The RFM-based Institutional Customers Clustering: Case Study of a Digital Content Provider

Spring C. Hsu
Department of Business Administration,
China University of Science and Technology, Taiwan

Abstract: The RFM (recency, frequency and monetary) model has been widely applied for personal customers’ analysis, but limited for the institutional customers. Therefore, the study takes a digital content provider with institutional customers of a variety of Small and Medium Enterprises (SMEs) in Taipei city for RFM clustering and deploys an innovative method combining RFM model and fuzzy c-means (FCM) for analysis. Clustering results suggest that September and October are the two busiest months of transactions for major customers. Moreover, major customers have twice of transactions and total transaction amount of 6,462 USD; i.e., an average transaction amount of 3,231 USD. Consequently, for a digital content provider, the enough staff and equipment preparations must ready before September and October to make sure a high quality service for major customers. The average transaction amount implies that the approximate cost of 3,250 USD is the upper bound of willing to pay for many SMEs. Therefore, to deliver the customized service packages with the charges close to 3,250 USD are the first priority of services development for a digital content provider.

Key words: RFM, digital content providers, fuzzy clustering

INTRODUCTION

As the business environment changes from the product-oriented to the customer-oriented to understand the customer behavior is thus becoming more important. The RFM model has been widely applied in many practical areas and showed the ability to profile the customer behaviors (Bhensdadia and Kosta, 2011; Bizhani and Tarokh, 2011; Blattberg et al., 2008; Buttle, 2009; Kohavi and Parekh, 2004; Sekhavat et al., 2010). The RFM method is the popular method used in market segmentation that shows profitable groups of customers. Companies easily find valuable customer behaviors and then develop effectively corresponding marketing strategies by adopting the RFM model.

Literature characterized the general RFM definitions as follows (Hughes, 1994; Stone, 1995; Wei et al., 2010). \( R \) is the time of last transaction during an analyzing period and the latest transactions account for the bigger \( R \) values. \( F \) is the frequency of transactions during an analyzing period and the higher the frequency is, the bigger \( F \) will. \( M \) is the total amount of spent money during an analyzing period and the much the monetary is, the bigger \( M \) will. Chao et al. (2008) indicated that the \( F \) and \( M \) values are more important for profiling customers. Moreover, Eq. 1 defines the RFM Score, where \( \alpha, \beta \) and \( \gamma \) denote the weights of \( R \), \( F \) and \( M \), respectively. For these weights, Hughes (1994) simply regarded the RFM Score with equal weights, but Stone (1995) regarded that the weights are not equal according to different industries:

\[
\text{RFM Score} = \alpha \cdot R + \beta \cdot F + \gamma \cdot M
\]  

(1)

Some of the variant studies relevant to the RFM model accompanied with neural network methods. However, because of the case-adapted assumptions, these studies showed their limit for general applications (Geyik, 2007; Lee and Hong, 2008). Liu and Shih (2005) tried to rate Customer Lifetime Value (CLV) by the RFM model. Their study clustered customers according to their lifetime value expressed with weighted RFM. However, Li et al. (2006) developed a timely RFM (TRFM) model that conducts the relationship of product properties and purchase periodicity. However, their model is also limited to specific case-adapted assumptions.

Moreover, the clustering methods play important roles to carry out the RFM model. Except for the conventional hierarchical method, the widely applied clustering methods are the \( k \)-means and fuzzy \( c \)-means (FCM) (Aggelis and Christodoulakis, 2005; Alvarez et al., 2007; Cheng and Chen, 2009; Hanzehei et al., 2011;
Fuzzy c-means is a method of clustering that allows an observation to belong with two or more clusters. This method is often used in pattern recognition. Similar to k-means method, a prior confirmation of cluster number is needed to process the fuzzy c-means method. However, not as the k-means method, the membership of input data points is not clearly belonged to specific clusters in fuzzy c-means method. This study initializes the cluster number via subtractive clustering before conducting fuzzy c-means method. Because subtractive clustering is a fast, one-pass algorithm for estimating the cluster number in a set of data (Abu-Qhosh et al., 2010; Chiu, 1996), and the cluster numbers from the subtractive clustering are usually used to initialize other fuzzy clustering or forecasting methods (Castellanos and James, 2009; Singh and Nagraja, 2011). Moreover, since the subtractive clustering is a commonly used method, this study makes the subtractive clustering directly via the GUI (Graphical User Interface) of mathematical software MATLAB 7.4.

