Customer Classification and Market Basket Analysis Using K-Means Clustering and Association Rules: Evidence from Distribution Big Data of Korean Retailing Company

Article · December 2018

DOI: 10.15813/kmr.2018.19.4.004

3 authors, including:

Young-Chan Lee
Dongguk University Gyeongju
70 PUBLICATIONS **1,003** CITATIONS

SEE PROFILE

http://dx.doi.org/10.15813/kmr.2018.19.4.004

Customer Classification and Market Basket Analysis Using K-Means Clustering and Association Rules: Evidence from Distribution Big Data of Korean Retailing Company[1]

군집분석과 연관규칙을 활용한 고객 분류 및 장바구니 분석: 소매 유통 빅데이터를 중심으로

Run-Qing Liu (리우룬칭) Master's Degree Student of Global Business, Dongguk University[2]
Young-Chan Lee (이영찬) Professor of Business Analytics, Dongguk University[3]
Hong-Lei Mu (무홍레이) Ph.D. Student of Global Business, Dongguk University[4]

ABSTRACT

With the arrival of the big data era, customer data and data mining analysis have gradually dominated the process of Customer Relationship Management (CRM). This phenomenon indicates that customer data along with the use of information techniques (IT) have become the basis for building a successful CRM strategy. However, some companies can not discover valuable information through a large amount of customer data, which leads to the failure of making appropriate business strategy. Without suitable strategies, the companies may lose the competitive advantage or probably go bankrupt. The purpose of this study is to propose CRM strategies by segmenting customers into VIPs and Non-VIPs and identifying purchase patterns using the the VIPs' transaction data and data mining techniques (K-means clustering and association rules) of online shopping mall in Korea. The results of this paper indicate that 227 customers were segmented into VIPs among 1866 customers. And according to 51,080 transactions data of VIPs, home product and women wear are frequently associated with food, which means that the purchase of home product or women wears mainly affect the purchase of food. Therefore, marketing managers of shopping mall should consider these shopping patterns when they build CRM strategy.

Keywords: *Patent strategy, Patent portfolio, New product introduction, Patent rearrangement*

1) 논문접수일: 2018년 8월 29일; 1차 수정: 2018년 11월 13일; 게재 확정일: 2018년 11월 25일
2) 제1저자 (rachellau0813@gmail.com)
3) 교신저자 (chanlee@dongguk.ac.kr)
4) 공동저자 (434478018@qq.com)

1. Introduction

With the arrival of the big data era, the combination of customer data analysis and data mining techniques has gradually occupied a leading position in the CRM domain. In fact, analyzing customer data is very helpful for managers to make correct decisions. However, according to the previous literature review, most of the studies focus on large companies while ignore the small and medium sized enterprises(SMEs) environment. In addition, according to the development situation report of SMEs(KOSIS, Korean Statistical Information Service) and Global Bankrupt Report in 2016(Asia Pacific Partnerships), there are about 67,000 SMEs have been in a "halt production" or "bankrupt" situation compared with 2015 in Korea. They also found out that the average lifetime of Korean SMEs is just 10.62 years, which is very short compared to SMEs in Europe and Japanese. The report points out that the slump is ascribed to the lack of technological innovations and adoption of incorrect management style, especially the management style between companies and their customers. What is more, managers in SMEs also lack the ability to analyze customer data and discover useful information hidden in the data.

Customer value is the core of CRM for companies. In particular, high customer value could bring much more benefit than the customer value who is low. Analyzing and understanding customer behaviors and characteristics is the basis for developing competitive CRM strategies and maximizing the customer value. In addition, according to the marketing axiom of 80/20 rule, 80% of the enterprise's benefit are mainly come from 20% of the loyal customers or the VIPs. Therefore, this study presents the following research questions from the perspective of SMEs: (1) How to segment customers into VIP or Non-VIP customers? (2) What are the purchase patterns of VIPs? (3) How to use the customer data to propose CRM strategies?

Specifically, the purpose of this study is to segment the overall customers of a given Korea SME's customer data into VIPs and Non-VIPs through the method of recency, frequency, and monetary (hereafter RFM). Then, analyze VIP's transaction data to identify the most effective rules and patterns by adopting various data mining techniques such as K-means clustering algorithm, association rules. Since appropriate data mining tools are one of the best supporting tools for developing different CRM decisions and generating suitable CRM strategies(Berson et al. 2000), and the application of data mining tools for CRM is worth pursuing in a customer-centric economy. And this study aims to facilitate SMEs to maintain the loyal customers or VIPs by providing effective rules and patterns. Un-satisfaction or customers' churn will lead to unexpected loss, including both of the financial loss and nonfinancial loss. Thus, it is vital for SMEs managers to make appropriate strategies and decision-makings to manage the relationship with their customers.

2. Literature Review

2.1 CRM and Data Mining

Customer relationship management (CRM) is defined as a "business approach to understanding customer behavior through the meaningful communications in order to improve customer's acquisition, retention, loyalty, and profitability" (Swift, 2001). From an architectural point of view, the CRM framework is vital to company and can be divided into operational CRM and analytical CRM(He et al. 2004; Teo et al. 2006; 권재현·최영준, 2016; 강수영 등, 2011). Operational CRM refers to the automation of business processes, whereas analytical CRM is based on the enterprise data set, and supports the organization's customer management strategies through the analysis of customer characteristics and behavior. Data mining tools are popular methods for the analytical CRM framework, and are mainly used to discover useful information and knowledge hidden in the data. Data mining is defined as "the process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequently gain knowledge from large databases"(Turban et al. 2011). Many previous studies have used data mining techniques to analyze customer behavior and most of them have worked well. For example, Keramati et al. (2014) employed data mining classification techniques including decision tree, artificial neural networks to predict customer churn in the telecommunications business. Sheu et al. (2017) applied the decision tree algorithm to analyze the target attributes on measures of customer loyalty for animations, comics, and games consumers (ACG) industry, and identify the factors influencing internal and external influences on the ACG industry.

