USE OF DATA MINING TO DERIVE CRM STRATEGIES OF AN AUTOMOBILE REPAIR SERVICE CENTER IN KOREA

Youngsam Yoon and Yongmoo Suh, Korea University, {mryys, ymsuh}@korea.ac.kr

ABSTRACT

Problems of a Korean automobile repair service center stem from the fact that a large portion of customers stop visiting the center for unknown reasons. So, we collected and analyzed the customer-related data in order to resolve those problems. We first defined customer class based on the R-F-M variables. Then, we built a customer classification model (i.e., a decision tree). Finally, we extracted characteristics (i.e., sequential patterns of services) of each customer class for each kind of automobile. What we learned is that the results of the analysis can be used together in order to provide better services so that total revenue of the center can be increased.

Keywords: data mining, CRM, R-F-M variables, decision tree, association rule, sequential pattern

INTRODUCTION

Firms in the world are facing competitive environments that get more and more complicated with various needs and wants of customers. Such complicated environment made them analyze and utilize various data and information. To survive such an environment, firms make use of data mining techniques to transform data into useful information and knowledge.

Automobile repair service centers are not exception in the above situation. Problems of a Korean automobile repair service center stem from the fact that a large portion of customers stop visiting the center for unknown reasons. In order to cope with the problems, we collected and analyzed the customer-related data provided by the automobile repair service center. We defined several customer classes based on R-F-M variables and built a customer classification model using decision tree algorithm (C5.0 node of Clementine). Then, we extracted characteristics of each customer class in terms of services, using association rule mining (Apriori node of Clementine), also, we were able to find differences in service patterns for each kind of automobile using sequential pattern mining (Sequence node of Clementine). What we learned is that the results of analysis can be used together in order to provide better services for the customers so that total revenue of the center can be increased.

The rest of this paper is organized as follows. First, we briefly introduce similar researches and related data mining techniques. Then, we describe the above data mining processes in detail. Lastly, we conclude with summary and limitations of the current work, and with future plan.

LITERATURE REVIEW

R-F-M (Recency, Frequency, and Monetary value) variables are used frequently in direct marketing, because they let marketers easily extract characteristics existing in customer purchase behaviors (Bauer, 1998). Recency means the time since the last purchase, frequency means the number of purchase occasions during a given period of time and monetary value means currency amount of purchase. Based on the values of R-F-M variables, marketers can calculate the customer score, assign classes to customers according to the calculated scores, and provide differentiated customer services for different classes of customers. R-F-M variables were also applied to solving diverse customer-related problems such as finding target customers in Kaymak et al. (2001), constructing customer response model in Suh et al. (1999), customer scoring in Lee et al. (2002), and proposing database marketing strategies in Ha et al. (1998).

Decision tree is used frequently for classifying by decomposing heterogeneous groups into homogeneous ones. The classification model constructed using train dataset can be used for predicting the value of target variable by classifying new records. Although there are various decision tree algorithms such as C5.0, CHAID, and CART, etc. their fundamental concepts are the same. But they differ in input data types and criteria for splitting and pruning a node while building a tree. Although we used C5.0 algorithm to build a customer classification model, it is used for various tasks such as finding loyal customers (Yada et al., 2000), predicting the probabilities of hypertension (Chae et al., 2001), predicting bankruptcy of firms (Lin et al., 2001) and predicting investment risk of nations (Becerra-Fernandez, 2002), etc.

Association rule mining developed by Agrawal et al. (1993) searches for interesting relationships among items in a given data set. Association rules are considered interesting if they satisfy both a minimum support and a minimum confidence. Although many other efficient algorithms have been developed since then, Apriori algorithm is still applicable to various problems, say, to get on-line recommendation in a retailing company (Wesley et al. 2001), knowledge from POS data in Taiwan (Lin et al. 2002) and personalized recommendation in an e-shop (Cho et al. 2002).

Sequential pattern mining is a technique which finds association rules from time sequence data such as telecommunication records, weather data, and production process data (Han and Kamber, 2001). It can be used to find *cause and effect* relationship from medical care records of patients (Berry et al. 1997).

DERIVING CUSTOMER CLASSIFICATION MODEL AND CUSTOMER SERVICE PATTERNS

Data

The data to which we have applied data mining techniques consists of four relational database tables, including records for 67 months from 1997/04/01 to 2002/10/31. *Owner* table includes demographic variables such as owner ID, social security number, etc. *Automobile* table includes variables related to automobiles such as the automobile registration number, the name of automobile, etc. *Customer registration* table includes variables related to registered events such as the date of registration, classification of repair types, etc. And *detail works after registration* table includes working codes, the name of work, etc.

