Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 55 (2015) 1340 – 1347

Procedia Computer Science

Information Technology and Quantitative Management (ITQM 2015)

A 2-tuple Fuzzy Linguistic RFM Model and Its Implementation

Ramón A. Carrasco[a,]*, María Francisca Blasco[a], Enrique Herrera-Viedma[b]

[a]Department of Marketing and Market Research, Complutense University, Madrid 28223, Spain
[b]Department of Computer Science and Artificial Intelligence, University of Granada, Granada 18071, Spain

Abstract

RFM is a model used to analyze the behavior of customer by means of three variables: Recency, Frequency and Monetary. The scores of these variables are expressed by an integer number, typically, in the range 1..5. The fuzzy linguistic approach is a tool intended for modeling qualitative information in a problem. In this paper, we propose to manage these RFM scores using the 2–tuple model which is a fuzzy linguistic model of information representation that carries out processes of "computing with words" without the loss of information. The proposed model permits us an easy linguistic interpretability and let us obtain a more precise representation of the RFM scores. Therefore, by interpreting these linguistic scores, 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. As a particular case study, we show an application example in order to select the customer of a loyalty campaign.

Peer-review under responsibility of the Organizing Committee of ITQM 2015

Keywords: RFM model; fuzzy linguistic model; 2-tuple linguistic model; IBM SPSS Modeler;

1. Introduction

RFM (Recency, Frequency and Monetary) is a model used to analyze the behavior of customer proposed by Hughes in 1994. "Recency" represents the length of a time period since the last purchase, while "Frequency" denotes the number of purchase within a specified time period and "Monetary" means the amount of money spent in this specified time period. In fact, these three variables belong to the behavioral variables and can be used to make predictions based on the conduct in the transactional database [1]. Therefore, in a RFM process, the goal is to obtain the customer purchase behavior (the most loyal customers, dormant customers…) from transactional data to proactively trigger appropriate marketing actions (retention, reactivation campaigns…) [2].

* Corresponding author. Tel.: +34 91 394 24 69.
E-mail address: ramoncar@ucm.es

Peer-review under responsibility of the Organizing Committee of ITQM 2015
doi:10.1016/j.procs.2015.07.118

The scores of these RFM variables are expressed by an ordinal number. 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 [3].

The fuzzy linguistic approach is a tool intended for modeling qualitative information in a problem. It is based on the concept of linguistic variable and has been satisfactorily used in many problems. The 2-tuple fuzzy linguistic approach [4, 5] is a model of information representation that carries out processes of "computing with words" without the loss of information.

In this paper, we propose to represent these RFM scales using the 2–tuples model. This yields a greater interpretability of these scores. Therefore, interpreting these linguistic scores, decision makers can effectively identify valuable customers and consequently develop more effective marketing strategy. Another further point is the fact that using the 2-tuple model we also obtain better accuracy of the results. Additionally, we present an IBM SPSS Modeler [6] implementation of this model. This makes us able to be widely applied at a practical level and not only at a theoretical one. As an example of use, we apply the new model to select the customer of a loyalty campaign.

The rest of the paper is organized as follows: Section 2 revises the preliminary concepts, i.e., the RFM model and the 2-tuple linguistic modeling. In Section 3 we propose to incorporate the 2-tuple model to RFM analysis and we show an implementation and use case of this new model using IBM SPSS Modeler. Finally, we point out some concluding remarks and future work.

2. Preliminaries

2.1. RFM Model

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 [2]:

- **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** (M): The total value of the purchases within the period examined.

Following, we explain the typically stages of a RFM analytic based on [2]:

1. **Data acquisition and preparation**. Once, we have chosen the period to analyze, the customers are selected if they have at least shopped during this period. Transactional data on these customers must be retrieved, audited, cleaned, and prepared for subsequent operations. Let *Transactions (CustomerID, Date, Amount)* be the table where these transactional data on purchases are included.
2. **RFM aggregation**. Transactional data is aggregated at a customer level, i. e., on the *CustomerID* attribute. Thus, we obtain the table *CustomerTransactions (CustomerID, Recency, Frequency, Monetary)* with the RFM information summarized for each customer.
3. **RFM scores computation**. Customers are sorted according to the respective RFM measure and are grouped in classes of equal size, typically quintiles. Customers are sorted independently according to each of the individual RFM components and then binned into five groups of 20%. This result is included into the table *CustomerRFM (CustomerID, RecencyScore, FrequencyScore, MonetaryScore, RFMScore)* with *RecencyScore, FrequencyScore, MonetaryScore* $\in$ {1,…,5}. Therefore the RFM measures are transformed into ordinal scores such that the value 1 includes the 20% of customers with the worst values and the 5 the 20% of customers with the best values in the corresponding measure. Especially for the *Recency* attribute, the scale of the derived ordinal score, *RecencyScore*, should be reversed so that larger scores represent the most recent buyers. Sometimes it can be useful to have a unique measure, *RFMScore*, which characterizes together the RFM scores. In order to provide this continuous RFM score, the R, F, and M bins are summed, with