However, these situations often appear in large enterprises. In many SMEs, managers always lack capabilities in innovation, information technology and data integration analysis, so data resources cannot be configured in an optimal manner. On the other hand, since customers' individual information and transaction data are very difficult to obtain, only a few studies used data mining techniques to segment customers and identify association rules. Hence, this study segments the target customers into VIPs and Non-VIPs using RFM model first, then using K-means algorithm to classify VIPs, and then identify association rules based on VIPs and finally develop CRM strategies.

2.2 RMF Model

Kaymak(2001) noted that an RFM model is one of the widely used customer value analysis approaches for Target Customer Segment because it can reduce the complexity of the model of customer value analysis. R means the time interval between the last purchasing behavior and present purchasing behavior, F is the number of transactions among a certain period of time, and M refers to the amount of money spent on products or services over a certain period of time. The smaller of R with the higher of F and the higher of M, the more likely corresponding customer to buy products or services again from

the same companies(Wu et al. 2005). RFM value has a very important performance for customer segmentation. Some studies of data mining have adopted RFM values to analyze the relationship between enterprises and customers. For example, Chen et al. (2009) verified purchasing patterns in the form of sequential patterns based on RFM values. Ravasan and Mansouri(2018) proposed a brand new and practical fuzzy analytic network process based weighted RFM (recency, frequency, monetary value) model for application in K-means algorithm for auto insurance customers' segmentation.

2.3 K-means Clustering Algorithm

The purpose of K-means clustering is to partition observations into K clusters and each of the observation belongs to the cluster which possesses the nearest mean. K-means clustering algorithm has been utilized widely, including data mining, statistical data analysis, and other business context. The processes for K-means clustering are as follows: (1) partition the items into K initial cluster by associating every observation with the nearest mean; (2) assign an item to the cluster with the closest centroid and recalculate the centroid for the cluster after adding or losing an item; (3) repeat step 2 until reassigning is completed. Previous studies have employed K-means clustering with CRM successfully. For example, Bansal et al. (2017) adopted K-means clustering to classify customer usage space based on tens of dimensions, for multiple duty cycles, and over years of operation. Hosseini et al. (2010) adopted K-means algorithm to classify customer product loyalty in a B2B context based on the expanded RFM model.

2.4 Association Rules

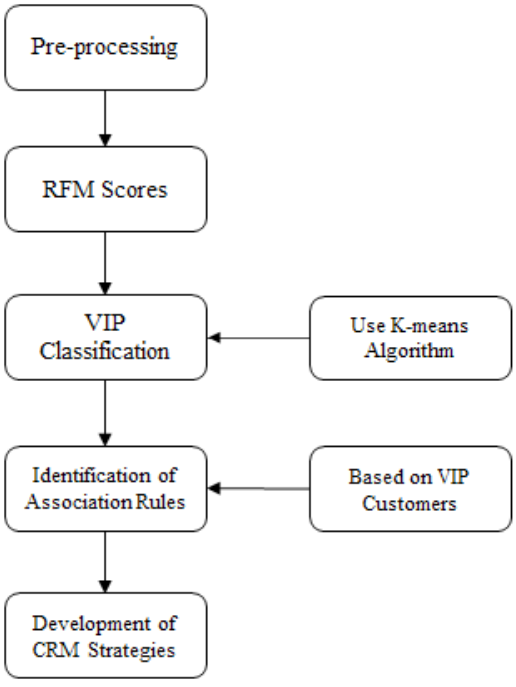
After segmenting customers into VIPs using RFM values and K-means clustering, our study would use the market basket analysis to identify VIPs characteristics and purchase patterns to help SMEs develop strategies. The most important application of the market basket analysis is the association rule. Association is one of the major techniques to detect and extract useful information from large scale transaction data(Hahsle and Karpienko 2017). It can answer some questions like what kind of products tend to be purchased together by customers. Rules with high confidence and strong support can be referred to as strong rules(Kantardzic 2003). Previous studies have adopted association rule with CRM area. For example, Tsai and Chen(2010) selected variables for customer churn of multimedia on demand according to association rules. Shim et al. (2012) proposed CRM strategies based on association rules and sequential patterns to analyze the transaction data of small-sized online shopping malls.

3. Research Design

3.1 Research Framework

<Figure 1> shows the framework of this research. First of all, we pre-processed the customer's data by deleting the duplicated records and

those with many missing values or inaccurate values. Then, we defined the RFM values of each customer based on the data provided by the target firm. Secondly, prior to the model construction, it is necessary to define the RFM scores and identify distribution features of these RFM scores. Then, standardized RFM scores to build clustering models. Thirdly, using K-means algorithm to evaluate and choose the best clustering model to determine the VIPs by selecting the important variables and drop those with low correlations with the target variables. It is important to identify and select variables which are taken on a good explanatory power. Fourthly, using the market basket analysis method to identify association rules from the transactions of VIPs which are determined by best predict model. Finally, used these patterns to propose CRM strategies for managers to provide better help.