Pre-Processing

Before applying data mining techniques to the data, it is necessary to conduct pre-processing such as deriving new fields, selecting input variables, changing different values which represent the same thing into a unified value, and eliminating useless records, etc.

First, we derived seven new fields such as *age, sex, R-F-M variables, and transaction period* from existing fields. Then, we selected nine variables as candidate input variables according to the opinion of domain expert and used chi-square test to check whether there is an association between the nine fields and the *customer class* field. After these statistical tests, we found that six variables out of nine, (i.e., *age*, *kilometers of traveling, customer transaction period, sex, classification of repair types,* and *regions*) were statistically significant under 95% significance level. In addition, we had to change the values of *service* field, since different values were input for the same repair service due to the lack of standard service names used in the center. Last, we eliminated useless records including null values, outliers, or redundant values.

Defining Customer Class

Another new field *customer class* is derived from *R-F-M variables* with the help of a domain expert as in Table 1, though there are many other ways for segmenting customers using these variables (Suh, 1999). Note that customers who belong to A, B or C class are differentiated as inactive customers according to the opinions of the domain expert, because they visited the center more than a year ago for the last time.

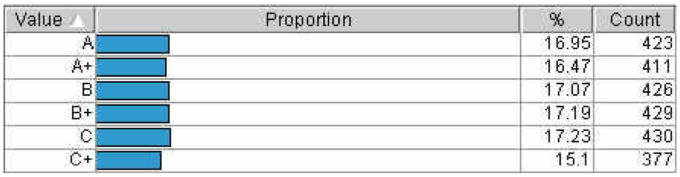
Table 1: classes of customers

Initial customer class	Description of initial customer class	Final customer class	Description of final customer class
A class	Top 20% of F and Top 20% of M	A+ class	Initial class = A and R ≤ 1 year
		A class	Initial class = A and R > 1 year
B class	Next 30% of F and Next 30% of M	B+ class	Initial class = B and R ≤ 1 year
		B class	Initial class = B and R > 1 year
C class	The Rest	C+ class	Initial class = C and R ≤ 1 year
		C class	Initial class = C and R > 1 year

R: Recency, F: Frequency, M: Monetary value

Customer Classification Model

In order to build customer classification model, we first split the input data of 7,148 records into a training dataset of 4,641(about 65%) records and a test data set of 2,507(about 35%) records. The number of records for each class is expressed as Figure 1 for test sample.



Value	Proportion	%	Count
A		16.95	423
A+		16.47	411
B		17.07	426
B+		17.19	429
C		17.23	430
C+		15.1	377

Figure 1: Number of records for each customer class

Each record consists of six input variables and one output variable. Then, C5.0 algorithm generated the prediction accuracy table, Table 2, after 10 times of boosting as we set. Compared with the prediction rate of a neural network model we had created, the prediction rate of the decision tree model was much better, especially for active customers.

Table 2: Prediction accuracies using C5.0 (test sample, measure: %)

Actual Predicted	A+	B+	C+	A	B	C
A+	68.764	20.783	0.375	1.942	2.454	0.165
B+	26.247	75.602	1.498	4.272	11.350	0.330
C+	0.000	0.000	98.127	0.000	0.000	18.997
A	2.820	2.108	0.000	65.049	20.859	0.495
B	2.169	1.506	0.000	28.738	65.337	10.561
C	0.000	0.000	0.000	0.000	0.000	69.472

Characteristics for Each Customer Class of Specific Kinds of Automobiles

Next, we derived characteristics of a customer class in terms of association rules of repair services or in terms of sequential patterns of repair services.

