discovery is automation of finding unknown patterns [28].

### 2.1.3 CLUSTERING ALGORITHMS

According to Lefait and Kechadi [28], clustering consists of “creating groups of objects based on their features in such a way that the objects belonging to the same groups are similar, and those belonging to different groups are dissimilar. Clustering analysis is one of the most important and prominent market segmentation techniques, and it has long been the dominant and preferred method for market segmentation [38], [24]. For example, D’Urso [37] used FCM method to cluster potential Chinese travelers. The FCM method combines partitioning and hierarchical clustering procedures. Khavand and Tarokh [24] proposed a data mining tool to prepare a framework for segmenting consumers based on their estimated future CLV value in an Iranian private bank, and the method was implemented in a health and beauty company, as well [26]. In retail sales clustering methods have been applied at least in groceries [30], online retail [8] and for identifying strategies for new ventures [7].
Mirkin [36] proposed a framework for partitional fuzzy clustering which suggests a model of how the data are generated from a cluster structure to be identified. R.Suganya [41] extend the FCMP framework to a number of clustering criteria and study the FCMP properties on fitting the underlying proposed model from which data is generated. D'Urso, [37] used FCM method to cluster potential Chinese travelers. The FCM method combines partitioning and hierarchical clustering procedures. Consumer segmentation to help to analyze transaction data with Fuzzy C-Means for clustering and Fuzzy RFM to identifying the consumers who have high and low loyalty.

Fuzzy C-Means algorithm (FCM), is one of the best known and the most widely used fuzzy clustering algorithms [41]. FCM allow each point to have a degree with every cluster center. Each data points are given with a value between 0 and 1 memberships to determine the degree of belonging to each group. The performance of FCM clustering depends on the selection of the initial cluster center and the initial membership value [89]. FCM has a wide domain of applications such as agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, and target recognition [44].


In this algorithm, two subsequent steps are iterated until there are no more changes in the current hypothesis [23]. In the Expectation-step (E-step) the probability that each observation is a member of each of the chosen class is calculated. Maximization-step (M-step) alters the parameters of each class with the objective to maximize those probabilities. The iteration is then repeated until converging to a (local) optimum.

3. APPLIED METHODOLOGIES

This section presents those data mining techniques used in this study. The analysis process takes four phases. The first phase focuses on data preprocessing, i.e., data cleansing, feature selection, and data transformation. Regression and clustering algorithms are applied in the second and third phase, respectively. This is enhanced to comprise six different variables, (R, F, M, P, Q and C). The proposed variables PQ, the P variable represents the average purchase power per consumer per all transactions. The Q variable represents the average purchase power per all products purchased per consumer. The C is the Consumer Change Rate (a trend) that shows the quantity and sign of a change of consumer purchase behavior. The enhanced regression algorithm finds this trend. The output is then fed into the clustering algorithms with RFMPQ data. Fuzzy c-means (FCM) and EM clustering algorithms are used for the segmentation. At the final phase, the accuracy of these partitions is measured by the cluster quality assessment introduced by Drăghici [13].

The point-of-sales database consists of all product sales and shows that the client sells diverse products like clothing, shoes, schools supplies, and accessories. Transaction data from three years are retrieved and extracted. Each transaction represents a purchase event, with each line consisting of a transaction cashier, store code, item code, brand code, products (quantity) sold, unit price and the total (quantity multiplied by the unit price), date and time of the purchase and information about the consumer. Both active and inactive consumers are included in the study. This is vital to enable the marketing team to develop appropriate marketing and retention campaigns.
3.1 Phase 1: Data Pre-Processing

To prepare for this stage, several interviews are conducted with marketing experts, sales directors, the IT manager, in-store employees, and POS engineers. Interviews intend to maximize variation in responses to gain a deep understanding on the challenges experienced at the client company in general and, more specifically, to find information and company insights about the retail industry, the market, and the consumer base.

3.1.1 Data Transformation Using Enhanced RFM Model

Based on the client information, RFM values are assigned. The latest purchase date of the consumer R is found from a set of 1095 days (records from 2013 to 2016), a number of transactions during this 1095-day period F comes from totally 11550 transactions, and total amount purchased M comes from the total sales of 1,300,000 KD. The RFM attributes are weighted with category 10. The data is extracted from a POS database and fed into a unified, generic POS data warehouse in a common format with consistent definitions for keys and fields. In this phase, the string variables must be converted to numeric variables. The missing values are checked and deleted or replaced by default or mean values of each parameter.