<Figure 1> Research Framework

3.2 VIP Classification

In the context of customer management, the individual customers are the central object, which is always analyzed by data mining methods. In general, a very simple and flat data model is chosen as the basis for predictive modeling. In this representation, all data belonging to an individual customer is involved in an observation (row). Single columns (such as variables or fields) represent the conditions at particular points in time. Descriptive statistics such as sums, mean, median, and standard deviation will be employed to capture features of the related time series. Therefore, the selection of feature variables is critical to build a classification model. Another key variable to be created during this step is the target or dependent variable, needed for predictive modeling. Once a satisfactory definition of the target variable is achieved, its values will be generated for all customers, while added these values to the existing data tables. The next step is to select the best model to predict the dependent variable. The final step in the data mining project is acting based on final results. In this step, customers are scored and ranked to identify the right customers to target.

3.3 Market Basket Analysis

Market Basket Analysis (MBA) also known as association rule learning or affinity analysis, is a data mining technique that can be utilized in different fields, for example marketing, education field, nuclear science etc. The main purpose of MBA in marketing is to provide the information

to the retailer to understand the purchase behavior of the customer, which can help the retailer to make correct decision-making(Kaur and Kang 2016). The output of MBA is a set of rules that indicate the products that are purchased simultaneously. The result of output will be used as input variables for the prediction. In this study, once the VIP classification model is built, we attempt to use MBA method to find association rules among the item categories from VIPs created by the classification model.

4. Experimental Design and Analysis Results

4.1 Data Description

The dataset in this study is adopted from a shopping mall in Korea that is willing to provide their customer's transaction data. The shopping mall mainly sells items such as food, goods, cosmetic, and wear etc. online or in the physical store and the entire data includes customer's basic information and their purchase behavior information from June 2015 until June 2016. This study extracted 2,000 customers and 51,080 transaction data as basic data from customer groups of target shop. Contents of the dataset are summarized in <Table 1>.

4.2 Pre-processing and VIP Classification

After deleting duplicated and missing values or inaccurate values, 1866 customers were obtained. The descriptive information of 1866 customers are listed in <Table 2> analyzed through SPSS 21. According to the table, females are more than male and most of the customers are middle-aged people. Then prior to customer segmentation, RFM values need to be defined. <Table 3> shows partial data of RFM values. <Table 4> shows the

<Table 1> Data Record

Customer ID	Branch Code	Age
Gender	First Purchase Date	Final Purchase Date
Total Purchase Frequency	Total Purchase Amount	Average Purchase Amount
Average purchase cycle	Membership card	Online membership
Customer Relationship Type	Membership Card Join Date	Membership Self-Identification
Card Classification	Foreign Expensive Good Preference	Normal Good Preference Type
Most Expensive Good Preference	Multi-Branch Visit	Sensitive to Discount
Main Branch	Biz-Area	
Food	Goods	Cosmetic
Women Wear	Men Wear	Kids
Home Product	Home Appliance	Furniture
Total Purchase Category	Main Item Category	

descriptive statistical of the RFM values by analyzing the RFM values. Minimum and maximum of Recency is 1 and 366 respectively, which means some customers' last purchase was the day before cut-off date and some customers' last purchase was a year ago. Similarly, with Frequency and Monetary.

<Table 2> Descriptive Statistical of 1866 Customers

Variables	Groups	Frequency	Percentage %
Gender	Male	332	12.80
	Female	1534	82.20
	Total	1866	100
Age	<=20	1	0.10
	21~30	151	8.10
	31~40	538	28.80
	41~50	662	35.50
	51~60	370	19.80
	61~70	105	5.60
	71~80	35	1.90
	81~90	4	0.20
	Total	1866	100

<Table 3> Partial Data of RFM Values

ID	Recency	Frequency	Monetary
6141	129	6	261.57
9937	49	14	270.08
19053	7	30	888.60
22258	24	30	765.28
28214	12	75	2447.44
…	…	…	…

** Unit: 10 Thousand Won (KRW)

<Table 4> Descriptive Statistical of RFM Values

RFM Indicators	Minimum	Maximum	Mean	Std.
R	1	366	48.07	67.54
F	2	227	26.25	20.60
M	188.63	6915.51	659.68	655.57

However, it is not accurate enough to estimate VIPs only based on RFM values. Thus, this study adopted Analytic Hierarchy Process (AHP) and Expert Consultation Method (Delphi) to determine the weight of each indicator. To conduct an AHP survey, we invited 15 professors who are CRM expert. After pairwise comparison, we checked the consistency ratio and calculated geometric mean of the number of pairwise comparisons. Finally, we calculated the weight of RFM and obtained the result of 11, 19, and 70 using Expert Choice 2000. Divided the RFM score by five scaling, and finally, the RFM scores are shown in <Table 5> below. In addition, we could obtain 125 modules by dividing the RFM score by five levels. Then we obtained a descriptive statistical analysis and established Pivot tables based on the RFM score by using the Tableau software. The distribution features are shown in <Figure 2> below. It could be found that the RFM scores were roughly distributed across each module.

