Abstract

In the literature there are some proposals for integrated schemes for campaign management based on segmentation from the results of the RFM model. RFM is a technique used to analyze customer behavior by means of three variables: Recency, Frequency and Monetary value. It is very much in use in the business world due to its simplicity of use, implementation and interpretability of its results. However, RFM applications to campaign management present known limitations like the lack of precision because the scores of these variables are expressed by an ordinal scale. In this paper, we propose to link customer segmentation methods with campaign activities in a more effective way incorporating the 2–tuple model both to the RFM calculation process and to its subsequent exploitation by means of segmentation algorithms, specifically, k-means. This yields a greater interpretability of these results and also allows computing these values without loss of information. Therefore, marketers can effectively develop more effective marketing strategy.

Keywords

2-Tuple Model, Campaign Management, Relational Strategy, RFM.

I. Introduction

Most marketers have difficulty in identifying the right customers to engage in successful campaigns. Customer segmentation is a popular method that is used for selecting appropriate customers for a launch campaign. In order to link customer segmentation methods with campaign activities, Chan [1] presents an approach that combines customer targeting and customer segmentation for campaign strategies based on RFM model.

RFM (Recency, Frequency and Monetary) is a model used to analyze customer behavior proposed by Hughes in 1994 [2]. “Recency” represents the length of a time period since the last purchase, while “Frequency” denotes the number of purchases within a specified time period and “Monetary value” means the amount of money spent in this specified time period. The RFM models were developed as a logical step in the evolution of marketing segmentation techniques. When the shotgun approaches (marketing everything to everyone) proved inefficient in terms of returns, the marketing campaigns started separating customers in segments based on socio-demographics attributes [3]. Commonly, RFM methods have been used to measure the Customer Lifetime Value (CLV), i.e., the predicted value a customer is going to generate in his entire lifetime [4]-[6].

With RFM analysis, organizations could discover these most valuable customers easily by observing their past behaviors [7]. In fact, these three variables belong to behavioral variables and can be used to make predictions based on the behavior in the transactional database [8]. Therefore, in a RFM process, the goal is to obtain the customer purchase behavior (the most loyal customers, dormant customers…) from these transactional data to proactively trigger appropriate direct marketing actions (retention, reactivation campaigns…) [9], [10].

In recent years, more sophisticated statistical and data-mining techniques have been employed in direct marketing field: chi-squared automatic interaction detection (CHAID), logistic regression, neural network models, etc. Despite the deployment of these methods, marketers continue to employ RFM models. There are several reasons for the popularity of RFM among which the following are worth mentioning [11]: It is easy to use; it can generally be implemented very quickly; and it is a method that managers and decision makers can understand.

McCarty and Hastak [11] have compared RFM, CHAID, and logistic regression as analytical methods for direct marketing segmentation, using two different datasets. It turns out that CHAID tends to be superior to RFM when the response rate to a mailing is low and the mailing would be send to a relatively small portion of the database. However, RFM is an acceptable procedure in other circumstances.

RFM approaches present known limitations like the lack of precision. Indeed, the scores of these RFM variables are expressed by an ordinal scale. The most common scale is the set {1,...,5} that refer to the customer contributions to revenue for enterprises. The 5 refers to the most customer contribution to revenue and 1 refers to the least contribution to revenue [12].

On the other hand, the fuzzy linguistic approach is a tool intended for modeling qualitative information in a problem. It is based on the concept of linguistic variable [13] and has been satisfactorily used in many problems [14]-[17]. The 2-tuple fuzzy linguistic approach is a model of information representation that carries out processes of “computing with words” without the loss of information [18] that has been widely used in many business and management applications [19]-[25].
In this paper, we propose to link customer segmentation methods with campaign activities in a more effective way incorporating the 2-tuple model both to the RFM calculation process and to its subsequent exploitation by means of segmentation algorithms, specifically, k-means. This yields a greater interpretability of these results and also allows computing these values without loss of information. Therefore, interpreting these linguistic results, decision makers can effectively identify valuable customers and consequently develop more effective marketing strategy. Additionally, we present an IBM SPSS Modeler implementation of this model. This enables us to be more applicable at the practical level and not remain solely confined to the theoretical one.