Fuzzy c-means algorithms: Equation 2 represents the objective function J of fuzzy c-means method, where n is the size of data points, k is the cluster number, x_j is jth input vector, c_i is ith cluster center and u_i is point j’s membership of cluster i. In addition, m is fuzziness exponent and usually equal to two (Abraham et al., 2006; Miyamoto et al., 2008):

\[ J = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{k} u_{ij}^m \| x_j - c_i \| \]  

(2)

It is based on minimization of the objective function to do the fuzzy c-means method. Moreover, the procedure of fuzzy c-means algorithm is specified as follows (Celikyilmaz and Turksen, 2009; Mirkin, 2011; Pedrycz, 2007):

- Identify cluster number k and initialize the membership function values that satisfy the Eq. 3:

\[ \sum_{i=1}^{n} u_{ik} = 1 \quad \text{for } 1 \leq j \leq n \]  

- Calculate the cluster centers by Eq. 4:

\[ c_i = \frac{\sum_{k=1}^{n} u_{ik}^{-\beta} x_k}{\sum_{k=1}^{n} u_{ik}^{-\beta}} \quad \text{for } 1 \leq i \leq k \]  

- Calculate the objective function with Eq. 2. This objective function varies from process to process of clustering. The iteration is stopped when goal ε is achieved; otherwise, go to procedure (4)

- Update membership function as Eq. 5 to cluster data under the constraint of Eq. 3. \(|x_j - c_i|\) stands for distance from point j to current cluster center i. \(|x_j - c_s|\) stands for distance from point j to other cluster centers s. After updating the membership function, go back to procedure (2):

\[ u_{ik} = \frac{1}{\sum_{k=1}^{n} \left( \frac{|x_j - c_i|}{|x_j - c_s|} \right)^{\frac{1}{\beta}}} \quad \text{for } 1 \leq i \leq k, 1 \leq j \leq n \]  

Moreover, this study makes the fuzzy c-means clustering via the program code developed by this study and run it in the mathematical software MATLAB 7.4.

THE CASE STUDY

The case-company, DIGITO Technology Co. Ltd., is a digital content provider in Taipei city. With experienced marketing integrators and e-commerce technicians, the company provides application software and a variety of integrated services of inter-net digital content for many SMEs and some large enterprises. Aim to offer customers a one-step total solution for digital content, major service patterns of the case-company are listed in Table 1.

Customers of the case-company come from a variety of industries, such as entertainment, media, food, furniture and beverages. As the result of complexity of customer composition, finding the core and clear behaviors of customers is thus an important issue to give more suitable services to customers. Therefore, according to the complex operation conditions of the case-company, the fuzzy clustering method is deployed to analyze the customer behaviors of the case-company.

Pre-processing of data: The analyzed raw data concern the time during January 1, 2008 and March 31, 2011, with the total of 273 data points of customers. For the pre-processing of the data, the RFM variables are first redefined or calculated as follows.

- R is the last transaction date of customers. Since the R value contributes to the RFM scoring determination, a numeric value is necessary. Therefore, R is redefined as the sequential number after the first date of concerned analysis time 2008/1/1 to 2011/3/31. For example, a customer that has conducted last transaction on 2008/1/1 is characterized by \( R = 1 \) and the other customer that has conducted last transaction on 2009/1/31 is characterized by \( R = 366+31 = 397 \).

- F is the count of transactions the customer conducted within the analysis period. M is the total amount of transaction the customer made within the analysis period. Moreover, RFM Score is summation of R, F and M values. Fragmented samples of the pre-processed R, F and M values are listed in Table 2.

Customers clustering: The two stages clustering process with subtractive clustering and fuzzy c-means are deployed to learn the cluster numbers and then to carry out the clustering. The subtractive clustering function of mathematical software MATLAB 7.4 is used learning the proper cluster numbers. In the subtractive clustering, the influence range is set as 0.5, squash is set as 1.25 and accept and reject ratios are 0.5 and 0.15, respectively. Moreover, to make sure all the important variables or variable sets are not ignored, two phases of subtractive clustering are made, shows as Table 3.
In phase-I, individual RFM variables and simple RFM score with equal weights are considered.

