Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty

Seyed Mohammad Seyed Hosseini *, Anahita Maleki, Mohammad Reza Gholamian

Industrial Engineering Department, Iran University of Science and Technology, Tehran, Iran

ARTICLE INFO

ABSTRACT

Data mining (DM) methodology has a tremendous contribution for researchers to extract the hidden knowledge and information which have been inherited in the data used by researchers. This study has proposed a new procedure, based on expanded RFM model by including one additional parameter, joining WRFM-based method to *K*-means algorithm applied in DM with *K*-optimum according to Davies–Bouldin Index, and then classifying customer product loyalty in under B2B concept. The developed methodology has been implemented for SAPCO Co. in Iran. The result shows a tremendous capability to the firm to assess his customer loyalty in marketing strategy designed by this company in comparing with random selection commonly used by most companies in Iran.

1. Introduction

In a B2B environment, suppliers and/or service providers usually need to understand the nature and characteristics of their customers. As customer attraction and satisfaction are the main objectives of any leading company, so the main objective of this article is to provide an effective and efficient methodology to be used for implementing the firm's objective to the best of possible. This part mainly reviews the studies related to customer relationship management, customer loyalty, RFM model, *K*-means algorithm.

1.1. Customer relationship management

Since the early 1980s, the concept of customer relationship management in marketing, and consists of four dimensions: customer identification, customer attraction, customer retention and customer development has gained its importance. It is difficult to find out a totally approved definition of CRM. We can describe it as a comprehensive strategy and process of acquiring, retaining and partnering with selective customers to create superior value for the company and the customer (Parvatiyar & Sheth, 2004). CRM is a comprehensive business and marketing strategy that integrates technology, process, and all business activities around the customer (Anton, 1996; Anton & Hoeck, 2002). Brown points out that CRM as "the key competitive strategy you need to stay focused on the needs of your customers and to integrate a customer-facing approach throughout your organization" (Brown, 2000). Chatterjee also defines CRM as a discipline which focuses on automating and improving the business processes associated with managing customer relationships in the area of sales, management, customer service, and support (Chatterjee, 2000). According to Feinberg and Kadam, profits increase by 25–80% when customer retention rates increase by five points (Feinberg & Kadam, 2002). CRM projects often fail and only about 40% of CRM implementations are successful (Feinberg & Trotter, 2001).

1.2. Customer loyalty

Creating a loyal B2B customer base is not only about maintaining numbers of customer overtime, but it is creating the relationship with business customers to encourage their future purchase and level of advocacy. Equipped with the knowledge of their business customers' loyalty levels, a supplier will be able to figure how their endeavors to maintain good relationships can contribute to its profit levels. Some authors believe that loyal customers offer a steady stream of revenue for a company by remaining with the brand/supplier and rejecting the overtures of competitors (Lam, Shankar, Erramilli, & Murthy, 2004; Reichheld & Teal, 1996). Considering this with the nature of large purchase and transactions in a B2B setting; there are gigantic rewards for those suppliers who succeed in creating and maintaining loyal customers.

Some Authors have proposed several theories to link variables that one usually finds in relationship marketing and business marketing to the loyalty construct. In the B2B context, evidence shows that relationship elements affect customer loyalty. For example, Ricard and Perrien found that relationship practices have a direct impact on customer loyalty (Ricard & Perrien, 1999). Other Authors