The rest of the paper is organized as follows: Section II revises the preliminary concepts, i.e., the integrated scheme of customer segmentation with campaign activities using the RFM model and 2-tuple model. In Section III we propose to modify this integration scheme by incorporating the 2-tuple model in two directions: in the RFM scores computation and the subsequent segmentation algorithm. Additionally we show an implementation and use case of this new model using IBM SPSS Modeler comparing it with the previous one. Finally, we point out some concluding remarks and future work.

II. Preliminaries

In this section we present the basic elements needed to understand our new proposal: an integrated scheme of customer segmentation with campaign activities based on the RFM model and the 2-tuple fuzzy linguistic approach.

A. Integrated Scheme of Customer Segmentation with Campaign Activities

The RFM analytic approach is a common model that identifies customer purchase behavior, i.e., that differentiates important customers from large data by three variables [9]:

- **Recency** (R): The time (in units such as days, months, years…) since the most recent purchase transaction or shopping visit.
- **Frequency** (F): The total number of purchase transactions or shopping visits in the period examined.
- **Monetary** value (M): The total value of the purchases within the period examined.

In order to link customer segmentation methods with campaign activities [1] the following scheme that integrates the RFM model (Fig. 1):

\[
RFM_{Score} = Reancy_{Score} \times w_R + Frequency_{Score} \times w_F + Monetary_{Score} \times w_M.
\]

5. **Segment.** Once the results of the previous step are validated, the marketers of the enterprise apply this RFM knowledge in order to search the most suitable customer group for each plan campaign. For this, segmentation or clustering techniques are especially useful. Clustering or segmentation is the process of grouping a set of objects into groups of similar objects. In this way, clustering based on RFM scores of the table CustomerRFM provides more behavioral knowledge of customers' actual marketing levels than other cluster analyses [27].

K-means is one of the well-known algorithms for clustering [28], [29] of which various modifications have been proposed including fuzzy logic [30]. In this algorithm each cluster is characterized by its center point i.e. centroid. K-means is a partitioning cluster algorithm by grouping n vectors (customers in our case) based on attributes
into \( k \) partitions, where \( k < n \), according to some measure, usually Euclidean distance. The name comes from the fact that \( k \) clusters are determined and the center of a cluster is the mean of all vectors within this cluster. The algorithm starts with \( k \) initial centroids, then assigns vectors to the nearest centroid using Euclidean distance and re-computes the new centroids as means of the assigned data vectors. This process is repeated over and over again until vectors no longer changed clusters between iterations \([27],[31]\). Thus, using a \( k \)-means algorithm, the centroid results of this algorithm are: \( v_i = (v_{i1}, v_{i2}, v_{is}), \) with \( s = 1..k \), one for each cluster. These centroids are quite interpretable from the point of view of business as explained in the previous stage (5 best values and 1 the worst).

6. Target. It is necessary to identify the most profitable groups of customers for each campaign plan. Thus, once the segmentation is concluded and validated, marketers should determine the targeted clusters that can be associated with the subsequent campaign and then get the clients that belong to those groups which are stored in the table CustomerTarget (CustomerID, CampaignID).

7. Action. The last step is to implement effective campaign management oriented selected target. Much of aforementioned approach can be solved with several data science or data mining tools. In Fig. 2 we show an example using IBM SPSS Modeler \([26]\).

![Fig. 2. Integrated scheme of customer segmentation with campaign activities with IBM SPSS Modeler.](image)

Following, we explain each stage of this stream:

1. **Plan.** Based on Ref. \([1]\), in Table I we present the set of campaigns to be carried out.