Table 3: Sequential patterns when time window=180 days

Consequent	Antecedent	Support	Confidence
Anti-freeze	Tensioner & engine tune-up > oil(engine/gear)	0.00216	100.0
	Ignition plug & tensioner > oil(engine/gear)	0.00216	60.0
	Control arm > air conditioner	0.00144	100.0
oil(engine/gear)	Door> muffler	0.00216	75.0
	Oil(engine/gear) > lamp	0.00144	100.0
	Oil(engine/gear) > alternator belt	0.00144	66.7
	Combination lamp > oil(engine/gear)	0.00144	66.7
	Windshield fluid > distributor	0.00144	66.7
Fender	Cap > axle shaft	0.00144	100.0
	Tensioner > axle shaft	0.00144	100.0
Air conditioner	Brush > wiring	0.00144	100.0
	Brush > engine tune-up	0.00144	66.7
	Brush > ignition plug	0.00144	66.7
	Brush > ignition plug & engine tune-up	0.00144	66.7
Engine cleaner	Brush > ignition plug	0.00144	66.7
	Brush > engine tune-up	0.00144	66.7
Alternator(assay)	Rotary > anti-freeze	0.00144	66.7

We used Apriori node of Clementine to find association rules. One of the rules we found for the class A+ is 'engine oil, anti-freeze → water pump' whose support and confidence are 10.9% and 62.7%, respectively. However, the support and the confidence of most of those rules are very low, because the number of input records is not large enough for this kind of analysis. We had better collect sufficient data from several repair service centers, if we are to get actionable rules. Such rules can be used as follows. For example, suppose there is a characteristic rule for class A+ such as (S1, S2 → S3), which implies that a customer who got services S1 and S2 also got another service S3. If we can find such a rule with high support and high confidence, we may recommend a service appearing in the right hand side of the rule to a specific customer who already visited the center for services appearing in the left hand side of the rule.

In order to do sequential pattern mining, we first transformed our data into sequence data and then selected three fields such as *customer-ID*, *sequence*, and *service*. We decided to derive sequential patterns of service only for the kinds of automobiles each of which takes more than 5% of records of the input data, considering the usefulness of the derived sequential patterns. For this process, we set minimum support to 0.1%, minimum confidence to 60%.

Table 3 shows the resulting sequential patterns obtained, when we set time window to 180 days. Each row of the table can be represented as X & Y > Z → W, where X & Y > Z corresponds to antecedent and W consequent in the table. The sequential pattern represented by the row can read as "60% of those who had got services X and Y at the same time and service Z within 180 days since then got another service W within 180 days.".

Use of Derived Knowledge to Provide Better Services

As a result of applying data mining techniques to a customer data, we obtain knowledge that can be used in business activities. Here, we suggest a few ways of using derived knowledge to provide better services to the customers of the repair service center.

First, when a customer comes in the center, the customer's class can be determined directly using customer classification model. So, the center may provide differentiated services to its customers, according to their class, such as assigning both specific service sectors and employees for customers of class A+, providing services such as oil or windshield fluid change, free of charge, for customers of class B+, and recommending future services using SMS (Short Message Service) for customers of class C+. Also, customer class information may be used to turn inactive customers into active by sending an SMS(Short Message Service) or e-mail, saying that customers of class A will be treated as VIP, some service charges will be discounted for the customers of class B, and presents will be given to the customers of class C just for visiting, using the media.

Table 4: A part of sequential patterns that have "air conditioner" as a consequent

number	Automobile ID	Antecedents
1	A1	Anti-freeze > clamp
		Clamp > light
		Heater > light
2	A2	Brush > engine tune-up
		Brush > ignition plug
		Brush > ignition plug & engine tune-up
		Brush > wiring

Second, if we compare the sequential patterns for each automobile class, we can find the differences among them. Table 4 shows a part of the sequential patterns that have "air-conditioner" as a consequent for two different kinds of automobiles. As we expected, the sequence of repair services in automobile type A1 is quite different from that in A2. Especially for automobile type A2, we can see that most customers who got air conditioner repair service received brush service and ignition and/or engine tune-up service in that order. In general, this kind of sequential patterns can be used to recommend a specific repair service to those customers who are selected based on the rules. Also, this result can show the differences in problems which occurs for each automobile.

Third, we can also generate sequential patterns for each customer class for each automobile type. Table 5 shows a part of the resulting sequential patterns for a specific kind of automobile for customer classes A+ and B+. From the table, employees can learn there are different service patterns for each customer class for each automobile type and recommend different future services for different customer classes. According to the first pattern for customers of class A+, when a customer of that class is provided with "ignition plug" service, the employee can recommend that the customer may need "alternator" service within next 6 months. After providing "alternator" service next time, the employee can recommend "oil (engine/gear)" service to that customer.