---

* Corresponding author. Tel.: +982177240447; fax: +982177240482.
E-mail addresses: seyedhoseini@iust.ac.ir (S.M.S. Hosseini), anahitamaleki@yahoo.com (A. Maleki), gholamian@iust.ac.ir (M.R. Gholamian).
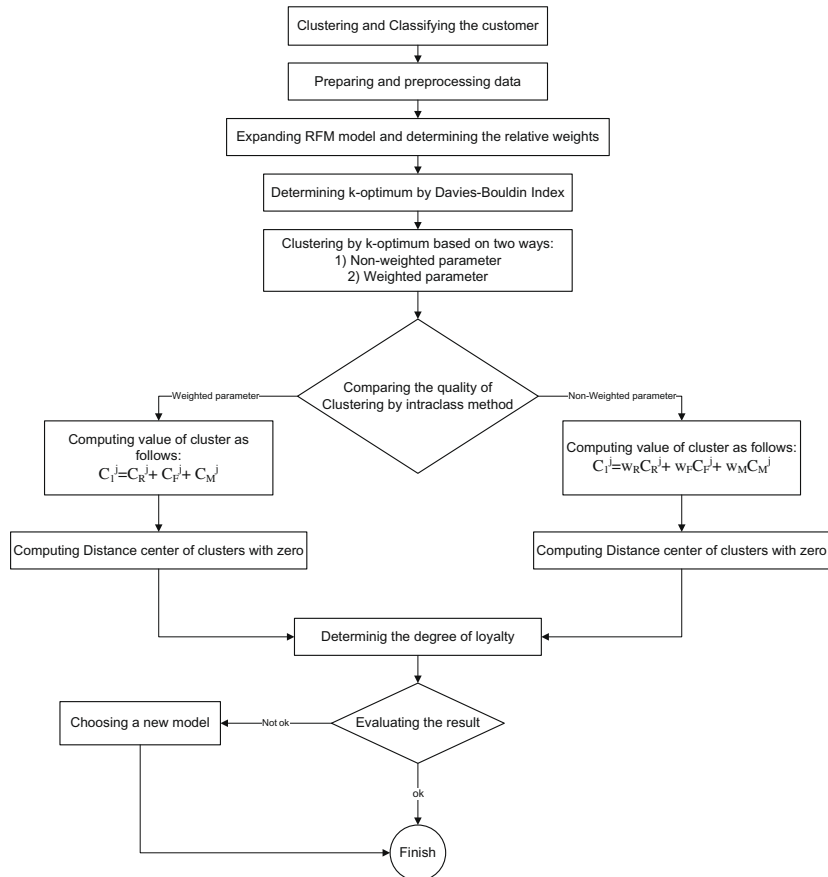
Fig. 1. Proposed method.

provide empirical evidence linking several constructs such as relationship quality, trust, involvement, satisfaction, purchase development, organizational change, and switching costs to influence B2B customer loyalty and retention (Chow & Holden, 1997; Eriksson & Vaghult, 2000).

In some researches, customer lifetime value (CLV) or loyalty is evaluated in terms of recency, frequency, monetary variables namely the integrating rate of each cluster, that is $C_1^j = w_R C_R^j + w_F C_F^j + w_M C_M^j$, where w_R, w_F, w_M are the relative importance of the RFM variables (Liu & Shih, 2005).

Other researches for classifying customer value have proposed a model based on computing the distance between the center of cluster and zero point as high value refers to most customer loyalty (Cheng & Chen, 2008).

1.3. RFM analysis

RFM analysis[1] has been used in direct marketing for several decades (Baier, Ruf, & Chakraborty, 2002). This technique identifies customer behavior and represents customer behavior characteristics by three variables as follows:

(1) Recency of the last purchase which refers to the interval between the time that the latest consuming behavior happens and present.
(2) Frequency of the purchases which refers to the number of transactions in a particular period.

(3) Monetary value of the purchase which refers to consumption money amount in a particular period.

RFM analysis is utilized in many ways by practitioners; therefore, RFM analysis can mean different things to different people. Classic RFM implementation ranks each customer on valuable parameters against all the other customers, and creates an RFM score for each customer/product.

The first step is to sort the customer file according to how recently customers have purchased from the firm. Then database divided into equal quintiles and these quintiles are assigned the numbers 5 to 1. Therefore, the 20% of the customers who most recently purchased from the company are assigned the number 5; the next 20% are assigned the number 4, and so on. The next step involves sorting the frequently, monetary. Therefore, the database is divided into 125 roughly equal groups (cells) according to recency, frequency, and monetary value. Customers/products with high scores are usually the most profitable (Hughes, 1994; Stone, 1995).

1.4. K-Means algorithm

Clustering is the process of grouping a set of physical objects into similar groups. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters (Han & Kamber, 2001). K-Means is one of the well-known algorithms for clustering which is very sensitive to the choice of a starting point for partitioning the items into K initial clusters. We can compare the performance of different clustering methods using intraclass method when the number of fixed cluster of K value is defined as $F(K) = 1 / K \sum_{n=1}^{k} \sum_{c_i} Dist(c_i, c^n)$. (Michaud, 1997; Shina & Sohnb, 2004).

---

[1] Recency, Frequency and Monetary.

2. Customer loyalty assessment model

We briefly introduce the proposed procedure for classifying customer value based on products.

The proposed model is based on the use of some data mining techniques and customer lifetime value analysis to improve CRM for enterprises, based on RFM attributes and K-means method for clustering product value. General purpose of this study is to determine loyalty degree of product to achieve an excellent CRM which are believed to maximize profits with win-to-win situation under B2B concept. Fig. 1 illustrates the proposed method in this study.

