academicJournals

Vol. 7(24), pp. 2309-2318, 28 June, 2013
DOI: 10.5897/AJBM10.1465
ISSN 1993-8233 © 2013 Academic Journals
http://www.academicjournals.org/AJBM

African Journal of Business Management

Full Length Research Paper

Applied data mining techniques in insurance company: A comparative study of rough sets and decision tree

Kun-Shan Wu[1], Fang-Kuo Wang[2,3*] and Jhieh-Yu Shyng[4]

[1]Department of Business Administration, Tamkang University, Tamsui, Taipei 251, Taiwan
[2]Graduate Institute of Management Sciences, Tamkang University, Tamsui, Taipei 251, Taiwan.
[3]Department of Risk Management and Insurance, Ming Chuan University, Taipei, 111 Taiwan.
[4]Department of Information Management, Lan-Yang Institute of Technology, I-Lan 621, Taiwan

Accepted 13 March, 2013

Nowadays, customers are the essential elements of marketing for business operation. It is a critical and unignorable task in exploring valuable customers for companies and estimating customer values. According to the definition of Customer Life Value (denoted as CLV), a suitable model was found in this study and customers' present values were estimated by given data from insurance company. Two data mining technologies (Rough Sets Theory and Decision Tree) were introduced and applied to find the rules and factors which might have influence on customers' values. The comparing results of two technologies revealed that the influential and important factors for both technologies were similar but not for the minor factors. Both technologies performed well in efficiency of analysis but were different in interpreting results. It suggested that the rules generated from both technologies could serve as the auxiliary factors in practical marketing strategies.

Key words: Customer lifetime value, insurance industry, rough sets theory, decision tree.

INTRODUCTION

Customer relationship management was focused on in marketing since the early Eighties, which was another approach for business to access customers. As well known, customers were a valuable asset to companies and so became a popular concept that is accepted universally (Blattberg and Deighton, 1996; Srivastava et al., 1998; Reinartz and Kumar, 2000). Previous researches focused on establishing the relationship between business and customers in order to enhance the retention rate and gain benefit for the company. However, some customers are profitless to companies because their costs are too high, and might lead to a negative return on investment (Reinartz and Kumar, 2000). So, it is clear that evaluation of each customer's value is an important key for a company that wants to succeed.

Many researches, based on Customer Life Value (denoted as CLV) definition, developed CLV related models which included RFM (Recency/ Frequency/ Monetary) model, probability model, econometric model, persistence model, computer science model and diffusion/growth model (Gupta et al., 2006). RFM model was applied in direct marketing over thirty years and was only good for retailing. Probability models, that is, Pareto/NBD model and Markov chains, assumed that behaviors of customers are different across the population based on some probability distributions. Econometric model used hazard models to estimate customer retention. Persistence models focused on the behavior of components, and treated these components as part of a dynamic system if long-time series were available. Computer science models applied computer technologies, such as data mining and machine learning, to acquire

*Corresponding author. E-mail:- fkwang@mail.mcu.edu.tw. Tel: +886-2-0919935669. FAX: +886-2-28809755.

predictive ability. Diffusion/growth models focused on aggregating individual CLV to customer equity. Donald (1989) calculated customer value based on health insurance and proposed that the insurance customers' value could be calculated by deducting claim and marketing costs from annualized premium; then, adding invest income as well as considering lapse rate to evaluate policyholder value. As well known, assumptions vary in different insurance products which can not calculate customers' values for all companies. Other researchers focused on the prediction of customers' potential value (Verhoef and Donkers, 2001), the estimation of customers' values of banks' enterprise clients by using customer relationship management skills (Ryals, 2005) and the comparison of customers' value by using different models (Donkers et al., 2007).