<table>
<thead>
<tr>
<th>CampaignID</th>
<th>RelationalStrategyDES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Best</strong></td>
<td>They are the most valuable clients for the company. The relational strategy will be aimed at managing the relationship in order to maintain the value of the clients. For example, we could offer them free services.</td>
</tr>
<tr>
<td><strong>New/Reactivates</strong></td>
<td>They are new clients or ex-clients that have been reactivated in the period analyzed. Welcome gifts, bonus... could be offered associated with the next purchase.</td>
</tr>
<tr>
<td><strong>Growing</strong></td>
<td>They are a low-value client which it is necessary to carry out a growth strategy by up-selling or cross-selling.</td>
</tr>
<tr>
<td><strong>Churn</strong></td>
<td>The customers who are possible to leave or turn to other competitors. The strategy could be retention by means special discounts...</td>
</tr>
<tr>
<td><strong>Worse</strong></td>
<td>They are clients with a minimum degree of relationship with the company so they can be considered ex-clients. The relational strategy will be similar to that of leads, that is, acquisition. Free trial could be offered for example.</td>
</tr>
</tbody>
</table>

2. **Data collect and preparation.** Transactional data of the two last years were retrieved, audited, cleaned, and prepared (casting to date type) for subsequent stages. Inactive customers with no purchases during this period were not included into the Transactions table. We have based this example on a file obtained based on data referenced in \([32]\) with 69659 purchase transactions, corresponding to \( n = 23570 \) distinct customers.

3. **RFM aggregation.** For this step we use the RFM Aggregate node (labeled as RFM Aggregate in Fig. 2) that simplifies the computation of this stage. We only have to designate the required transaction fields (CustomerID, Date and Amount) and the fixed date to compute Recency as the time of difference (days, hours, minutes or seconds) between Date and this date (2018-01-01), see Fig. 3.

![Fig. 3. Detail of the RFM Aggregate node settings.](image)

4. **RFM scores computation.** IBM SPSS Modeler also offers a node named RFM Analysis (called RFM scores computation in Fig. 2) that can directly group the R, F, and M measures into the selected number of quantiles (five in our case). This node also computes the \( RFMScore \) using the Eq. (1) with \( w_R=1/3, w_F=1/3 \) and \( w_M=1/3 \), see Fig. 4.

![Fig. 4. Detail of the RFM Analysis node settings.](image)

Before performing segmentation using scores may address certain plan campaigns, for example, in order to identify the most valuable customers (campaign identified by Best in Table I). This could be solved with the RFM model sorting in descending order by the field \( RFMScore \). As can be seen in Table II, there are many clients with equal score because when grouping customers in quintiles the procedure results in a total of \( 5 \times 5 \times 5 = 125 \) distinct values as much of \( RFMScore \). This lack of precision can be a problem when selecting customers for the different campaigns.

<table>
<thead>
<tr>
<th>CustomerID</th>
<th>Recency</th>
<th>Frequency</th>
<th>Monetary</th>
<th>RFM Score</th>
<th>RFMScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23524</td>
<td>7153</td>
<td>15</td>
<td>105.950</td>
<td>5.950</td>
</tr>
<tr>
<td>2</td>
<td>8758</td>
<td>7243</td>
<td>26</td>
<td>394.150</td>
<td>5.759</td>
</tr>
<tr>
<td>3</td>
<td>10096</td>
<td>7199</td>
<td>42</td>
<td>630.640</td>
<td>5.654</td>
</tr>
<tr>
<td>4</td>
<td>7617</td>
<td>7066</td>
<td>4</td>
<td>240.800</td>
<td>5.536</td>
</tr>
<tr>
<td>5</td>
<td>9769</td>
<td>7251</td>
<td>26</td>
<td>102.545</td>
<td>5.125</td>
</tr>
<tr>
<td>6</td>
<td>9982</td>
<td>7186</td>
<td>34</td>
<td>477.410</td>
<td>5.072</td>
</tr>
<tr>
<td>7</td>
<td>9856</td>
<td>7133</td>
<td>27</td>
<td>377.410</td>
<td>5.067</td>
</tr>
<tr>
<td>8</td>
<td>9988</td>
<td>7163</td>
<td>116</td>
<td>255.140</td>
<td>4.977</td>
</tr>
<tr>
<td>9</td>
<td>1015</td>
<td>7121</td>
<td>23</td>
<td>113.460</td>
<td>4.937</td>
</tr>
</tbody>
</table>