The proposed procedure is divided into six steps: (1) selecting the dataset and preparing data; (2) defining recency, frequency and monetary and period of product activity as the parameters of model then determining weight of each parameter by eigenvector technique and yielding the scale of RFML attributes for the purpose of clustering, then partitioning all parts of firm into 34 classes

while we evaluate the optimum K based on Davies–Bouldin Index in order to use in K-means algorithm; next the parts have been clustered in two ways which are based on weighted-parameters and non-weighted-parameters in order to compare the quality of clusters based on $F(K)$; (3) Computing the integrated rate of each cluster namely F that is $C_1^j = w_R C_R^j + w_F C_F^j + w_M C_M^j$ where w_R, w_F, w_M and w_L are the relative importance of the RFML variables; (4) computing distance between center of each cluster with zero point namely D; (5) achieving appropriate equation for assessing CLV ranking under two parameter namely D and F; (6) finally, splitting dataset into training data and testing data and validating data in order to evaluate the results.

The computed process is introduced step by step as follows:

Step 1: Preparing and preprocessing data.

At first, select the dataset for empirical study. To preprocess the dataset to make knowledge discovery easier is needed. Thus, we firstly delete the records which include missing values or inaccurate values, eliminate the redundant attributes and transform the datum into a format that will be more easily and effectively processed for clustering customer value.

Step 2: Clustering customer value by K-means algorithm.

The following step is to define the variables of model and the scale of R–F–M–L attributes based on Hughes then yield quantitative value of RFML attributes as input attributes for clustering customer value by using K-means algorithm with K optimum according to Davies–Bouldin Index and compare the performances of these clustering methods using 'intraclass' method.

The detail process of this step is expressed into these sub-steps:

(1) Defining the R–F–M–L variables for empirical study.
(2) The R–F–M–L attributes have different weight which is obtained by filling the questionnaire and using eigenvector technique for group decision makers in firm which is computed as follow: $w_R = 0.25$, $w_F = 0.15$, $w_M = 0.5$ and $w_L = 0.1$.
(3) Defining the scale of four R–F–M–L attributes, which are 5, 4, 3, 2 and 1 and referring to the customer contributions to revenue for enterprises for each product. For example '5' refers to the most customer contribution to revenue and '1' refers

Table 1
The matrix of group decision making.

$$\begin{bmatrix} 1 & 3.9 & 3.8 & 3.4 \\ 0.25 & 1 & 4.2 & 2.4 \\ 0.26 & 0.23 & 1 & 3.8 \\ 0.29 & 0.41 & 0.26 & 1 \end{bmatrix}$$

Table 2
The partial data of SAPCO Co.

ID	R	F	M	L
⋮	⋮	⋮	⋮	⋮
72316870	4.63	2	108,864,000	2
96468606XX	99.82	8	409,864,886	9
K9118422980	30.01	9	1,784,428,920	7
9420229517	100.00	14	127,618,560	2
YG202992L2	0.00	1	120,750,000	6
K914561018A	57.14	2	818,210,016	5
⋮	⋮	⋮	⋮	⋮

Table 3
The real scaling of R–F–M–L attributes.

Scaling	R	F	M	L
5	89–100	13–14–15–16	270,316,800–43,865,483,200	9
4	68–88	10–11–12	49,264,128–268,853,200	8
3	43–67	7–8–9	10,432,800–48,875,000	6–7
2	16–42	4–5–6	1,571,760–10,403,228	3–4–5
1	0–15	1–2–3	2000–1,569,600	1–2

Table 4
Determine optimum K.

Method	$F(34)$
Clustering with non-weighted-parameter	0.186094
Clustering with weighted-parameter	0.181831

Table 6
First evaluation.

	R^2/F	R^2/D	$R^2/F+D$
F	0.95	0.85	0.91
D	0.85	0.95	0.93
$F+D$	0.92	0.94	0.96
Mean	0.90	0.91	0.93

Table 7
Second evaluation.

Method	Artificial Neural Network	Decision Tree
Proposed model	99.10	99.72
Base model	95.01	96.63

Table 5
The cluster result by K-means with 34 classes on output based on D, F, $F + D$.