appropriate user-defined weights, i.e., w_R, w_F, w_M. The RFM score is the weighted average of its individual components and is calculated as follows:

$$RFMScore = RecencyScore \times w_R + FrequencyScore \times w_F + MonetaryScore \times w_M \quad (1)$$

4. **RFM deployment**. Once the results of the previous step are validated, the marketers of the enterprise apply this RFM knowledge obtained in some business processes: retention, churn, segmentations… Often, these RFM results are used for a new data mining process.

This described approach can be solved with several data mining tools. In the Fig. 1 we show an example using IBM SPSS Modeler [6]. Following, we explain each stage of this stream:

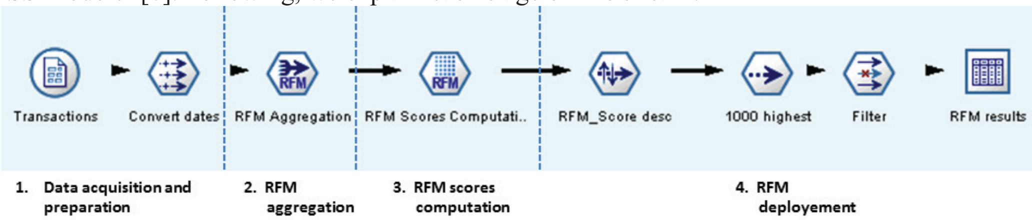

Fig. 1. Example of a RFM analytic with IBM SPSS Modeler

1. **Data acquisition and preparation**. Transactional data of the last year (2014) were retrieved, audited, cleaned, and prepared (casting to date type) for subsequent stages. Inactive customers with no purchases during 2014 were not included into the *Transactions* table.
2. **RFM aggregation**. For this step we use the *RFM Aggregate* node (see 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 (2015-01-01).

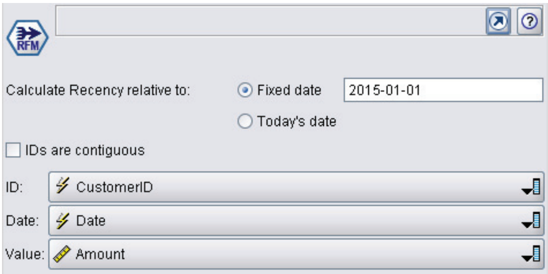

Fig. 2. Detail of the *RFM Aggregate* node settings

3. **RFM scores computation**. IBM SPSS Modeler also offers a node named *RFM Analysis* (Fig. 3) 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=.15, w_F=.1 and w_M= .75).

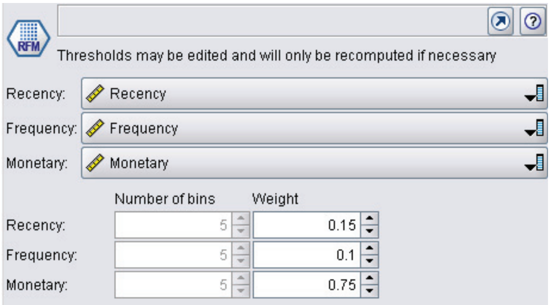

Fig. 3. Detail of the *RFM Analysis* node settings

4. **RFM deployment**. We apply the RFM results to select the customer of a loyalty campaign choosing the 1000 highest *RFMScore*. As can be seen in Fig. 4, there are many clients with equal score because when grouping customers in quintiles the procedure results in a total of 5 × 5 × 5 = 125 distinct values as much of *RFMScore*. This lack of precision can be a problem when selecting customers.