The rest of the data preprocessing phase handle noisy data, missing values and makes attributes reduction and transformation. In this step, the data must be transformed into an appropriate format, to make the discovery of patterns easier. Continuous consumer-related attributes are encoded by decreasing the original values into a small number of value ranges. The age of the consumer was encoded into six categories.

Gender attribute was encoded as 1 for Male, 2 for Female and 3 for Companies. Furthermore, demographic concepts like cities and countries are replaced by higher level concept nationality. For normalizing the RFMPQ score, uses the following rescaling method (2)

\[
RFMPQ_{\text{score}}^\text{scaled} = \frac{RFMPQ_{\text{score}} - \min(RFMPQ_{\text{score}})}{\max(RFMPQ_{\text{score}}) - \min(RFMPQ_{\text{score}})}
\]

3.1.2 Phase 2: Steps of Data Transformation to RFMC

As usual, the client company has limited resources, and they are not able to serve all consumers with the same intensity. To manage the available resources more efficiently, the client wants to distinguish between relatively low and high shoppers, choose only profitable market segments and concentrate all efforts on the strategic value of these segments to increase profitability and consumer retention. This study decided to incorporate RFMPQ scores with CLV matrix before running the modified regression algorithm to generate the C data set. The CLV matrix will divide the RFMPQ data set into four categories.

The CLV matrix calculates the number of purchases by taking the total number of purchases for the consumer and then dividing by the total number of consumers in the consumer database. The average purchase amount is derived by taking the total revenue and dividing it by the total number of purchases. Comparing the average number of purchases, F between consumers and the average purchase amount, M with total average values is the next step. M and F are used to classify each consumer into one of the four fields Best, Spender, Frequent and Uncertain.
3.1.3 Enhanced Regression Algorithm to Generate Consumer Purchase Curve (C Data Set)

This study proposes a combination of two analytical data mining steps. For finding C, the change in consumer purchase behavior, it uses supervised linear regression method. Then C dataset is then put into the unsupervised clustering algorithms, to split the consumers into different groups based on pattern dissimilarities. The variable C should answer a very important question if the consumer is at high risk of shifting his or her service to another retailer. One of the most common indicators of high-risk consumers is a drop off in purchase power and a decrease of visits.

A major limitation of RFM and market segmentation models is that they ignore behavioral changes of consumers during the period of analysis. Although the recency parameter is one of the indicators of such behavior, it suffers from the transient behavior of the consumer. Therefore, introducing a new analysis parameter is essential for retailers to narrow down a group of consumers at high risk or consumers that should be prompted to a higher category. From a retailer perspective, purchase power differs from one consumer to the other. If the amount of purchase power decrease continuously, then the consumer is on the verge of shifting his or her services to another retailer or falling from the beneficial segment to the non-beneficial segment. Similarly, a consumer with an increasing purchase power during the period of analysis shows the potential that could be harnessed with appropriate marketing actions.

To capture this change of behavior, we first calculate purchase amounts of all consumers in each selected period of analysis. The parameter C for time k is defined as

\[ C(k) = \begin{cases} \frac{p_a_k - p_a_{k-1}}{p_a_{k-1}}, & p_a_{k-1} \neq 1 \\ 1, & \text{else} \end{cases} \]  

where \( p_a_k \) indicates the total amount purchased by a consumer at time k.

If the data is divided into \( n+1 \) analysis periods, \( n \) changes are calculated. If the change rate values of the latest analysis periods have the same sign (negative or positive), then the average of these values will be used as the final change rate parameter C for each the consumer. Consumers changing signs are assigned neutral \( C = 0 \). The final set consists of positive, negative and neutral change rate values.

Furthermore, since the C dataset consists of numbers with six or more digits, the logarithm of C is useful to obtain normalized C data the analysis database.

\[ C_{\text{scaled}} = \log_b C, b>0, b\neq 1 \]  

The next step is to apply the modified best-fit regression algorithm on each segment separately by using the demographic variables Age, Gender and Nationality. Then the purchase behavior change rate C values are prepared for market segmentation by applying clustering algorithms for every segment.