Table5: A part of sequential patterns for customer classes A+ and B+

Class	Consequent	Antecedents	Support	Confidence
A+	oil(engine/gear)	ignition plug > alternator	0.025424	0.75
A+	air conditioner	brush > wiring	0.016949	1
A+	fuel system	anti-freeze & ignition plug	0.016949	0.666667
A+	fuel system	belt & cam	0.016949	0.666667
B+	Distributor	belt & cylinder	0.012903	1
B+	ignition plug	Break & pump	0.012903	0.666667
B+	ignition plug	anti-freeze & break	0.012903	0.666667
B+	Wiring	Air conditioner & pump	0.012903	0.666667

There are many combinations of customer class, kind of automobile and time window. In each case, we can generate sequential patterns. Then, after analyzing the individual pattern carefully, the center can utilize it to provide better timely services to its customers, thereby reducing the rate of inactive customers.

CONCLUSION

We have created customer classification model using selected variables including derived ones. Using data mining techniques, we then generated diverse sequential patterns, which can be applied to a specific customer class and/or to a specific kind of automobiles. However, usefulness of those generated rules is in doubt currently, because the minimum support and minimum confidence are quite low. This happens due to the fact that the location of the repair service center is limited to a specific city of Korea.

Note that data mining is a cyclic process of getting knowledge to solve problems, acting on that knowledge, measuring the results of action, and identifying new problems. As we collect more and more data and analyze them, we expect more useful knowledge to be generated next time. It should be also mentioned that it is more important to examine carefully the derived knowledge before taking an action than just getting knowledge using data mining techniques. Also, measuring the effectiveness of the action is very important. We can measure customer satisfaction, revenue increase, the rate of inactive customers, etc.

REFERENCES

1. Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *SIGMOD'93*, Washington, D.C.
2. Barry, M. and Linoff, G. (1997), *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, pp.47~50.
3. Bauer, C. L. (1988), A Direct mail customer purchase model, *Journal of Direct Marketing*, 2:16-24.
4. Becerra-Fernandez, Irma, Zanakis, Stelios H. and Walczak, Steven (2002), Knowledge discovery techniques for predicting country investment risk, *Computers & Industrial Engineering*, Volume 43, Issue 4, Pages 787-800.
5. Chae, Young Moon, Ho, Seung Hee, Cho, Kyoung Won (2001), Dong Ha Lee and Sun Ha Ji, "Data mining approach to policy analysis in a health insurance domain", *International Journal of Medical Informatics*, Volume 62, Issues 2-3, Pages 103-111.
6. Changchien, S. Wesley and Lu, Tzu-Chuen (2001), Mining association rules procedure to support on-line recommendation by customers and products fragmentation, *Expert Systems with Applications*, Volume 20, Issue 4, Pages 325-335.
7. Cho, Yoon Ho, Kim, Jae Kyeong and Kim, Soung Hie (2002), A personalized recommender system based on web usage mining and decision tree induction, *Expert Systems with Applications*, Volume 23, Issue 3, Pages 329-342.
8. Han, J. and Kamber, M. (2001), *Data Mining: Concepts and Techniques*, Morgan Kauffmann.
9. Kaymak, U. (2001), "Fuzzy Target Selection using RFM variables", *IFSA World Congress and 20th NAFIPS International Conference*, 2001. Joint 9th, Volume: 2, Pages 1038-1042.
10. Lee, Yong Gu, and Jung, Sung Won (2002), Customer Information Analysis Model in Retailing, *Proceedings of The Korean Data Mining Society*, Pages 84-96.
11. Lin, Feng Yu and McClean, Sally (2001), A data mining approach to the prediction of corporate failure, *Knowledge-Based Systems*, Volume 14, Issues 3-4, Pages 189-195,
12. Lin, Qi-Yuan, Chen, Yen-Liang, Chen, Jiah-Shing and Chen, Yu-Chen (2003), Mining inter-organizational retailing knowledge for an alliance formed by competitive firms, *Information & Management*, Volume 40, Issue 5, Pages 431-442.
13. Suh, E. H., Noh, K. C., and Suh, C. K. (1999), "Customer list segmentation using the combined response model", *Expert Systems with Applications*, Volume 17, Issue 2, Pages 89-97.
14. Sung, Ho Ha and Sang, Chan Park (1998), Application of data mining tools to hotel data mart on the Intranet for database marketing, *Expert Systems with Applications*, Volume 15, Issue 1, Pages 1-31.
15. Yada, K.; Ip, E.H.; Hamuro, Y.; Katoh, N. (2000), "The discovery of Business Chance from Customer Knowledge"; Industrial Electronics Society, 2000, *IECON 2000*, 26th Annual Conference of the IEEE, Volume: 3, Page(s): 1638 -1643.