In order to further segment customer base, we employed IBM SPSS Modeler 14.1 to cluster the three fields of RFM. The clustering analysis method mainly includes Kohonen, K-means and Two-step algorithms. Before clustering, we normalized RFM values first and derive the Z-score for RFM by SPSS 21. The Z-scores of RFM are used as input variables for clustering analysis. When the number of clusters are determined to be 4, the segmentation result is the best according to the K-means results of SPSS Modeler 14.1. <Table 6> shows the results of clustering. <Figure 3> also shows the difference curve by using C5.0 algorithm.

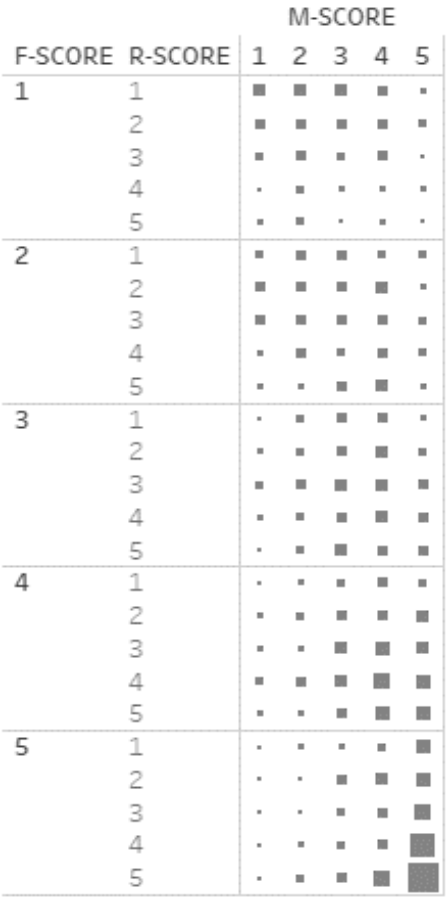
<Table 5> Real Scaling of RFM Attributes

Scaling	ID	R Score	F Score	M Score	RFM Score
5 Score	6141	1	1	1	100
4 Score	9937	2	2	1	130
3 Score	19053	4	4	5	470
2 Score	22258	3	4	4	389
1 Score	28214	4	5	5	489

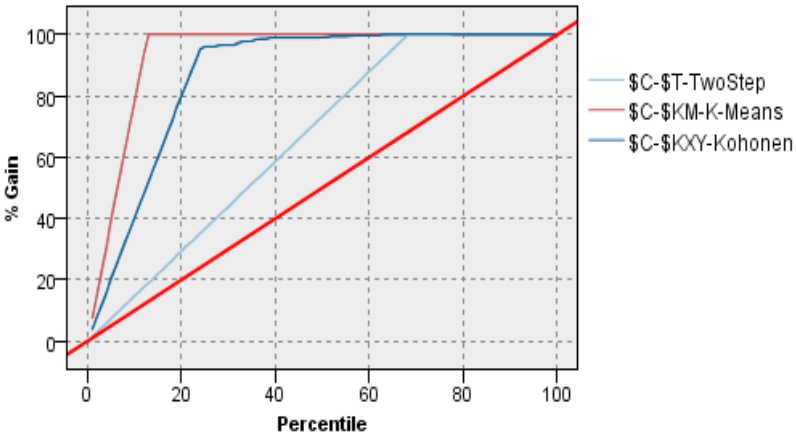
<Table 6> Clustering Situation of Three Algorithms

Algorithm	Cluster Quality	Clusters
Kohonen	0.3	12
K-means	0.6	4
Two Step	0.6	3

		M-SCORE				
F-SCORE	R-SCORE	1	2	3	4	5
1	1	6,000	7,650	5,760	4,340	1,520
	2	4,107	4,163	5,522	4,815	2,737
	3	2,806	3,840	3,406	5,644	804
	4	1,064	3,451	1,365	1,715	1,239
	5	2,160	3,424	852	1,416	848
2	1	2,737	4,536	4,662	3,619	3,591
	2	4,420	4,000	5,670	6,800	2,050
	3	4,230	4,642	5,620	5,265	3,789
	4	2,280	3,996	3,504	4,344	3,456
	5	2,119	2,097	3,939	5,968	2,215
3	1	966	2,912	5,560	4,524	1,672
	2	1,490	3,723	4,046	6,103	2,574
	3	2,400	4,830	6,900	7,030	4,400
	4	2,223	3,615	5,598	6,858	5,412
	5	1,092	3,780	5,796	3,920	5,544
4	1	1,099	1,816	2,376	4,771	3,496
	2	2,184	2,380	4,312	4,158	5,824
	3	1,432	2,241	6,699	8,947	5,967
	4	2,660	4,940	6,270	10,800	9,400
	5	2,211	1,626	5,456	9,453	9,139
5	1	528	1,476	2,212	3,088	8,208
	2	187	771	4,251	5,955	8,406
	3	396	804	3,380	4,080	12,428
	4	836	1,395	2,792	5,028	26,406
	5	440	2,320	4,680	10,750	43,000



<Figure 2> Distribution Features of RFM Scores



Algorithm	Accuracy
Kohonen	99.41%
K-means	99.73%
Two Step	96.78%

<Figure 3> Accuracy of Three Algorithms

According to <Figure 3>, K-means obtained the best curve, thus we established model based on K-means clustering algorithm. <Table 7> shows the feature of clusters. According to the results, cluster 3 could be selected as VIPs since value of R ($\bar{X}$= 15.586) was lower than the average R, and F ($\bar{X}$= 62.132), M ($\bar{X}$= 1093.515) were above average F and M values. What is more, the R value of cluster 3 is less than cluster 2 which means it could bring more economic benefits than cluster 2. Finally, the VIPs are 227 among the 1,866 customers.