The solution starts with the calculation of the purchase amount trend of the regression algorithm in the time series of the known \( k \) and known \( M_k \), purchases at \( k \). The result is \( m_1 \), the slope of the linear regression curve.
\[ M_k = m_0 + m_1 k + \varepsilon_k, \quad (5) \]

where \( m_0 \) is the intercept and \( \varepsilon_k \) is the random parameter.

The \( m_1 \) rate determines the expected value and sign of the change rate \( C \), based on past purchases. To obtain the regression curve the purchase slope discount rate \( w_k \) to modifies the effect of the purchase \( M_k \) for each period \( k \). More recent spending gets a higher rate. For instance, for a consumer with 4-time periods in the analysis, with purchase slope discount rate 0.7, the latest time \( M_{k-1} \) is obtained by multiplying the actual purchase amount by \( w_{k-1} = 0.7^1 \), \( M_{k-2} \) by \( w_{k-2} = 0.7^2 = 0.49 \), and so on. This method to compute the total purchase amount curve (slope) leads to decreasing effect from older purchase and captures the alternative cost of consumer capital.

Next, we introduce a new computed parameter named Consumer Purchase Curve (CPC). the (CPC) parameter multiplies the \( M_k \) and \( k \) to obtain the correct calculated purchase curve (slope).

The final formula is shown in Equation (6).

\[ \text{(cpc)} \ P = \frac{n \sum (\Delta k \Delta M) + (\sum \Delta k)(\sum \Delta M)}{n \sum \Delta k^2 - (\sum \Delta k)^2} \quad (6) \]

Next, applies the ND Y3, to remove the Outliers extracted from the dataset.

Here, \( Y3 \) represents the Normal Distribution in the newly modified equation.

\[ \Delta \bar{k} = \Delta k \times Y3 \quad \text{and} \quad \Delta \bar{M}_k = \Delta M_k \times Y3. \]

Figure 3. Sample of customers Purchase history

Sample of one consumer’s purchase history using different normal distribution values. As shown in Fig 3 we have different shapes indicating the percentage of outliers in the dataset by using different variable for each consumer.

3.2 Phase 4: Market Segmentation Using Clustering Algorithms

3.2.1 Fuzzy C-Means Soft Clustering

Fuzzy c-means (FCM) is unsupervised Clustering of numerical data method developed by Dunn in 1973 [41]. The algorithm allows one piece of data corresponding to two or more clusters with the
same weight, and it uses the concepts from the field of fuzzy set theory and fuzzy logic [41]. The FCM is frequently used in pattern recognition and employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1 [42]. The Fuzzy c-mean algorithm is composed of the following steps

**Step 1.**

Initialize \( U = [x_{ij}] \) matrix, \( U^{(0)} \) \hspace{1cm} (7)

**Step 2.**

At k-step: calculate the centers vectors

\[
C^{(k)} = [c_j] \quad \text{with} \quad U^{(k)}
\]

\[
9C_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}}
\]

**Step 3.**

Update \( U^{(k)}, U^{(k+1)} \)

\[
d_{ij} = \sqrt{\sum_{i=1}^{n} (x_i - c_j)^2}
\]

**Step 4.**

\[
u_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}}
\]

**Step 5.**

if \( || U^{(k+1)} - U^{(k)} || < \epsilon \) then STOP; otherwise return to step 2

(11)

the m is any real number greater than 1, \( u_{ij} \) is the degree of memberships of \( x_i \) in the cluster \( j \), \( x_i \) is the \( ith \) of d-dimensional measured data, \( c_i \) is the d-dimension center of the cluster. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. After each iteration membership and cluster centers are updated according to the formula.

### 3.2.2 EXPECTATION-MAXIMIZATION (EM) HARD CLUSTERING

The Expectation-Maximization (EM) is an iterative estimation algorithm developed by Dempster, Laird, and Rubin [11]. EM is historically very important algorithm in market segmentation and
data mining. EM algorithm has also proven its efficiency in a good performance, decreasing sensitivity to noise and estimation problem involving unlabeled data.