**TABLE II.**

**Detail of the CustomerRFM Table Ordered by RFM Score Descending (the First 2640 Clients Have the Same RFMScore = 5)**
information by means of 2-tuple \((s_i, \alpha)\), \((s_i, \alpha)\) is a numerical value and 2-tuple: expressing the value of the translation from the original result to the closest index label, \(\alpha\) is the value of the symbolic translation. For all \(\Delta\), there exists \(\Delta^4\), defined as:

\[
\Delta^4(s_i, \alpha) = i + a.
\]

The negation operator is defined as:

\[
\neg(s_i, \alpha) = \Lambda(T - \Delta^1(s_i, \alpha)).
\]

Information aggregation consists of obtaining a value that summarizes a set of values. Hence, the result of the aggregation of a set of 2-tuples must be a 2-tuple. Using the functions \(\Delta\) and \(\Delta^4\) that transform numerical values into linguistic 2-tuples and vice versa without loss of information, any of the existing aggregation operators can be easily extended for dealing with linguistic 2-tuples. Below, we describe the aggregation operators which we use in our model:

**Definition 3.** Let \(A = \{(l_1, a_1),..., (l_n, a_n)\}\) be a set of linguistic 2-tuple and \(W = \{w_1, ..., w_n\}\) be their associated weights. The 2-tuple weighted average \(\bar{A}\) is:

\[
\bar{A} = \left(\frac{\sum_{i=1}^{n} a_i \cdot w_i}{\sum_{i=1}^{n} w_i}\right)
\]

**Definition 4.** Let \(A = \{(l_1, a_1),..., (l_n, a_n)\}\) be a set of linguistic 2-tuple. The 2-tuple average \(\bar{A}\) is:

\[
\bar{A} = \left(\frac{\sum_{i=1}^{n} a_i}{n}\right)
\]

### III. Applying the 2-Tuple Approach for Campaign Management

As explained in Section II.A, although RFM analysis is a very useful tool for campaign management, it has its limitations such as its lack of precision in the calculation of scores. This is due to the representation as an ordinal number of these RFM scores (for example, see the Table II where you cannot identify which are really the best customers). In this section, we propose to incorporate the 2-tuple model in order to improve this campaign management. This is possible because by incorporating the 2-tuple model we will get results from the RFM model with more linguistic interpretability and above all with more precision.

The campaign management scheme followed is the same as shown in Fig. 1 and explained in Section II.A but changing stages 3 and 4 as explained in the following two sub sections in which we also show the implementation in SPSS Modeler of the model:

**A. RFM Scores Computation**

The basic idea is to compute and store the scores included into the output table of this step (CustomerRFM), i.e., RecencyScore, FrequencyScore, MonetaryScore and RFMScore using the 2-tuple model [25].

First, we need to define the symmetric and uniformly distributed domain \(S\) using five linguistic labels. These labels have a semantic
meaning for these four variables of the RFM model referred to the degree of agreement on the goodness of the variable:

Let $S = \{s_0, ..., s_4\}$, $T = 4$; $s_0 = Strongly\ Disagree = SD, s_1 = Disagree = D, s_2 = Neutral = N, s_3 = Agree = A,$ and $s_4 = Strongly\ Agree = SA$, with the definition showed in Fig. 5.

Therefore, we have the variables to calculate: Recency Score, Frequency Score, Monetary Score, RFM Score $\in S \times [-0.5, 0.5]$.

For each customer $i = 1, ..., n$, we obtain $A_i = (A_{i1}, A_{i2}, A_{i3})$ with $A_{i1}$ = Recency Score, $A_{i2}$ = Frequency Score, and $A_{i3}$ = Monetary Score.

Firstly, customers are sorted in ascending order according to each of the individual RFM components $B_i = (B_{i1}, B_{i2}, B_{i3})$, with $B_{i1}$ = Recency, $B_{i2}$ = Frequency, and $B_{i3}$ = Monetary, stored in CustomerTransactions (obtained as explained in phase 3 of Section II.A). Now, we define $\text{rank}_i \in \{1, ..., n\}$ as the ranking of each client respect to each of these variables:

$$\text{percent\ rank}_i = \frac{\Delta(\text{percent\ rank}_0)}{(n-1)}$$

with $\text{percent\ rank}_0 \in [0, 1], i = 1, ..., n, j = 1, ..., 3$ and $n > 1$. The final 2-tuple score $A_{ij}$ is obtained as following:

$$A_{ij} = \begin{cases} 
\Delta(\text{percent\ rank}_0), & \text{if } j \neq 1 \\
\neg\Delta(\text{percent\ rank}_0), & \text{if } j = 1 
\end{cases}$$

(7)

where $\Delta(\cdot)$ and $\neg(\cdot)$ have been defined in Section II.B (Eq. 2 and 4 respectively). We use the negation function on Recency because the larger scores represent the most recent buyers.