Other than traditional statistic methods described above, some newly aroused computer technologies, such as data mining, played a nonignorable role in data analysis which could shorten computational time as well as improve precision of prediction. Dwyer (1997) reported that insurance depended on empathetic and competent sales support, innovative technical capability. To apply new technologies efficiently in insurance has become an important method in this competitive environment. Two data analysis technologies were applied in many fields: Rough Set Theory (RST) and Decision Tree (DT). DT was popular and applied in predicting airline customers' future values (Tirenni et al*.,* 2007), discovering knowledge for the insurance industry by using ID3 algorithm (Wu et al*.,* 2005), image classification (Yang et al*.,* 2002), implementing a recommender system (Cho, et al., 2002). Rough sets theory is a newer data mining technology compared to DT. It has been applied in the research fields of medical diagnosis (Tsumoto, 2004), engineering reliability, expert systems, machine diagnosis (Zhai et al., 2002), failure prediction (Beynon and Peel, 2001; Dimitras et al., 1999); some researchers reported that RST had a good results in analyzing insurance customers' characteristics (Shyng et al., 2007).

The insurance market sale was based on agent and customers in face-to-face interaction, which needed to deliver time and energy to break ice between them. As the progress of information technologies and society changes in recent years, direct marketing became an important element for insurance industry gaining profits, with a large portion of transaction. Data collection and analysis were the basis of direct marketing, especially for insurance industry. It was known that insurance company held vast customers' data including demographics of customers and transaction details; it was not applied in marketing and its precious value was not known.

In this study, we collected the valuable data which was provided by a life insurance company in Taiwan as the research target. A CLV model proposed by Gupta and Lehmann (2003) served as a base to calculate CLV. Two data mining technologies, RS and DT, were used to extract variables and generate rules. The analyzed results were interpreted and compared in way of convenience and precision. Both results could provide a good insight for initiating marketing strategies for insurance practicing.

Customer Lifetime Value (CLV)

Many models were proposed to deal with CLV calculation which was defined as the net profit or loss to the firm from a customer over the entire life of transactions of that customer with the firm. Thus, the CLV for a firm could be addressed as the net of the revenues obtained from that customer over the lifetime of transactions with that customer minus the cost of attracting, selling, and servicing that customer, taking into account the time value of money (Berger and Nasr, 1998; Gupta et al., 2006). CLV definition could be expressed as:

$$CLV = \sum_{t=0}^{T} \frac{(p_t - c_t) r_t}{(1+i)^t} - AC \quad (1)$$

where

t = time horizon for estimating CLV;

p_t = price paid by a consumer at time *t*;

c_t = direct cost of servicing the customer at time *t*;

i = discount rate or cost of capital for the firm;

r_t = probability of customer repeat buying or being "alive" at time *t*;

AC = acquisition cost.

Equation (1) just was a simple mathematic form that might not fully reflect CLV. Various models were developed to estimate CLV. RFM Model, as depicted above, had been used extensively more than 30 years that applied three variables (Recency, Frequency, and Monetary) into model to set cells and predict the future values of customers by given different weights. RFM model had been applied in direct marketing and retail business but it came with limitations (Fader et al., 2005; Kumar, 2006). First, the model only could predict behavior in the net period. Secondly, it ignored the fact that consumers' past behavior might be a result of firm's past marketing activities. Furthermore, it was useful for commodities prediction but not suitable for low purchasing frequency, such as insurance. Gupta and Lehmann (2003 and 2005) reported that if margins $(p_t - c_t)$ and retention rates were constants over time and used as an infinite time horizon, then CLV could be simplified to:

$$CLV = m * \frac{r}{(1+i-r)} \quad (2)$$

where *m* = margin profit;
r = retention probability of a customer;

i = discount rate or cost of capital for the firm.

Accordingly, equation (2) was adopted to calculate CLV in this study for three reasons:

1. Purchase patterns of customers were relatively stable in insurance market;
2. Finance services had fewer frequencies of use than other commodities;
3. Interest and retention rate were the important factors affecting premium, and were critical considered in determining the decision of purchasing.

Rough Sets Theory (RST)

RST was proposed by Pawlak in 1982; afterward, Walczak and Massart gave detailed discussion and presented RST foundation in 1999. RST reflected the nature of human, using the indiscernibility relation and perceptible knowledge to process the classification ability. Over two decades of development, RST was applied to the management fields of: data analysis and knowledge discovery; empirical study of material data (Jackson et al., 1996), machine diagnosis (Zhai et al., 2002), activity-based travel modeling (Witlox and Tindemans, 2004), business failure prediction (Beynon and Peel, 2001; Dimitras et al., 1999), solving linear programs (Azbi and Vanderpooten, 2002), data mining (Shyng et al., 2007; Li and Wang, 2004; Hu et al., 2003; Chan, 1998). Also, it was useful in exploring data patterns because of its ability to search through a multi-dimensional data space and determine the relative importance of each attribute with respect to its output.