**Step 1. Expectation- Maximization (EM) Clustering Initialization, the E-step**

Every class $j$, of $M$ classes (or clusters), is formed by a vector parameter ($\theta$), composed by the mean ($\mu_j$) and by the covariance matrix ($P_j$), which defines the Gaussian probability distribution (Normal) features used to characterize the observed and unobserved entities of the data set as shown in Equation (11)

$$\theta(t)=\left(\mu_j(t), P_j(t)\right), j=1,\ldots,M.$$  

(12)

On the initial instant ($t=0$) the implementation can generate randomly the initial values of mean ($\mu_j$) and of the covariance matrix ($P_j$). The EM algorithm aims to approximate the parameter vector ($\theta$) of the real distribution.

Fraley and Raftery [18] suggest another alternative to initialize (EM) with the clusters obtained by a hierarchical clustering technique. The relevance degree of the points of each cluster is given by the likelihood of each element attribute in comparison with the attributes of the other elements of cluster $C_j$ as shown in Equation (12) The E-step

$$C_j|x=\frac{\sum_{k=1}^{N} \frac{1}{2} \log\det P_j(t)(x_k-\mu_j(t))}{\sum_{k=1}^{M} \sum_{j} \frac{1}{2} \log\det P_k(t)}.$$  

(13)

**Step 2. M-Step**

First is computed the mean ($\mu$) of class $j$ obtained through the mean of all points in function of the relevance degree of each point, as shown in Equation (13)

$$\mu_j(t+1)=\frac{\sum_{k=1}^{N} P(C_j|x_k) x_k}{\sum_{k=1}^{N} P(C_j|x_k)}.$$  

(14)

To compute the covariance matrix for the next iteration the with the Bayes Theorem, $P(A|B) = P(B|A) * P(A) / P(B)$ conditional probabilities of the class occurrence are calculated, as shown in Equation (14)

$$\Sigma_j(t+1)=\frac{\sum_{k=1}^{N} P(C_j|x_k) (x_k-\mu_j(t))(x_k-\mu_j(t))}{\sum_{k=1}^{N} P(C_j|x_k)}.$$  

(15)

The probability of occurrence of each class is computed through the mean of probabilities ($C_j$)

$$P_j(t+1)=\frac{1}{N} \sum_{k=1}^{N} P(C_j|x_k).$$  

(16)

**Step 3. Cluster Convergence**

After performing each iteration, a convergence inspection which verifies if the difference of the attributes vector of an iteration to the previous iteration is smaller than an acceptable error tolerance, given by parameter.
4. Analysis Results and Discussion

A sample data of the generated segment Best is shown in tables Table 2 and include information such as the number of samples for each cluster, the average purchase change rate of the cluster as well as the average age, gender and nationality of the cluster.

The Fuzzy c-means (FCM) analysis shows cluster 4 is the best cluster with positive purchase slope (37.12), while cluster 3 (-44.41) has the worst negative purchase slope.

Table 3 shows the four clusters generated by EM. The analysis shows cluster 2 is the least cluster with negative purchase slope (+3.82) and it is the most beneficial segment because it is superior to the other clusters 2. The analysis indicates that consumers between the age of (18-24) to (25-34), 3. Average Gender (1.74) indicates that Female consumers and a very small percentage of Male consumers and 4. Average Nationality (9.11) citizens of Kuwait and Saudi Arabia are best consumers in this cluster. The analysis also shows cluster 1 (-150.98) has the worst negative purchase slope.

4.1 Accuracy and Effectiveness Determination (Inter-Cluster Distance)

Consumer dataset can be clustered in many ways, but how can one know the clusters are accurate and meaningful? A way to answer this is here suggested. Clustering has become a key technique in analyzing quality assessment in a variety of recent studies. There are several studies supporting suggestions for measuring the similarity between clustering algorithms. Those measures are used to compare how accurate different clustering algorithms are on a dataset. Accuracy is usually tied to the type of benchmark being considered. The approach of Draghici [13] is to compare the size of the clusters vs. the distance to the nearest cluster (the inter-cluster distance vs. the size (diameter) of the cluster). That is the distance between the members of a cluster and the cluster’s center, and the diameter of the smallest sphere containing the cluster. If the inter-cluster distance is much larger than the size of the clusters, then the clustering method is trustworthy. Table 4 shows that the quality can be assessed simply by looking at cluster diameter. The cluster is created by means of a heuristic even when there is no similarity between clustered patterns. This is occurring because the algorithm forces K clusters to be created. Comparing both algorithms using cluster quality assessment then the Best segment shows EM clustering with the size of the cluster (2.45967329) is more accurate than the Fuzzy c-means (FCM) algorithm with the size of the cluster (0.001661999).