The variable R is subtracted into three clusters and the others are subtracted into a cluster. The clustering results of phase-I infer that variable R plays an important role in the case. Accordingly, a new weighted RFM score with weights of \( \alpha = 3/(3+1+1) = 3/3, \beta = 1/(3+1+1) = 1/5, \gamma = 1/(3+1+1) = 1/5 \) is deployed in phase-II clustering. Meanwhile, to make sure that the inherent different scales of variables R, F and M do not interfere with calculation of RFM scores, a normalized RFM score is also developed in this study. To calculate this normalized RFM score, variables R, F and M are first normalized by Eq. (6), where \( x_i^j \) denotes rescaled/normalized values of data point \( x_i \). Furthermore, a RFM score with normalized data points and different weights is also proposed as “weighted normalized RFM score.” Finally, the clustering with R, F and M three variables together is also made in phase-II. It is concluded from Table 3 that accounts for the variables R, F and M together makes a subtractive clustering suggestion with three clusters. Therefore, it is confirmed that this is the key result that deserved an in-depth analysis:

\[
x_i^j = \frac{x_i}{\sqrt{\sum x_i^j}}
\]  

Consequently, the three variables R, F and M are selected together to make clusters and the scatter plot of original data are shown as Fig. 1. To do the fuzzy clustering, this study writes the fuzzy c-means clustering program codes that executed under the mathematical software MATLAB 7.4. In the fuzzy c-means clustering, exponent for the partition matrix \( m \) is set as 2 and least amount of improvement is set as 1e-5. The fuzzy c-means program then runs 98 iterations and converges with the objective function value of \( 2.88 \times 10^{-4} \), Fig. 2. Meanwhile, Table 4 shows fragment of memberships of fuzzy partitioning and the fuzzy clustering results are shown in Table 5. The visualized clustering results are shown in Fig. 3.

However, it is easily to spot an apparent outlier point on the right back corner in Fig. 3. With the corresponding R, F and M values of this outlier point, it is found in the original data set as customer number No. 228 (with the values of R = 1095, F = 22 and M = 31,147,000, respectively). Therefore, it is necessary to remove the outlier from original data set and executes the fuzzy clustering process with the trimmed data set. With the trimmed data set, the program then runs 78 iterations and converges with the objective function value of \( 1.09 \times 10^{-4} \) (Fig. 4).
Fig. 2: Iterations of objective function for original data

Fig. 3: Clustering and centering plot of original RFM data

Table 5: Clustering outcomes with original RFM data

<table>
<thead>
<tr>
<th>Clusters</th>
<th>R</th>
<th>F</th>
<th>M</th>
<th>%b</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁</td>
<td>6.1538529965891235e+002</td>
<td>1.8984570072188056e+000</td>
<td>2.2445507954496346e+005</td>
<td>89.7435897435897</td>
</tr>
<tr>
<td>C₂</td>
<td>7.4514710631667779e+002</td>
<td>9.1091357769247477e+000</td>
<td>5.3561130092610887e+006</td>
<td>7.32600732600733</td>
</tr>
<tr>
<td>C₃</td>
<td>9.0042169787439598e+002</td>
<td>1.3505960229940110e+001</td>
<td>1.425509590473837e+007</td>
<td>2.93040293040293</td>
</tr>
</tbody>
</table>

Note: percentages stand for the data points partitioned to clusters

Comparing Fig. 2 with 4, removing the outlier data point facilitates efficiency of the convergence process. Finally, the fuzzy clustering results with trimmed data set are shown in Table 6. The visualized clustering results with trimmed data set are shown in Fig. 5.
Fig. 4: Iterations of objective function for trimmed data

Fig. 5: Clustering and centering plot of trimmed RFM data

Table 6: Clustering outcomes with trimmed RFM data

<table>
<thead>
<tr>
<th>Clusters</th>
<th>R</th>
<th>F</th>
<th>M</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_1)</td>
<td>6.1347717788700198e+002</td>
<td>1.81330821089285800e+000</td>
<td>1.901139345569749e+005</td>
<td>99.0735294117647</td>
</tr>
<tr>
<td>(C_2)</td>
<td>6.935973439562101e+002</td>
<td>8.154399201912653e+000</td>
<td>4.1493802171554156e+006</td>
<td>6.61764705882353</td>
</tr>
<tr>
<td>(C_3)</td>
<td>8.634653285345729e+002</td>
<td>1.1731004779725149e+001</td>
<td>1.2258541022436669e+007</td>
<td>3.30882352041176</td>
</tr>
</tbody>
</table>

Note: percentages stand for the data points partitioned to clusters.