	CustomerID	Recency Score	Frequency Score	Monetary Score	RFM Score
1	C0100109460	5	5	5	5.000
2	C0100730638	5	5	5	5.000
3	C0101373994	5	5	5	5.000
4	C0105231950	5	5	5	5.000
5	C0105748652	5	5	5	5.000
6	C0101373743	5	5	5	5.000
7	C0105398565	5	5	5	5.000
8	C0103902748	5	5	5	5.000
9	C0101235072	5	5	5	5.000
10	C0102880841	5	5	5	5.000

Fig. 4. Detail of the *RFM results* table (extract of the *CustomerRFM* table)

2.2. The 2-Tuple Fuzzy Linguistic Approach

Let $S = \{s_0, \ldots, s_T\}$ be a linguistic term set with odd cardinality, where the mid-term represents a indifference value and the rest of terms are symmetric with respect to it. We assume that the semantics of labels is given by means of triangular membership functions and consider all terms distributed on a scale on which a total order is defined, i.e. $s_i \leq s_j \Leftrightarrow i < j$. In this fuzzy linguistic context, if a symbolic method aggregating linguistic information obtains a value $b \in [0,T]$, and $b \notin \{0,\ldots,T\}$, then an approximation function is used to express the result in S.

DEFINITION 1 [4]. *Let b be the result of an aggregation of the indexes of a set of labels assessed in a linguistic term set S, i.e. the result of a symbolic aggregation operation, $b \in [0,T]$. Let $i = round(b)$ and $\alpha = b-i$ be two values, such that $i \in [0,T]$ and $\alpha \in [-0.5,0.5)$, then α is called a Symbolic Translation.*

The 2-tuple fuzzy linguistic approach [2] is developed from the concept of symbolic translation by representing the linguistic information by means of 2-tuple (s_i, α_i), $s_i \in S$ and $\alpha_i \in [-0.5,0.5)$, where s_i represents the information linguistic label, and α_i is a numerical value expressing the value of the translation from the original result b to the closest index label, i, in the linguistic term set S. This model defines a set of transformation functions between numeric values and 2-tuple:

DEFINITION 2. *Let $S = \{s_1,\ldots,s_T\}$ be a linguistic term set and $b \in [0,T]$ a value representing the result of a symbolic aggregation operation, then the 2-tuple that expresses the equivalent information to b is obtained with the following function:*

$\Delta: [0,T] \rightarrow S \times [-0.5,0.5)$

$$\Delta(b) = (s_i, \alpha), \text{ with } s_i,\ i = round(b) \text{ and } \alpha = b - i,\ \alpha \in [-0.5,0.5) \tag{2}$$

where round(·) is the usual round operation, s_i has the closest index label to b and α is the value of the symbolic translation.

$$\text{For all } \Delta, \text{ there exists } \Delta^{-1}, \text{ defined as } \Delta^{-1}(s_i, \alpha) = i + \alpha. \tag{3}$$

$$\text{The negation operator is defined as } neg((s_i, \alpha)) = \Delta(T - (\Delta^{-1}(s_i, \alpha))). \tag{4}$$

Below, we describe the aggregation operators which we use in our model:

DEFINITION 3 [5]. *Let $A = \{(l_1, \alpha_1),\ldots, (l_n, \alpha_n)\}$ be a set of linguistic 2-tuple and $W = \{w_1,\ldots, w_n\}$ be their associated weights. The 2-tuple weighted average $\bar{A}^w$ is:*

$$\bar{A}^w[(l_1, \alpha_1),\ldots, (l_n, \alpha_n)] = \Delta\left(\frac{\sum_{i=1}^{n}\beta_i \cdot w_i}{\sum_{i=1}^{n} w_i}\right), \tag{5}$$

3. Using the 2-Tuple Fuzzy Linguistic Approach to RFM analysis

Although RFM analysis is an useful tool, it does have 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 Fig. 4 where you cannot identify which are really the best customer). In this section, we propose to incorporate the 2-tuple model to RFM analysis in order to obtain a higher precision and an easier linguistic interpretability of the model results. Additionally, we present an IBM SPSS Modeler [6] implementation of this model.

The basic idea is to compute and store the scores included into the output table, i.e., *RecencyScore, FrequencyScore, MonetaryScore* and *RFMScore* using the 2-tuple model. First, we need to define the symmetric and uniformly distributed domain *S* using five linguistic labels. These labels must have a semantic meaning for the four variables to model: $S = \{s_0,\ldots, s_T\}$, $T = 4$: s_0 = *Very Low* = *VL*, s_1 = *Low* = *L*, s_2 = *Medium* = *M*, s_3 = *High* = *H*, and s_4 = *Very High* = *VH* (see Fig. 5).