The 2-tuple RFM Score, which characterizes together the RFM scores, is calculated for each $i$-customer using the Eq. (5) as follows:

$$\text{RFM Score}_i = \tilde{A}^3[A_{ij}]$$

(8)

with the user-defined weights $W = \{w_R, w_F, w_M\}$.

In a previous paper [24] the authors have proposed both a representation data type 2-tuple as the implementation of the functions $\Delta$ and $\Delta^1$ using IBM SPSS Modeler. Using these tools, the 2-tuple approach proposed in this paper has been implemented. Thus, the stream to solve the example of the Section II.A is showed in the Fig. 6.

We execute the new version of the stream (until 4th phase in Fig. 6) on the same input data and the same user-defined weights (wR, wF, wM) used on the conventional stream (Fig. 2). The selection of the best customers, i.e., the highest $\Delta^4(\text{RFM Score})$ is presented in the Table IV. Also in this table we show the RFM scores that were obtained according to the conventional process. In the 2-tuple implementation the interpretability of the scores is easier as they are expressed by linguistic labels instead of ordinal numbers. Also the accuracy of such scores is greater, owing the 2-tuple model, allowing a better prioritization (selection) of the best customer in order to identify the most valuable customers (campaign Best in Table I).

**TABLE IV. RESULTS OF THE CONVENTIONAL RFM PROCESS VS 2-TUPLE RFM PROCESS ORDERED BY VS 2-TUPLE RFM SCORE DESCENDING (TOP 20 CUSTOMERS)**

<table>
<thead>
<tr>
<th>Conventional RFM process</th>
<th>2 tuples RFM process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer</td>
<td>Score</td>
</tr>
<tr>
<td>1000</td>
<td>1576.00</td>
</tr>
<tr>
<td>1005</td>
<td>1564.00</td>
</tr>
<tr>
<td>1010</td>
<td>1552.00</td>
</tr>
<tr>
<td>1015</td>
<td>1540.00</td>
</tr>
<tr>
<td>1020</td>
<td>1528.00</td>
</tr>
<tr>
<td>1025</td>
<td>1516.00</td>
</tr>
<tr>
<td>1030</td>
<td>1504.00</td>
</tr>
<tr>
<td>1035</td>
<td>1492.00</td>
</tr>
<tr>
<td>1040</td>
<td>1480.00</td>
</tr>
<tr>
<td>1045</td>
<td>1468.00</td>
</tr>
<tr>
<td>1050</td>
<td>1456.00</td>
</tr>
<tr>
<td>1055</td>
<td>1444.00</td>
</tr>
<tr>
<td>1060</td>
<td>1432.00</td>
</tr>
<tr>
<td>1065</td>
<td>1420.00</td>
</tr>
<tr>
<td>1070</td>
<td>1408.00</td>
</tr>
<tr>
<td>1075</td>
<td>1396.00</td>
</tr>
<tr>
<td>1080</td>
<td>1384.00</td>
</tr>
<tr>
<td>1085</td>
<td>1372.00</td>
</tr>
<tr>
<td>1090</td>
<td>1360.00</td>
</tr>
<tr>
<td>1095</td>
<td>1348.00</td>
</tr>
<tr>
<td>1000</td>
<td>1336.00</td>
</tr>
</tbody>
</table>

**B. Segment**

The main problem with the previous approach (4th phase in Section II.A) is the lack of precision in the representation of each individual customer (RFM scores). Consequently the results (centroids) obtained in the next stage are also imprecise. In this section, we propose applying the 2-tuple fuzzy linguistic RFM approach to customer segmentation (using k-means) to obtain more accurate results. On the other hand, it will also increase the interpretability of these centroids as we use linguistic values.