Basically, RST presumed every object associated with information (data, knowledge) which constituted of similar indiscernible and available information. Any set of indiscernible objects was called an elementary set which formed a basic granule of knowledge (Dimitras et al., 1999). The data or information were grouped and extracted to generate rules which could illustrate each object set and corresponding attribute comprehensively.

From the information table, we understood that the 4 - tuple $IM = (U, A, V, g)$, was called information function (Pawlak, 1991). Where *U* denoted the universal object sets of *IM*, *U* consists of *n*'s objects, *A* was a finite set of attributes/features, $V = \bigcup_{i \in A} V_i$ and V_i was a set of values of the attributes. Let $g: U \times A \to V$ be a description function; and let *g*(*x*) be the description of *x* in *IM*, where $g(x,i) \in V_i$ was for every $i \in A$ and $x \in U$ (Pawlak, 2002). Assuming a family $Y = \{X_1, X_2, ..., X_m\}$ was a family of nonempty sets (classification) that $X_i \subseteq U$, $X_i \neq \varnothing$, $X_i \cap X_j = \varnothing$ were for $i \neq j$, $i, j = 1, 2, ..., m$ and $\bigcup_{i=1}^{m} X_i = U$. Any subset *B* of *A* determined a binary relation *IND(B)* on *U*, which we called an indiscernibility relation, and defined it as $a \in B$, if $g_{x_1}(a) = g_{x_2}(a)$ for every $a \in A$. The equivalence class of *IND*(*B*) was called an elementary set (of atoms) of *IM*. The partition of *U* with respect to *IND*(*B*) was denoted by *U*/*IND*(*B*). Thus, any x_i of *U* could be induced so that the value sets of attributes represented in *B* were in the same class.

Assuming $IND(A) = IND(A - a_2)$, the a_2 could be taken as a superfluous attribute. The superfluous attributes were removed, which simplified the information set and generated diagnostic values. If R was an equivalence relation, then, $\underline{R}X = \{x \in U : [x]R \subseteq X\}$ was the lower approximation; $\overline{R}$ X= $\{X \in U : [x]R \cap X \neq \emptyset\}$ was the upper approximation; $Bnd(RX) = \overline{R}X - \underline{R}X$, the boundary region of *X* that the objects were inconsistent or ambiguous (Shyng et al., 2009). The computation of lower and upper approximation was the major factor in decision rule extraction.

Assuming that the attribute set *A* of *IM* was divided into condition attributes (denoted as *CA*) and decision attributes (denoted as *DA*). The classification of *CA* and *DA* generated as condition and decision classes in the *IM* information (Shyng et al., 2009).

The purpose of reduction was to improve the precision of decisions and the process of attributes was to reduce elementary set numbers. By given an attribute space $A = (CA, DA)$, where $CA \neq \varnothing$ and $DA \neq \varnothing$; then, $CA \bigcup DA = A$ and $CA \cap DA = \varnothing$; which were the elements of the decision table. Let $RED(B) \subseteq A$; *RED*(*B*) was the reduction set which was composed of a set of attributes *B*. The intersection of all reduct attribute sets was the core attribute set which was the most important attribute in the decision-making process; $COR(C) = \bigcap RED(B)$ where *COR*(*C*) was the core composed of a set of attributes *C*. Applying the reduction set to the information, we could induce the decision rules. Due to the functional dependencies between conditions and decision attributes, a decision table might be deemed as a set of decision rules. The syntax could use the "if …., then…." rule to specify as "if …., then…".

Decision Tree (DT)

DT was one of the data mining techniques. It could compute quickly, attain nonlinear mapping, interpret rules easily, and had an embedded ability for feature selection (Hautaniemi et al., 2005). Comparing with traditional statistical analytical tools, such as Discriminate analysis and Regression analysis, the limitation of data input for DT was relatively less but much better in the prediction accuracy and interpretability. Therefore, the application of

DT became popular in all fields because of those advantages.