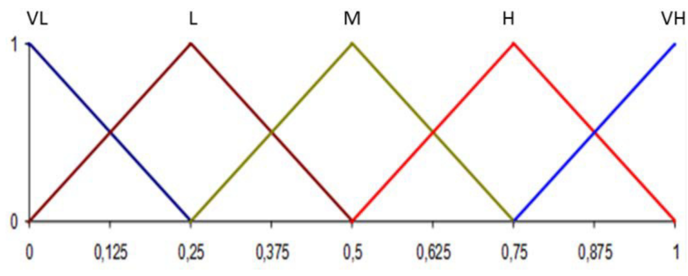

Fig. 5. Definition of the set *S*

The first two stages of the RFM analytic process are the same as we have shown in Section 2.1. Following, we explain the third and fourth stages:

3. **2-tuple RFM scores computation**. Customers are sorted according to the respective RFM measure and are grouped in classes of equal p size, i.e., p-quantiles. Customers are sorted independently according to each of the individual RFM components and then binned into p groups of $1/p$%. The higher the p value, the greater the accuracy of the results, although its maximum value depends on the distinct values of each RFM variable. Each B, with $B \in \{Recency, Frequency, Monetary\}$, measure is transformed into an ordinal number q, with $q \in \{1,..,p\}$. The final 2-tuple score A, with $A \in \{RecencyScore, FrequencyScore, MonetaryScore\}$, is obtained as following:

$$A = \begin{cases} \Delta(norm_{0T}(q)), \text{ if } B \in \{Frequency, Monetary\} \\ neg(\Delta(norm_{0T}(q))), \text{ otherwise} \end{cases} \tag{6}$$

where $norm_{0T}(\cdot)$ is the usual min-max (min=1 and max=p) normalization function to the interval $[0,T]$ and $\Delta(\cdot)$ and $neg(\cdot)$ have been defined in Section 2.2 (Eq. 2 and 4 respectively). We use the negation function on *Recency* because the larger scores represent the most recent buyers.

The 2-tuple *RFMScore* which characterizes together the RFM scores is calculated using the Eq. (5) as follows:

$$RFMScore = \bar{A}^{w}, \text{ with user-defined weights } W = \{w_R, w_F, w_M\} \tag{7}$$

4. **RFM deployment**. The scores obtained in the previous step are 2-tuple values, therefore before applying any conventional numerical operation we have to transform these values using the corresponding function Δ^{-1} (Eq. 3).

In [7] we have proposed both a representation data type 2-tuple as the implementation of the functions Δ and Δ^{-1} using IBM SPSS Modeler. Using these tools, the RFM 2-tuple approach proposed in this paper has been implemented. Thus, the stream to solve the example of the Section 1.1 is showed in the Fig. 6.

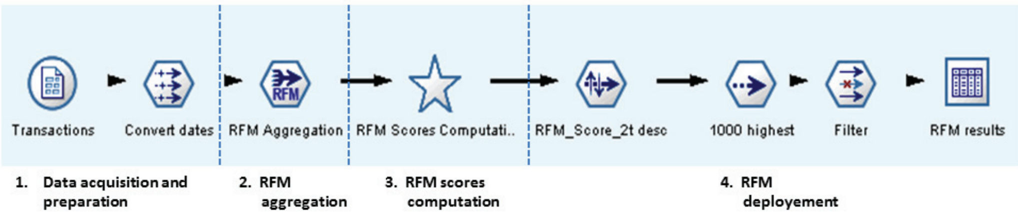

Fig. 6. Example of a 2-tuple RFM analytic with IBM SPSS Modeler

This stream is very similar to the presented in the Fig. 1 to solve a conventional RFM process. The main difference is the implementation of the third stage. For this purpose we have created a super node (symbolized by ) named *RFM Scores Computation* (shown in Fig. 7). A super node is similar to a procedure with inputs (labeled with *From Stream*) and/or outputs values (labeled with *To Stream*). This super node computes de 2-tuple RFM scores (using the Eq. 6 and 7) and it also computes the corresponding Δ^{-1} function (Eq. 3) on these values to apply conventional numerical operations necessary in the fourth stage.