Cluster no.	Member	R	F	M	L	D	D ranking	F	F ranking	F + D	F+D Ranking
1	38	4.526	3.079	4.895	4.842	8.7961	2	4.595	2	13.321	2
2	221	1.036	1.000	1.317	1.145	2.2625	34	1.182	34	3.444	34
⋮	–	–	–	–	–	–	–	–	–	–	–
33	10	3.100	1.000	1.200	4.800	5.9237	23	2.005	29	7.9286	26
34	55	3.164	1.036	1.073	2.236	4.1515	30	1.7065	31	5.8580	31

to the least contribution to revenue for m parameter. So we totally have 625 ($5 \times 5 \times 5 \times 5$) combinations in the proposed model.

(4) We can evaluate the optimum K by Davies–Bouldin Index, in order to use in K-means algorithm for the purpose of clustering.

(5) The parts have been clustered in two ways which are based on weighted-parameters and non-weighted-parameters.

(6) We can compare the quality of clusters based on $F(K)$.

Step 3: The CLV ranking was derived to help develop more effective strategies for retaining customers and thus identify and compare market segments. Clustering the integrated rate of each cluster refers to $C_1^j = w_R C_R^j + w_F C_F^j + w_M C_M^j + w_L CL^j$ namely F, where w_R, w_F, w_M and w_L are the relative importance of the RFML variables. Finally, sort the F_1 to F_{34} by descendant order which can be indicated the degree of loyalty.

Step 4: Base on Davies–Bouldin Index, let $K = 34$ for the purpose of clustering. Furthermore, so we can obtain the center of clusters as follows:

$C_1 = \{S_{11}, S_{12}, \ldots, S_{1P1}\}, \ldots, C_{34} = \{S_{341}, S_{342}, \ldots, S_{34PK}\}$ where S_{ij} denotes the jth element of C_i, P_i denotes the number of elements in C_i, so we obtain $c_1, c_2, \ldots, c_{34}$ as the center of the clusters as follow:

$$C_1 = (v_{11}, v_{12}, v_{13}, v_{14}), \ldots, c_{34} = (v_{341}, v_{342}, v_{343}, v_{344}). \tag{1}$$

Then compute the distance between c_i and zero point namely D.

Finally, sort the distance D_1 to D_{34} by descendant order which can be indicated the degree of loyalty.

Step 5: When we combine ranking of D and ranking of F for indicating the degree of loyalty, we can obtain an appropriate equation for customer loyalty assessment based on the degree of loyalty and number of clusters.

Step 6: Evaluating the result. First evaluation is R^2 value test in order to R^2 value which is shown on equation chart based on ranking of F with F, D, $F + D$ values, ranking of D with F, D, $F + D$ values and ranking of $F + D$ with F, D, $F + D$ values.

Second evaluation is the accuracy rate of the generated rules with Decision Tree and Artificial Neural Network methods;
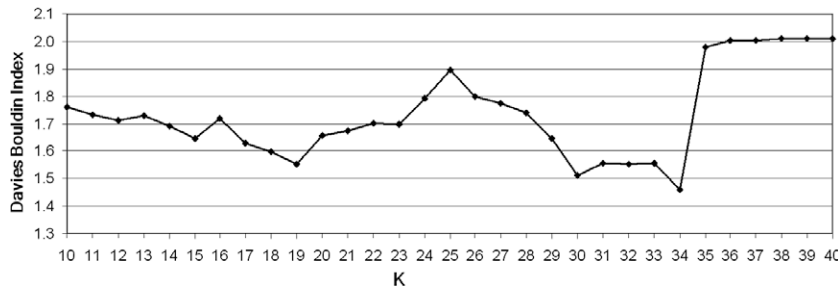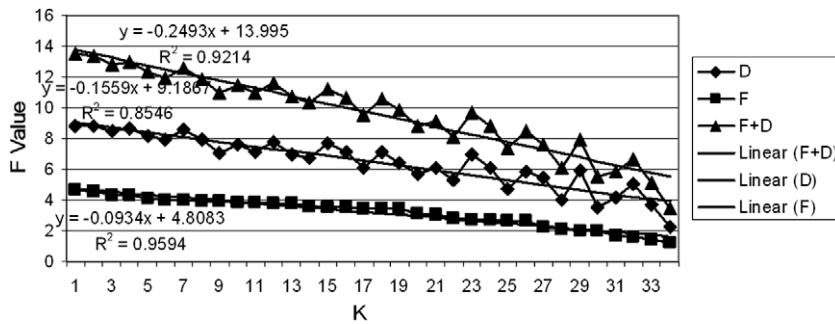

Fig. 2. Determine optimum K.