DT offered a nonlinear classification method that could avoid ineffective clustering resulting from linear method (Moshkovich et al., 2002), it was a tree-like flowchart which could classify data through chosen variables and designed goals. It tested root nodes and separated data sequentially; then, classified data by decision rules. Each branch and leaf node represented the testing results and the distribution of focal variables. DT rules could be extracted from each route which connected with root node and leaf node. It could form a hierarchical structure of classification system or prediction model that performed a structural knowledge. Also, DT might have different decision rules resulting from different tree-like flow charts (Han and Kamber, 2001; Chelghoum, 2002; Moran and Bui, 2002; Pal and Mather, 2003).

Nowadays, many DT algorithms were developed that included Interactive Dichotomiser 3 (ID3) (Quinlan, 1986), chi-squared automatic interaction detection (CHAID) (Kass, 1980), classification and regression tree (CART) (Breiman et al., 1984), C4.5 (Quinlan, 1993) and C5.0. ID3 algorithms selected attributes that were based on information entropy to classify the data table without considering the dependency between attributes and the importance of the attributes to the classification. Attributes produced the most rapid decrease in information entropy that did not always provide the maximal contribution to classification. C4.5 was the extension of ID3 algorithm that was used in a supervised classification problem; it used information gain rate to represent the information quantity in each layer and to serve as branch principles. Error Based Pruning (EBP) was used as a calculation basis in C4.5 (Lim et al., 2000) which could determine the optimal subset of a control system for selecting attributes among continuous attributes, noise data and alternative measures (Shiue and Guh, 2006).

C5.0 was a top-down tree and adopted the "divide and conquer" to establish DT to improve performance (Quinlan, 2003) and the choice of branch attribute principles was based on the attributes with the maximum information gain serving as the branch nodes. C5.0 was modified from C4.5 that could be operated either in Unix or Windows platform. There were some advantages of C5.0, such as dealing with continuous independent variables without branch limitations, using a boosting technique to generate and combine multiple classifiers, and improving prediction accuracy; it also could be represented as sets of "if – then" rules to improve human readability. Therefore, previous researches (Mooney et al., 1989; Weiss and Kapouleas, 1989; Fisher et al., 1989) presented various algorithms without assuring the best algorithms among them. Considering the data content and efficiency, C5.0 was chosen as the DT algorithm in this study.

The proposed DT was constructed from a data set *S* by using training criteria, Gain Ratio, that was a measure of incorporated entropy (Quinlan, 1993; Frey and Fisher, 2003; Ranilla et al., 2003; Dombi and Zsiros, 2005; Hu et al., 2007). The procedures of Gain Ratio are defined as follows (Chao et al., 2008). For simulated *IR* data, assuming a training data set *S* consisted of *C* class examples. The function *p(S,i)* was the ratio for the class number of an *IR* data set that belonged to class *i* of the total class number $|S|$ of an *IR* data set *S*, where $1 \le i \le C$. The entropy was defined as:

$$Entropy(S) = -\sum_{i=1}^{C} p(S,i)\log_2 p(S,i)$$

Presuming *T* was a feature containing total partitions *v*. The value *j* was any specific value of *v*, S_i was a subset of the *IR* data set *S* corresponding to the value *j* of *T*. The information gain, Gain(*S*, *T*), corresponding to the partitioning of *S* from feature *T*, was calculated:

$$Gain(S,T) = Entropy(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} Entropy(S_j)$$

where $|S_j|$ was the number of subsets S_j in the *IR* data, and *Entropy*(S_j) was calculated as the same as *Entropy(S)*. In order to obtain a good generation by reducing bias, wherein, the *SplitInfo* was defined as:

$$SplitInfo(S,T) = -\sum_{j=1}^{v} \frac{|S_j|}{|S|} \log_2 \left(\frac{|S_j|}{|S|} \right)$$

The function *Gain*(*S*, *T*) was very sensitive to the value of *v*, so the ratio of information gain was calculated as:

$$Gain\ Ratio(S,T) = \frac{Gain(S,T)}{SplitInfo(S,T)}$$

C5.0 algorithm of DT served as the classification method in this study because: it could provide valuable insight to carry out a classification task which had been proven as a good tool in previous researches (Huang and John, 1997; Eklund et al., 1998); it had rapid calculation ability compared to other classification tools (Tirenni, 2007); it had the better flexibility to fulfill our research needs.