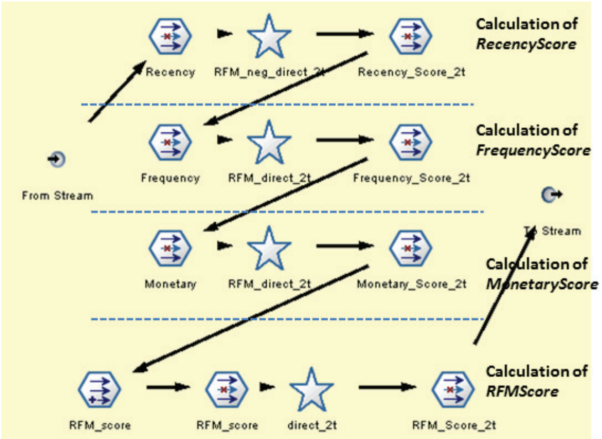

Fig. 7. IBM SPSS super node for RFM scores computation

We execute the new version of the stream (Fig. 6) with p = 100 on the same input data and the same user-defined weights (w_R, w_F, w_M) used on the conventional stream (Fig. 1). The selection of the 1000 best customers, i.e., the highest Δ^{-1}*(RFMScore)* presented in the Table 1 is different now. Also in the Table 1, we show the RFM scores that were obtained according to the conventional process (Fig. 1). 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.

Table 1. Comparison of the results of the conventional RFM process vs the 2-tuple RFM process

				Conventional RFM process				2-tuples RFM process			
CustomerID	Frequency	Monetary	Recency	Recency Score	Frequency Score	Monetary Score	RFM Score	Recency Score	Frequency Score	Monetary Score	RFM Score
C0103319423	28	1212.28	5	5	5	5	5	VH-0.040404	VH-0.034483	VH-0.030303	VH-0.032236
C0102509617	26	1423.8	11	5	5	5	5	VH-0.101010	VH-0.103448	VH-0.010101	VH-0.033072
C0102071070	27	2079.81	21	4	5	5	4.85	H+0.047980	VH-0.068966	VH	VH-0.037200
C0102323890	26	1311.28	10	5	5	5	5	VH-0.090909	VH-0.103448	VH-0.020202	VH-0.039133
C0104283094	28	1113.79	7	5	5	5	5	VH-0.060606	VH-0.034483	VH-0.040404	VH-0.042842
C0105242662	30	1045.74	1	5	5	5	5	VH-0.010101	VH	VH-0.060606	VH-0.046970
...											

4. Concluding remarks and future work

We have presented a general 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. 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 types of marketing problems: segmentation, retention, reactivation campaigns, etc. As an example, we have applied the implemented model to select the customer of a loyalty campaign, verifying the advantages of the new model regarding the conventional RFM model.

We are currently focusing on the use of this model to several marketing problems, especially for the calculation of the customer lifetime value.

Acknowledgements

This work has been developed with the financing of FEDER funds in National Project TIN2013-40658-P and the Andalusian Excellence Project TIC-5991.

References

[1] Jo-Ting W, Shih-Yen L, Hsin-Hung W. A review of the application of RFM model. African Journal of Business Management 2010;**4**(19):4199–4206.
[2] Tsiptsis K, Chorianopoulos A. *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons; 2011.
[3] Cheng CH, Chen YS. Classifying the segmentation of customer value via RFM model and RS theory. Expert systems with applications 2009;**36**(3):4176–4184.
[4] Herrera, F., Alonso, S., Chiclana, F., Herrera-Viedma, E. Computing With Words in Decision Making: Foundations, Trends and Prospects. Fuzzy Optimization and Decision Making 2009; **8**(4):337-364.
[5] Herrera-Viedma E, Herrera F, Martınez L, Herrera JC, López AG. Incorporating filtering techniques in a fuzzy linguistic multi-agent model for information gathering on the web. Fuzzy sets and Systems 2004;**148**(1):61–83.
[6] IBM SPSS Modeler. URL (consulted 2015): http://www.ibm.com/software/analytics/spss/products/modeler/
[7] Carrasco RA, Blasco MF, Herrera-Viedma E. An Implementation of a Linguistic Multi-Criteria Decision Making Model: An Application to Tourism. LNAI 2014 **8536**:232–239.