It is also necessary to find other variables that will affect customers' segmentation in addition to RFM variables. Thus, 12 variables were analyzed by adopting Gain Ratio attribute evaluator on the basis of ranked search method. <Table 8> gives the descending order of importance of the 12 variables. The meaning of each variable is shown in <Table 9>.

<Table 7> Feature of Clusters

Clusters	N	Percentage %	R	F	M	RFM Level
			$\bar{X}$	$\bar{X}$	$\bar{X}$	
1	1323	70.90%	27.954	19.766	678.161	↓↓↑
2	78	4.18%	19.282	54.590	3070.102	↓↑↑
3	227	12.17%	15.586	62.132	1093.515	↓↑↑
4	236	12.75%	200.328	18.748	464.947	↑↓↓
Total	1866	100%	48.072	26.246	659.678	

** Unit: 10 Thousand Won (KRW)

<Table 8> Entropy and Gain Ratio of Variables

Items	Entropy	Gain Ratio
M	0	0.534
F	0.178	0.356
R	0.442	0.092
MEMBERSHIP_SELF_IDEN	0.506	0.028
AGE	0.514	0.020
BRANCH	0.522	0.012
GENDER	0.522	0.012
ZZ_CARD_FLG	0.524	0.010
MULTI_BRANCH_VISIT_YN	0.526	0.008
MOST_EXP_GOOD_PREF_YN	0.529	0.005
ENURI_SENSI_YN	0.530	0.004
FOREIGN_EXP_GOOD_PREF_YN	0.531	0.003

<Table 9> Meaning of Independent Variables

Attribute	Meaning
R	The time interval between the last purchasing behavior and current
F	The number of transactions over a certain period of time
M	The amount of money spent on products or services
MEMBERSHIP_SELF_IDEN	Membership Self-identify (For their own use=0, for family members use=1)
AGE	Customer's age
BRANCH	Gangna=1, Konkuk=2, Gwanak=3, Gwangju Outlet=4, Gwangju=5, GimhaeOutlet=6, Nowon=7, Daegu=8, Daejeon=9, Dongnae=10, Myeong-dong=11, Mia=12, BusanJung-gu =13, Busanjin-gu=14, Bupyeong=15, Bundang=16, Sangin=17, Anyang=18, Yeongdeungpo=19, Ulsan=20, Incheon=21, Ilsan=22, Jamsil=23, Jeonju=24, Changwon=25, Cheongnyangni =26, Pohang=27, Haeundae=28
GENDER	The gender of customers (Male=0, Female=1)
ZZ_CARD_FLG	The classification of card (No card=0, Public card=1, holiday card=2, Ordinary card=3)
MULTI_BRANCH_VISIT_YN	Whether customers will go to multi-branches shopping (Single shop=1, Multiple shop=2)
MOST_EXP_GOOD_PREF_YN	Whether customers prefer to buy most expensive goods (No=0, Yes=1)
ENURI_SENSI_YN	Whether customers will be sensitive to discount (No=0, Yes=1)
FOREIGN_EXP_GOOD_PREF_YN	Whether customers prefer to buy foreign expensive goods (No=0, Discount sensitive =1, Prefer to buy overseas brand-name goods=2, Buy overseas brand-name goods once=3, Overseas brand-name freak =4)

4.3 Market Basket Analysis Results

Association rules means transactions of the database which contain item X tend to contain item Y(Lee 2003). To discover association rules, we could use Apriori algorithm and its numerous variations. Apriori algorithm refers to that minimum support and confidence for the large scale effect generated by the rules, so the two thresholds are very important. After some attempts with different parameter values, we finally set the minimum support to 10%, the minimum confidence to 85% and the maximum number of antecedents to 3. Finally, 14,104 transaction data of the VIPs (227 customers) based on the result of classification need to be analyzed. An example of the transaction data as shown in <Table 10>.

The total number of rules are 29, and all of the rules satisfied both minimum support (10%) and minimum confidence (85%) set by us previously. While, there are 10 rules' confidence are greater than 90%, which were selected and shown in <Table 11>. Due to the support of this rule is 44.611%, and lift is 1.243, it means that this rule is highly reliable. The 3rd rule has the highest confidence level, which means that customer who purchases kids' product and cosmetic at the same time generally tends to purchase women wear together with each of them. Because of the target firm's VIPs are mainly middle-aged women, they are naturally concerned and purchased of these items, such as women wear, kids' product, and cosmetic, etc.