RESEARCH METHODS

The study is designed as follows. First, reviewing of all CLV models published previously and using them to calculate current values of customers. Then, we applied RST and DT to select useful variables and found out the rules affecting customers' value factors, based on the customers' information (such as age, gender, annual

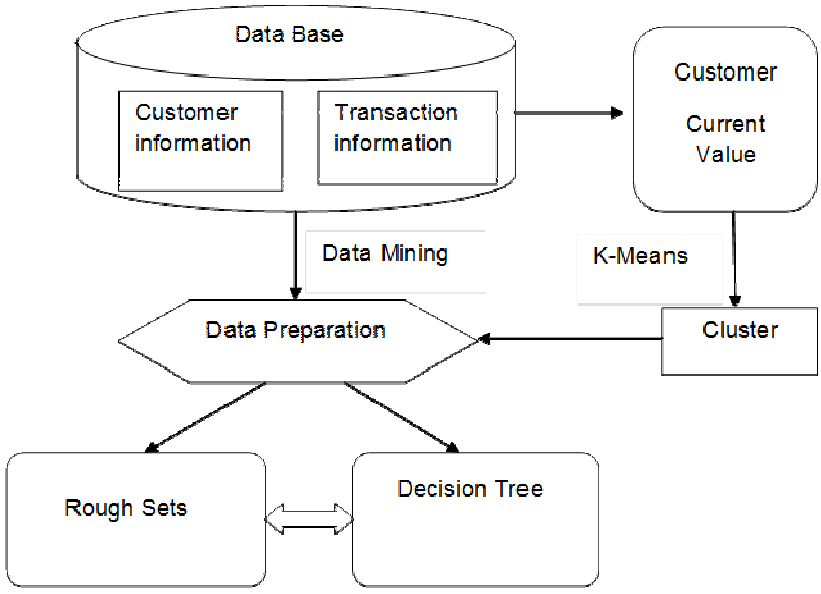

Figure 1. Research process.

income and occupation) and internal information (such as category of insurance, amount of insured and gross premium) provided by an insurance company. The results and efficiencies were compared as well as the advantages of these two methods in this study. The process is shown in Figure 1.

Case study

Data description

This study accessed customers' information provided by a life insurance company in Taiwan. Due to vast number of customers, we focused on the residence of customer located in metro Taipei area only. The reason we chose metro Taipei area was: Taipei was the major economic and political center with mass population that had significantly demographic characteristics. The customers' data were about 36,000 pieces, some of them were abnormal. After carefully screening, we got 25,584 customers' data. On the other hand, because of interest rate declining rapidly in recent years, insurance company needed to replace the policy of products with high interest rate by lower ones of which 33 insurance products were within data. In order to simplify the analytical procedure, we selected top 10 of most insurance products which took about 90% of total transaction, and, the top ten product policies are shown in Table 1.

Data Preparation

Socio-demographic variables, such as social hierarchies, life style, asset allocation would affect the purchase of insurance products (Wu et al., 2005; Verhoef and Donkers, 2001; Jackson, 1989). Because of the limitation of data content, the above variables could only be investigated individually that could not be acquired from internal database of insurance company. So, we selected basic customers' information (age, sex, occupation, annual income, etc.) and insurance transaction data (category of product, amount of insured, master contract premium, Gross premium, channel of payment, etc.) as the variables in this study (Wu et al., 2005).

During the information sorting, we processed the data granulation to avoid generating complicated principles which might make it difficult in taking continuous data. The data granulation procedure could transform continuous data into various categories. After granulation, eleven attributes were established and transferred to categories.

Cluster

Because of vast amount of data, we only focused on exploratory data to treat data classification. K-means (MacQueen, 1967) was applied to cluster data that was taken as an unsupervised method for data analysis. It was the most popular clustering algorithm that had been used in many domains, such as image segmentation (Marroquin and Girosi, 1993); marketing research (Punj and Stewart, 1983) and data miming (Fayyad, 1996).