The scores obtained in the previous step (Section III.A) are 2-tuple values: $A_i = (A_{i1}, A_{i2}, A_{i3})$ with $A_{i1}$ = Recency Score, $A_{i2}$ = Frequency Score,
and $A_3 = \text{MonetaryScore}_i$.

The objective is to obtain the centroids $v_s = (v_{s1}, v_{s2}, v_{s3})$ with $s = 1..k$, one for each cluster. The values of these centroids will be expressed using model 2-tuple model, thus we get a better linguistic interpretability. In order to apply the algorithm, we need to get the distance between customers and these centroids. We propose to use the Euclidean distance $d_s$ following:

$$d_s(i, s) = \sqrt{\sum_{i=1}^{n} (\Delta^I(A_{si}) - \Delta^I(v_{si}))^2 + (\Delta^2(A_{si}) - \Delta^2(v_{si}))^2 + (\Delta^3(A_{si}) - \Delta^3(v_{si}))^2}$$

for each customer $i = 1..n$ and for each cluster $s = 1..k$.

In each step of the k-means algorithm, we recalculate the new cluster center $v_s$ using the Eq. 6:

$$v_s = (\bar{A}_{s1}, \bar{A}_{s2}, \bar{A}_{s3})$$

with $r = 1..c_s$, that symbolizes the $r$-customer such that belongs to the $s$-cluster.

Following the example used previously, we show the result of our 2-tuple model in Table V after executing the proposed algorithm. With our model the linguistic interpretability of the clusters (centroids) is better (see the linguistic labels included in Fig. 5). But the main advantage is that these results are also more accurate as we have already commented. We can see how the distribution of groups with our model is more equitable than the conventional process, where the largest groups are those that contain the worst and best customers (see Table III).

### TABLE V. RESULTS OF THE 2-TUPLE K-MEANS CLUSTERING

<table>
<thead>
<tr>
<th>S</th>
<th>Recency Score</th>
<th>Frequency Score</th>
<th>Monetary Score</th>
<th>RFM Patter ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_{s1}$</td>
<td>$v_{s2}$</td>
<td>$v_{s3}$</td>
<td>[10]</td>
<td></td>
</tr>
<tr>
<td>cluster-1</td>
<td>3139</td>
<td>A-0.015419</td>
<td>N+0.018895</td>
<td>N+0.004251</td>
</tr>
<tr>
<td>cluster-2</td>
<td>5913</td>
<td>A+0.088279</td>
<td>A+0.111631</td>
<td>A+0.109542</td>
</tr>
<tr>
<td>cluster-3</td>
<td>3276</td>
<td>N-0.078436</td>
<td>A-0.065937</td>
<td>A-0.060768</td>
</tr>
<tr>
<td>cluster-4</td>
<td>5374</td>
<td>D-0.024133</td>
<td>D-0.099303</td>
<td>D-0.115425</td>
</tr>
</tbody>
</table>

Therefore, our model could get a more appropriate and effective campaign plan.

### IV. CONCLUDING REMARKS AND FUTURE WORK

RFM [2] is a technique widely used a lot more now in marketing due to its simplicity of use, implementation and interpretability of its results. Even its results are better, in a practical level, than other more sophisticated techniques as CHAID and logistic regression in specific circumstances [11]. However, RFM applications to direct marketing present known limitations like the lack of precision.

In this context, we have presented an integrated relational campaign management scheme based on RFM analytic process that incorporates the 2-tuple model in order to obtain a higher precision and an easier linguistic interpretability of the RFM model results and the subsequent segmentation, in order to develop a more effective campaign plan.

Additionally, we have presented an IBM SPSS Modeler implementation of this model. In such a way, our proposal could be widely applied at a practical level on several marketing problems of this type. As an example, we have applied the implemented model on a well-known data set verifying the advantages of the new model regarding the conventional campaign management scheme.

We are currently focusing on the use of this model to several marketing problems, especially in banking industry.

**Acknowledgment**

This paper has been elaborated with the financing of FEDER funds in the Spanish National research project (TIN2013-40658-P), Spanish Department for Economy and Competitiveness project (TIN2016-75850-R).

**REFERENCES**