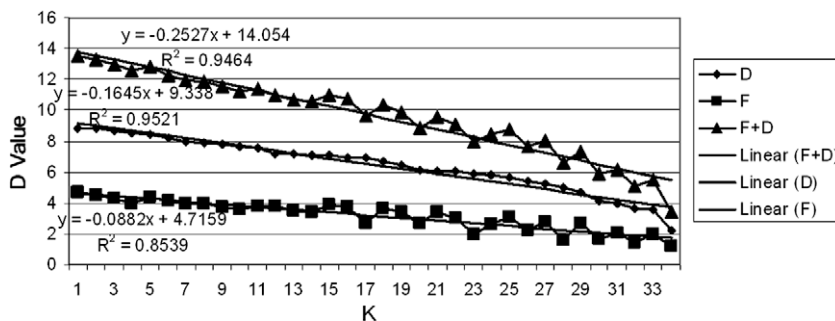
Fig. 3. Ranking of F with F, D, $F + D$ values.

Fig. 4. Ranking of D with F, D, $F + D$ values.
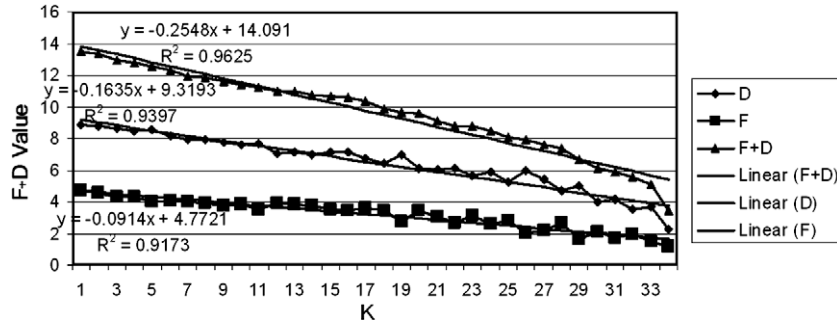
Fig. 5. Ranking of $F + D$ with F, D, $F + D$ values.
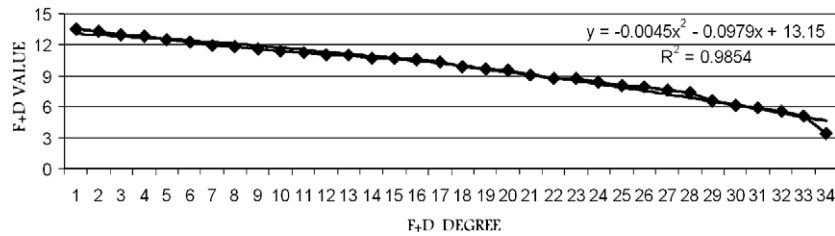

Fig. 6. Customer loyalty assessment function.

compare the proposed model with the non-weighted RFM model of three classes on output.

3. Empirical study

The developed methodology was implemented in SAPCO Co. This company was founded in 1993 and is one of the most leading supplier companies in Iran which provided items, equipment and car accessories for several car factories in Iran through some distributor agent companies such as ISACO Co. that usually provides the customer with product after-sale services. Recently she received the EFQM award by successfully auditing her company under Iranian National Productivity and Business Excellence Award.

A practical dataset of SAPCO Co. for automotive industry in Iran has been collected from 2008/3/20 to 2008/11/21.

The computing process using for SAPCO can be shown according to the Fig. 1 with these parameters:

R: supply percent of each product for ISACO;
F: frequency of the purchase of each product by ISACO;
M: monetary value of the order from ISACO site;

Fig. 7. The results of the application of proposed model in SAPCO.

L: period of product activity for above-mentioned period of time.

The matrix of group decision making for determining the relative importance weights of the RFML variables shown in Table 1.

Applying the concept shown in Fig. 1, the results have been demonstrated in Tables 2–7 and Figs. 2–5.

4. Discussion and finding

After assessing CLV ranking under two parameters namely D and F, the results indicate that when we combine these two parameters, we can obtain a better degree of loyalty which has been confirmed with R^2 value test and accuracy rate of the generated rules with Decision Tree and Artificial Neural Network methods, also we can obtain customer loyalty assessment function which has been shown in Fig. 6.

The application of proposed model is implemented in SAPCO, one of the leading car manufacturing supplying companies in Iran.

The application of clustering and classifying procedure used for the above-mentioned company indicates that out of 90% of total sale is classified into 40% of the target parts as shown in Fig. 7.