<Table 10> Example of the Transaction Data

Customer ID	Purchase Date	Purchase Items	Amount
9937	2015/06/06	(Food, Women Wear, Home Appliance)	186,742
	2015/07/03	(Cosmetic, Men Wear)	170,300
	2015/07/31	(Food, Men Wear)	154,325
	2015/08/04	(Food)	68,942
	2015/08/13	(Food, Cosmetic, Men Wear)	154,700
	2015/09/07	(Food, Women Wear, Men Wear)	196,980
	2015/09/23	(Cosmetic, Women Wear, Home Appliance)	175,490
	2015/10/17	(Food, Men Wear, Home Appliance)	287,843
	2015/10/23	(Food)	49,826
	2015/11/11	(Food, Cosmetic, Women Wear, Men Wear, Home Appliance)	435,630
	2015/12/26	(Food, Cosmetic, Women Wear, Men Wear, Home Appliance)	214,878
	2016/01/08	(Food, Home Appliance)	39,593
	2016/02/23	(Food, Home Appliance)	305,738
	2016/04/13	(Cosmetic, Women Wear, Men Wear)	259,853
Total	Purchasing Frequency: 14; The Number of Item Category:5		2,700,840

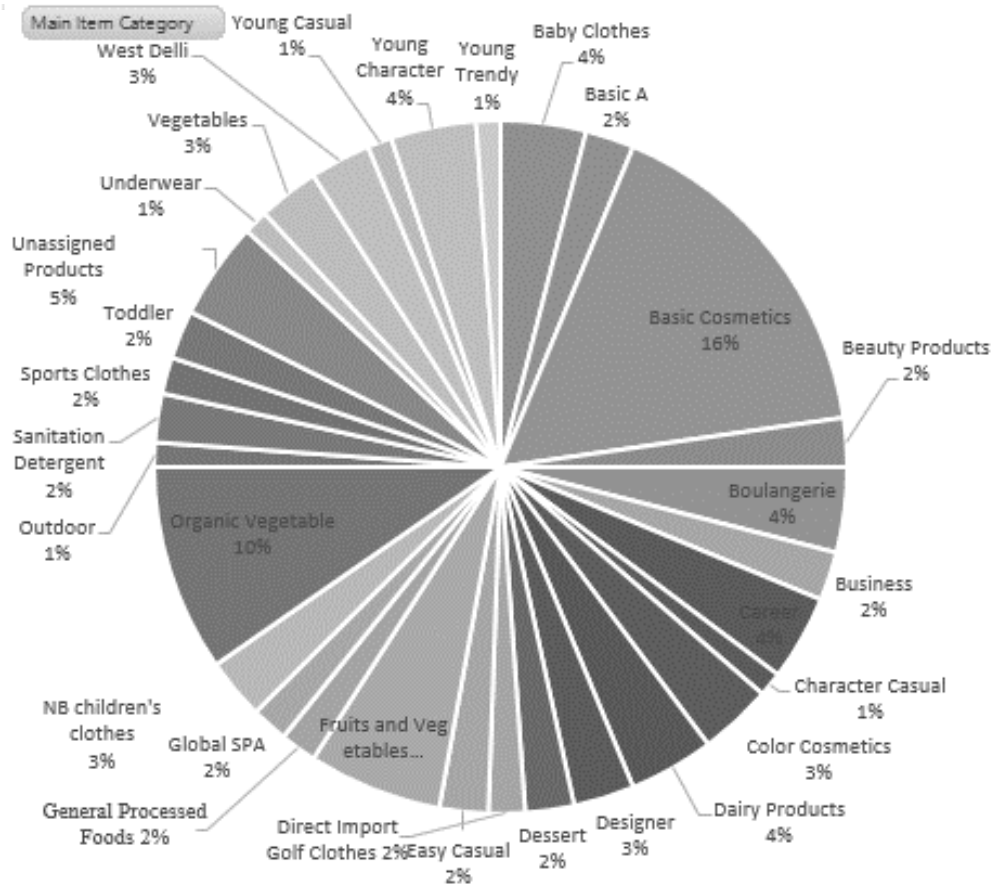
<Table 11> Top 10 Association Rules

Order	Consequent	Antecedent	Instances	Support %	Confid-ence %	Rule Sup-port %	Lift
1	FOOD	HOME_PRODUCT GOODS	6292	44.611	96.249	42.938	1.243
2	FOOD	WOMEN_WEAR HOME_PRODUCT	6350	45.023	96.220	43.321	1.242
3	WOMEN WEAR	KIDS COSMETIC	8337	59.111	93.175	55.077	1.203
4	WOMEN WEAR	MEN_WEAR KIDS GOODS	8205	58.174	92.724	53.942	1.197
5	FOOD	MEN_WEAR HOME_PRODUCT	8229	58.345	92.575	54.013	1.195
6	FOOD	WOMEN_WEAR HOME_PRODUCT COSMETIC	3079	21.831	92.043	20.094	1.188
7	FOOD	MEN_WEAR HOME_PRODUCT GOODS	2791	19.789	91.652	18.137	1.257
8	FOOD	HOME_PRODUCT KIDS WOMEN_WEAR	2791	19.789	91.508	18.108	1.182
9	COSMETIC	HOME_PRODUCT KIDS WOMEN_WEAR	8043	57.026	91.172	51.992	1.177
10	COSMETIC	KIDS WOMEN_WEAR FOOD	3451	24.268	90.351	22.107	1.239

<Table 12> reorganized the relationship between variables and consequents of the 10 rules. According to the table, food is frequently associated with other categories such as home product, women wear, men wear, goods, kids' product, and cosmetic. Of this, home product and women wear appeared the highest frequency, indicating that the purchase of them mainly affect the purchase of food. Similarly, women wear is frequently associated with other categories such as kids' product, men wear, goods, and cosmetic. Of this, kids' product appeared the highest frequency, indicating that the purchase of them mainly affect the purchase of food. In addition, cosmetic is frequently associated with other categories such as women wear, kids' product, home product, and food. Of this, women wear and kids' product appeared the highest frequency, indicating that the purchase of them mainly affect the purchase of cosmetic. This result also indirectly pointed out

<Table 12> Usage Frequency of Confidence Top 10 Rules

Variables	Frequency	Consequent
Home Product	6	Food
Women Wear	3	Food
Men Wear	2	Food
Goods	2	Food
KIDs	1	Food
Cosmetic	1	Food
KIDs	2	Women Wear
Men Wear	1	Women Wear
Cosmetic	1	Women Wear
Goods	1	Women Wear
KIDs	2	Cosmetic
Women Wear	2	Cosmetic
Home Product	1	Cosmetic
Food	1	Cosmetic



<Figure 4> Main Item Category

that customers of the target shop are mainly female customers.