4.1.1 RFMC Dataset Summary

According to the EM analysis results of the Best Segment, all four clusters in the Best segment have a negative purchase slope, with a total average purchase slope of (-147.82). This ranks the Best segment as the highest negative purchase curve slope. In the Spender segment, the EM analysis results show the effectiveness of the new adapted target marketing strategy. Showing a positive purchase curve (slope), yet the total average purchase slope is (-16.10), this is the segment with the biggest churn rate. The Frequent segment EM analysis results show cluster 2 has a positive purchase slope with (7.17), but the total average purchase slope is (-6.01). This ranks the Frequent Segment slightly above the Spender Segment with respect to the purchase curve (slope). In Fuzzy c-means (FCM) analysis, the Uncertain segment has the Best positive purchase slope, with a total average purchase slope of (1.78). The positive purchase slope is based on smaller purchases. Young females from the age 20 – 40 and citizens of the GCC region are the biggest
consumers with negative purchase slope. Consumers at high risk of canceling their services or falling to a less beneficial segment are easily identified based on the analysis from the C dataset. The retailer can develop a more targeted retail service model to retain such profitable consumers. These valuable insights are derived only from this RFMC method identifying consumer purchase trends.

Data analysis, though often rushed, is the most important stage in the market segmentation solution. The analysis is here conducted under the supervision of internal marketing and sales manager to identify which variable and segment make the most sense to focus all efforts on. Such a framework identifies the model strengths and weaknesses, with special attention paid to all implications stemming from each. The internal marketing and sales managers possess an understanding of the client capabilities and resources. It helps the client to focus on the consumer and develop marketing mixes for a very specific market segment.

Detailed description and specification of all segments in our case study are presented as well as useful strategies as proposed by the marketing experts.

Table 3 Four Clusters (Cl) Based On The Best Segment Generated By Fuzzy C-Means (FCM) Algorithm

<table>
<thead>
<tr>
<th>Cl</th>
<th>Customers</th>
<th>Purchase Change Rate</th>
<th>Av. Age</th>
<th>Av. Gender</th>
<th>Av. Nationality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>95</td>
<td>-29.01</td>
<td>3.22</td>
<td>2.03</td>
<td>2.43</td>
</tr>
<tr>
<td>2</td>
<td>47</td>
<td>-7.41</td>
<td>3.77</td>
<td>2.2</td>
<td>2.62</td>
</tr>
<tr>
<td>3</td>
<td>46</td>
<td>-35.19</td>
<td>3.13</td>
<td>2.03</td>
<td>11.02</td>
</tr>
<tr>
<td>4</td>
<td>73</td>
<td>59.24</td>
<td>3.02</td>
<td>2</td>
<td>7.11</td>
</tr>
<tr>
<td>Tot.</td>
<td>261</td>
<td>-32.37</td>
<td>85</td>
<td>2.065</td>
<td>5.795</td>
</tr>
</tbody>
</table>

Table 4 Four Clusters (Cl) Based On The Best Segment Generated By EM Algorithm.

<table>
<thead>
<tr>
<th>Cl</th>
<th>Customers</th>
<th>Purchase Change Rate</th>
<th>Av. Age</th>
<th>Av. Gender</th>
<th>Av. Nationality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>54</td>
<td>-150.98</td>
<td>3.80</td>
<td>1.09</td>
<td>2.37</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>-3.52</td>
<td>2.53</td>
<td>1.04</td>
<td>1.77</td>
</tr>
<tr>
<td>3</td>
<td>233</td>
<td>-199.13</td>
<td>3.81</td>
<td>1.94</td>
<td>17.83</td>
</tr>
<tr>
<td>4</td>
<td>580</td>
<td>-128.54</td>
<td>3.41</td>
<td>2.00</td>
<td>2.95</td>
</tr>
<tr>
<td>Tot.</td>
<td>865</td>
<td>-147.82</td>
<td>3.62</td>
<td>1.92</td>
<td>7.08</td>
</tr>
</tbody>
</table>
4.2 SUMMARY

This study uses two methods:

1. Draghici approach is to compare the size of the clusters vs. the distance to the nearest cluster.
2. Industry marketing experts as human judgment to validating the accuracy and intelligence of the results.