5. Conclusions

The proposed procedure has shown that for the purpose of clustering, when we combine the expanded WRFM model into K-means algorithm, we can see a tremendous improvement in classifying accuracy in order to reach to an excellent CRM.

As the distance and the integrated rate of each cluster commonly used by many researchers as a separate and independent parameters, in this study the combination of these two parameters has been considered in clustering and classification analysis. The result of statistical test for model validation has shown that the developed methodology for CRM has an acceptable result with a high level of confidence in comparing with other commonly used models by researchers.
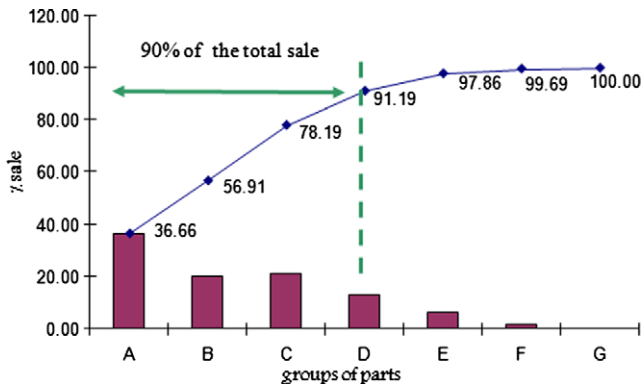
The proposed CRM methodology can be used by industries as well as service sectors in evaluating their customers loyalty in a most efficient and effective manners.

References

Anton, J. (1996). *Customer relationship management.* New York, NY: Prentice-Hall.

Anton, J., & Hoeck, M. (2002). *E-business customer service.* Santa Monica, CA: The Anton Press.

Baier, M., Ruf, K. M., & Chakraborty, G. (2002). *Contemporary database marketing: Concepts and applications.* Evanston: Racom Communications.

Brown, H. L. (2000). *"The three Ws of technology", customer relationship management: A strategic imperative in the world of e-business* (pp. 219–225). Toronto: Wiley.

Chatterjee, J. (2000). Managing customer relationships in the e-business economy. *Journal of Scientific and Industrial Research*, 749–752.

Cheng, C.-H., & Chen, Y.-S. (2008). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications*.

Chow, S., & Holden, R. (1997). Toward an understanding of loyalty: The moderating role of trust. *Journal of Management Issue, 9*(3), 275.

Eriksson, K., & Vaghult, L. A. (2000). Customer retention, purchasing behavior and relationship substance in professional services. *Industrial Marketing Management, 29*, 363–372.

Feinberg, R., & Trotter, M. (2001). Immaculate deception: The unintended negative effects of the CRM revolution: May be we would be better off without customer relations management. In D. H. Renner (Ed.), *Defying the limits, Vol. 2, Montgomery Research*, San Franscisco, CA.

Feinberg, R., & Kadam, R. (2002). E-CRM web service attributes as determinants of customer satisfaction with retail web sites. *International Journal of Service Industry Management, 13*(5), 432–451.

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques.* San Francisco: Morgan Kaufmann Publishers.

Hughes, A. M. (1994). *Strategic database marketing.* Chicago: Probus Publishing Company.

Lam, S. Y., Shankar, V., Erramilli, M. K., & Murthy, B. (2004). Customer value, satisfaction, loyalty, and switching costs: An illustration from a business-to-business service context. *Journal of the Academy of Marketing Science, 32*(3), 293.

Liu, D.-R., & Shih, Y.-Y. (2005). Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information & Management, 42*, 387–400.

Michaud, P. (1997). Clustering techniques. *Future Generation Computer System, 13*(2), 135–147.

Parvatiyar, A., & Sheth, J. (2004). Conceptual framework of customer relationship management. In J. N. Sheth, A. Parvatiyar, & G. Shainesh (Eds.), *Customer relationship management: Emerging concepts, tools and applications* (5th ed.). New Delhi: Tata McGraw-Hill Publishing.

Reichheld, F. F., & Teal, T. (1996). *The loyalty effect: The hidden force behind growth, profits, and lasting value.* Trans. Boston, MA: Harvard Business School Press.

Ricard, L., & Perrien, J. (1999). Explaining and evaluating the implementation of organizational relationship marketing in the banking industry: Clients' perception. *Journal of Business Research, 45*, 199–209.

Shina, H. W., & Sohnb, S. Y. (2004). Segmentation of stock trading customers according to potential value. *Expert Systems with Applications*, 27–33.

Stone, B. (1995). *Successful direct marketing methods* (pp. 37–57). Lincolnwood, IL: NTC Business Books.