This study also analyzed the main purchase item categories of VIPs to help better understand association rules, which are listed in <Figure 4>. According to the figure, Basic Cosmetics accounted for 16%, Organic Vegetable accounted for 10%, Fruits and Vegetable accounted for 6% which indicated that VIPs mainly focus on cosmetic, women wear, kids' products, food etc.

5. Conclusion

5.1 Academic Implications

The first academic implication is that this study successfully provides effective rules and new patterns by integrating the data mining techniques with CRM. The application of data mining techniques in CRM is an emerging trend in the industry. It has attracted the attention of practitioners and academics. In the development process of the big data era, data mining application can further improve the accuracy of rules and patterns, so that make it easier for enterprises to improve customers' satisfaction and loyalty, reduce customer churning intention. Second, this study mainly analyzes customers' behaviors from the perspective of SMEs, thus filled the gap in the SMEs context. Since most of the previous studies have employed CRM under the large companies' environment, while only a few studies focused on SMEs. When faced with fierce market competition, compared to SMEs, large enterprises dare to use new technologies to adapt to changing environment in the market, and formulate more active marketing strategies to make up a greater market share. However, that is not applicable for SMEs to only adopt conservative strategies. The result of this study is helpful for SMEs that posses small market share. It can also help SMEs correctly identify VIP customer groups to retain valuable customers, and provide a meaningful strategy to help enterprises to improve market share, firm market position and maintain a positive development process.

5.2 Practical Implications

According to the results, we defined VIPs on the basis of RFM values, and develop segmentation models that classify customers into VIPs or Non-VIPs groups. Simultaneously, we propose a decision-making framework which combines a trend prediction model. Then, we extracted 29 association rules from the transaction data of VIP customers and decided to come up with CRM strategies against VIP customers for the target firm as follows:

First of all, the classification model can be adopted to identify VIPs so that the target shopping mall can exercise marketing activities against them. We can design specific promotional activities for VIP customers and provide them with more accurate services to achieve long-term development of enterprises. For example, enterprises can identify VIPs and design specific service projects, such as sending promotional activities and related product catalogs via emails, which is

conducive to reduce costs and efficiently allocate resources.

Secondly, the results show that women wear is highly associated with the categories of men wear, kids' product, cosmetic, it is recommended that the target firm redesigns their website so that the page associated with women wear can be one-click away from men wear, kids' product, cosmetic, etc. Set up relevant marketing activities in accordance with the content of the association rules. It will improve customers' satisfaction, which is good for customer retention.

Thirdly, the target shopping mall can set recommendations on the portal and provide a preference for product advertising according to customers' search habits. The current keywords of the target firm in major search sites will suggest and induce customers to purchase product preferred. Therefore, it is recommended that the target firm resets current keywords would help to attract more potential customers.

Last but not the least. since VIPs are mainly middle-aged women who preferred cosmetics, women wear, home product, it is wise to encourage customers to buy products actively and set specific marketing activities for these VIPs. For example, when purchasing 50$ of A product, the target shop will present 5$ B product coupons as a gift. The target firm can improve customer satisfaction and loyalty by providing a personalized service, such as by offering customers a 20% discount on a certain product on the date of their birthday. Hopefully, the above suggestions could also be helpful to SMEs that aspire to increase their market share.

5.3 Limitations and Future Research

This study has some limitations. First of all, the history of the target firm is very short. Therefore, the dataset used for analysis is not sufficient to produce more meaningful results. Secondly, there are only 227 VIPs among 1866 customers, the representation of VIPs may exist bias. Thirdly, the prediction model construction is relatively simple. Every prediction model has its advantages, and we might obtain better results if we combine them together. Fourth, in the process of this study, it could be found that the lack of data would affect the accuracy of the prediction model. Finally, this study should continue and observe whether the strategies produced by actionable knowledge obtained by data mining are effective.

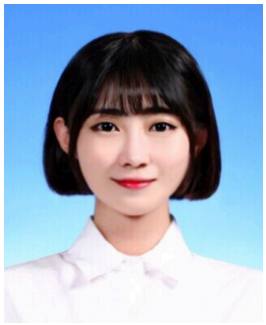
Based on the limitation of this study and the computational results, this study proposed some potential directions for future research. The first focus is on the construction of predictive models. The results show that each technique has its advantages and different approach often obtain different results. Hence, the present model in this study, which combines different prediction models, will become a valuable research direction. The second direction of future research is how to choose the most appropriate model on the basis of the real-world applications. Finally, this experiments show that the accuracy of the model is greatly influenced by the choice of parameters.