Analysis and experts agree that classifying consumer purchase behavior using a CLV matrix against normally distributed enhanced RFMPQ Dataset gives high accuracy about the entire consumer's journey information. The C variant takes into consideration the changes in the consumer average purchase power over the time compared to the standard RFM model. Age and gender variables show an estimated accuracy of 83%. However, the nationality variable gives low accuracy, possibly because of null data removed by the normal distribution formula.

5. CONCLUSION, CONTRIBUTION, AND ANSWERS TO THE STUDY QUESTIONS

In this study, we proposed new RFM variants (PQC) to the standard RFM model to segment a POS dataset using demographic variables. Suggested is a novel step by step approach with CLV and RFMPQC analysis in two data mining tasks. Regression and clustering methods apply separately for every RFMPQ variation. Treating consumers with different reflected purchase...
behavior is the objective of this study. The proposal is to make this into a computable parameter a
new modified regression algorithm is also proposed in this study.

To help mapping the entire consumer's journey, in this study existing consumer purchase profile of
demographic variables like age, gender and nationality are segmented and presented accordingly.
The study has shown that integrating CLV and enhanced RFM methods provide a credible base for
capturing consumer’s purchase trend and understanding market segmentation with different values
of Frequency, Monetary, average purchase power per consumer per all transactions, average
purchase power per all products purchased per consumer and Purchase Change Rate. The model
using the proposed techniques can identify VIP consumers who already shifted or at the high risk
of shifting their business to another competitor or those consumers falling from a higher segment
to a lower segment.

Two clustering algorithms were tested in this study. Hard Clustering EM, and soft clustering
Fuzzy c-means. The experiment showed Fuzzy c-means difficulty in selecting the initial cluster
centers and requires more computation time as it requires more iterations than the EM algorithm.
However, shows lower clustering errors. It is more efficient for large scale clustering, while the
EM is more suitable for small scale clustering. The model provides simplicity. However, the
enhanced regression algorithm failed to accurately measure consumer final purchase trends
(Slope) over the selected periods of time. For example, if a consumer with increasing (Positive)
purchases, and the most recent purchase made by the consumer is lower than the previous
purchase, then the final curve (slope) is negative, regardless of the consumer increasing purchase
history. This indicates that the most recent purchase has the final effect on the consumer purchase
curve (slope). The enhanced regression does not strengthen the effect of the most recent purchase
in the computation of the total purchase amount slope while justifying the importance of previous
purchase slopes. The second contribution is a novel market segmentation method using clv to
classify consumers into four quadrants (Best, Spender, Frequent, Uncertain), and enhanced
normal distribution formula to remove outliers in data and enhanced regression based on the
enhanced RFMPQC variables followed by clustering on demographic data. The study provides
key attributes describing consumer purchase behavior in different demographic characteristics
such as gender, age, and nationality. The analytical information gained from extracted data is
useful to adapt more targeted marketing strategies and for decision making. We can confirm that
proposed market segmentation using demographic variables can contribute to the body of
literature in consumer purchase behavior and assist SMR retailers in meeting the needs and
preference of consumers.

REFERENCES

Rowman Altamira.


[38] K.M. Bataineh, M.Naji, M.Saquer “A Comparison Study between Various Fuzzy Clustering Algorithms” vol.5 no.4 Aug. 2011


AUTHORS

Fahed Yoseph Business Intelligence and Big Data Consultant, Bachelor in Management information system. Master in computer Science in the area of artificial intelligence. Current doctoral student at Åbo Akademi University in Finland. Research interest in Computer Science, database and artificial intelligence.

Mohammad AlMalaily Financial, Marketing and BI Consultant, Bachelor in Electronic Engineering.University in Finland. Research interest in Computer Science, Marketing and Finance.