It is now concluded that for the case-company data, the clustering with three variables R, F and M is appropriately portioned into three clusters. Where cluster \(C_1\) approximately accounts for 90% of all data (approximate R = 613 - 2009/4, F = 2, M = 190,114; i.e., \(M/F = 95.057\)), cluster \(C_2\) approximately accounts for 7% of all data (approximate R = 694 = 2009/10/24, F = 8, M = 4,149,380; i.e., \(M/F = 518,673\)) and cluster \(C_3\) approximately...
accounts for simply 3% of all data (approximate \( R = 863 = 2010/5/12, F = 12, M = 12,258,541 \); i.e., \( M/F = 1,021,545 \)).

In brief, the cluster C, represents the “major customers” for the case-company. It has the average transactions of two times and average transaction amount of 95 thousand NTD (i.e., 3,231 USD; exchange rate NTD/USD=29.418, March 31, 2011, Central Bank of Taiwan). Moreover, the frequencies show that September and October are the two busiest transaction months of major customers. The cluster C, represents the “first-minor customers” for the case-company. It has the average transactions of eight times and average transaction amount of 519 thousand NTD (i.e., 17,631 USD). Moreover, the cluster C, represents the “second-minor customers” for the case-company. It has the average transactions of 12 times and average transaction amount of an million NTD (i.e., 34,725 USD). Observe that, the major customers (i.e., C,) account for 90% of all customers, but making lower contributions to revenues for the case-company. Comparatively, the minor customers (i.e., C, and C,%) account for 10% of all customers, but making higher contributions to revenues for the case-company. Finally, remind that since the case-company serves many SMEs and few large enterprise customers, the major customers of clustering results consist of SMEs and the minor consist of large enterprises. The results suggest that if the case-company wants to keep or even increase market shares, it should focus on the SMEs, otherwise, if the case-company wants to increase revenues, the large enterprises should be targeted.

CONCLUSIONS

The behaviors of institutional customers are not commonly mentioned in literature. This study analyzes the case-company that mainly provides digital content services to institutional customers. Meanwhile, this study deployed an innovative method combining RFM model and FCM clustering for analyzing the case-company. The analyzed period ranges from January 1, 2008 through March 31, 2011, with originally 273 raw customer data and finally to analyze 272 data after trimming an outlier point. This study deployed the two stages clustering process with subtractive clustering and fuzzy c-means to learn the cluster numbers and then to carry out the clustering via variables of RFM model. This study applied the creative way to decide the weights of RFM scores via importance ranking of R, F and M subtractive clustering outcomes. Without any assumptions or constraints, the clustering approach of this study effectively and easily makes simultaneously clustering on R, F and M variables.

Three customer categories named “major customers,” “first-minor customers,” and “second-minor customers” are clustered in this study. The major customers account for 90% of all customers, but making lower contributions to revenues for the case-company. Comparatively, the minor customers account for 10% of all customers, but making higher contributions to revenues for the case-company. The major customers of clustering results consist of SMEs and the minor consist of large enterprises. The results suggest that if the case-company wants to keep or even increase market shares, it should focus on the SMEs; otherwise, if the case-company wants to increase revenues, it should target the large enterprises.

Suppose that the major customers are critical to the case-company, it is found that September and October are the two busiest transaction months of major customers. Moreover, the major customers in the case-company have the approximate average transaction amount of 3,231 USD. This study concluded that for a digital content provider, the enough staff and equipment preparations must ready before September and October these two busy months to make sure the high quality service for major customers. Finally, since the average transaction amount is below 3,250 USD, it infers that the cost of 3,250 USD is the upper bound of willing to pay for many SMEs. Therefore, to deliver the customized service packages with the charges close to 3,250 USD are the first priority of services development.

REFERENCES