References

1. 권재현, 최영준 2016. "은행의 고객관계관리와 학습능력이 조직혁신성에 미치는 영향," *지식경영연구* (제17:3호) 227-248.
2. 강수영, 오평석, 김상만 2011. "고객 지식을 활용한 병원 CRM활동이 고객관계상태 및 향후 행동 의도에 미치는 영향," *지식경영연구* (12:3) 39-58.
3. Bansal, A., Sharma, M., and Goel, S. 2017. Improved K-means Clustering Algorithm for Prediction Analysis Using Classification Technique in Data Mining," *International Journal of Computer Applications* (157:6) pp. 0975-8887.
4. Chen, Y. L., Kuo, M. H., Wu, S. Y., and Tang, K. 2009. "Discovering Recency, Frequency, and Monetary (RFM) Sequential Patterns from Customers' Purchasing Data," *Electronic Commerce Research and Applications* (8:5), pp. 241-251.
5. Hahsler, M., and Karpienko, R. 2017. "Visualizing Association Rules in Hierarchical Groups," *Journal of Business Economics* (87:3) pp. 317-335.
6. He, Z., Xu, X., Huang, J. Z., and Deng, S. 2004. "Mining Class Outliers: Concepts, Algorithms and Applications in CRM," *Expert Systems with Applications* (27:4), pp. 681-697.
7. Hosseini, S. M. S., Maleki, A., and Gholamian, M. R. 2010. "Cluster Analysis Using Data Mining Approach to Develop CRM Methodology to Assess the Customer Loyalty," Expert Systems with Applications (37:7), pp. 5259-5264.
8. Kantardzic, M. 2003. *Data Mining-Concepts, Models, Methods, and Algorithms*, John Wiley & Sons.
9. Kaur, M., and Kang, S. 2016. "Market Basket Analysis: Identify the Changing Trends of Market Data using Association Rule Mining," *Procedia Computer Science* (85), pp. 78-85.
10. Kaymak, U. 2001. "Fuzzy Target Selection Using RFM Variables," IFSA World Congress and 20th NAFIPS International Conference, IEEE (2), pp. 1038-1043.
11. Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., and Abbasi, U. 2014. "Improved Churn Prediction in Telecommunication Industry Using Data Mining Techniques," *Applied Soft Computing* (24), pp. 994-1012.
12. Lee, Y. C., and Shin, S. I. 2003. "Mining Association Rules of Credit Card Delinquency of Bank Customers in Large Databases," *Journal of Intelligence and Information Systems* (9:2), pp. 135-154.
13. Ravasan, A. Z., Mansouri, T. 2018. "A Fuzzy ANP Based Weighted RFM Model for Customer Segmentation in Auto Insurance Sector," Intelligent Systems: Concepts, Methodologies, Tools, and Applications. IGI Global, pp. 1050-1067.
14. Sheu, J. J., Chu, K. T., and Wang, S. M. 2017. "The Associate Impact of Individual Internal Experiences and Reference Groups on Buying Behavior: A Case Study of Animations, Comics, and Games Consumers," *Telematics and Informatics* (34:4), pp. 314-325.

15. Shim, B., Choi, K., and Suh, Y. 2012. "CRM Strategies for A Small-sized Online Shopping Mall Based on Association Rules and Sequential Patterns," *Expert Systems with Applications* (39:9), pp. 7736-7742.
16. Swift, R. S. 2001. Accelerating Customer Relationships: Using CRM and Relationship Technologies, Prentice Hall Professional.
17. Teo, T. S. H., Devadoss, P., and Pan, S. L. 2006. "Towards A Holistic Perspective of Customer Relationship Management (CRM) Implementation: A Case Study of the Housing and Development Board, Singapore," *Decision Support Systems* (42:3), pp. 1613-1627.
18. Tsai, C. F., and Chen, M. Y. 2010. "Variables Selection by Association Rules for Customer Churn Prediction of Multimedia on Demand," *Expert Systems with Applications* (37:3), pp. 2006-2015.
19. Turban, E., Sharda, R., and Delen, D. 2011. *Decision Support and Business Intelligence Systems, Pearson Education India*.
20. Wu, C. H., Kao, S. C., Su, Y. Y., and Wu, C. C. 2005. "Targeting Customers Via Discovery Knowledge for the Insurance Industry," *Expert Systems with Applications* (29:2), pp. 291-299.

저 자 소 개



리우룬칭 (Run-Qing Liu)

동국대학교 경주캠퍼스에서 글로벌비즈니스협동과정 석사 학위를 취득하였다. 주요 관심 분야는 데이터마이닝, 지식경영, 사회적 자본, 혁신 및 전략경영 등이다.

이 영 찬 (Young-Chan Lee)

서강대학교 경영학사, 동 대학원에서 경영학 석사 및 박사학위를 취득하였다. 현재 동국대학교 경주캠퍼스 경영학부 교수로 재직하고 있으며, Annals of Management Science, The Open Operational Research Journal의 Editorial Board로 활동 중이다. 주요 관심 분야는 핀테크, 데이터마이닝, 다기준의사결정, 시스템 다이내믹스 등이다.

무홍레이 (Hong-Lei Mu)

현재 동국대학교 경주캠퍼스 글로벌비즈니스협동과정 박사과정으로 재학 중이다. 동국대학교에서 글로벌비즈니스협동과정 석사 학위를 취득하였다. 주요 관심 분야는 핀테크, 전자상거래, 모바일 쇼핑 등이다.