There was no consensus on the optima number of categories to choose properly in previous researches (Salazar et al., 2007; Ray and Turi, 1999; Fraley and Raftery, 1998). There were three criteria for choosing

Table 1. Product category.

Product	No. of customer	Ratio (%)
Term life (TL)	1244	4.9
Participating Increasing Sum Assured Whole Life (PL)	7398	28.9
Whole Life, whole life payment (WL)	2837	11.1
Endowment (E)	786	3.1
Participating Whole Life (LPL)	9230	36.1
Dread Disease Whole Life (PDL)	1142	4.5
New Limited Payment Whole Life (NLPL)	1165	4.6
Dread Disease Whole Life (Non-participating insurance) (XPDL)	520	2.0
Endowment Whole Life (LES)	653	2.6
Endowment Whole Life (LESA)	609	2.4

Table 2. Data granulation.

Attribute	Item	Principles of classification									
V_1	No. of policy	Identification									
V_2	insurance	1	2	3	4	5	6	7	8	9	10
V_5	Gender	1(Male)							0(Female)		
V_6	terms	1	2	3	4	5	6	7			
V_7	Age	1 (0-14)	2 (15-24)	3 (25-34)	4 (35-44)	5(45-)					
V_8	duration	1(1-10)	2(11-20)	3(20-)							
V_9	Amount of insured	1(1- 200K〕	2(200K – 500K〕	3(500K-1000K〕	4(1000K-2000K〕	5(2000K-above)					
V_{10}	Premium (Master contract)	1(0 - 2500)	2(2501-5000)	3(5001-7000)	4(7001-12000)	5(12001-above)					
V_{11}	Annual Income	1(1 – 360K〕	2(360K-500K〕	3(500K- 1000K〕	4 (1000K-above)						
V_{12}	Gross Premium	1(1 - 7000)	2(7001-12000)	3(12001- 16000)	4(16001- above)						
V_{13}	Occupation	1(Government / military/ Education)	2 (Medical and clinic)	3 (commercial and Finance)	4 (Engineering and information)	5(Others)					

(USD 1=NT $ 31).

category that were accuracy rate, speedup of the convergence for data clustering, and the each cluster evenly. Then, we used the discriminate analysis to serve as the basis of comparison for clustering relevance. In this study, we tried to use 2, 3, 4, and 5 categories to cluster 25,584 data and to determine the accuracy, shown in Table 3.

Based on the criteria depicted above, the more the categories, the accuracy rate becomes worse. The speedup of the convergence for data clustering was not significantly different among four categories while using computer to compute. The clusters showed an evenly distribution. Thus, we chose cluster 2 as the basis of data analysis.

DISCUSSION

Rough sets theory

Through the analysis procedure as depicted above, there were 10 attributes: 9 were condition attributes and one was a decision attribute. Those data could be divided into two groups, high and low CLVs, by k-means that covered 10 policies and 25,584 customers' information. There were no superfluous attribute. Afterward, 621 decision rules were generated and the accuracy rate of the classification was about 0.93, as shown in Table 4.

621 rules were generated from a total of 25,584 customers' information, but, it was too complicated to use

Table 3. Accuracy rate and distribution of clustering.

		Cluster 2	Cluster 3	Cluster 4	Cluster 5
Accuracy rate of clustering		**93.6%**	**89. 8%**	**86.4%**	**83%**
No. of each group	Group 1	11992(47%*)	9859(38%)	5012(20%)	3856(15%)
	Group 2	13592(53%)	8521(33%)	6072(24%)	6046((24%)
	Group 3		7204(28%)	7557(30%)	5450((21%)
	Group 4			6943(27%)	4530(18%)
	Group 5				5702(22%)

*Percentage means number of each group divide total observations.

Table 4. Classification of accuracy of rough sets.

Class	No. of objects	Lower approx.	Upper approx.	Accuracy
1	11992	10638	12463	0.8536
2	13592	13121	14946	0.8779

Quality of classification, 0.9287.

Table 5. Classification rules of rough sets.

Rule No.	No. of objects	Rules for reasons						Strength %	Expectation
		AI(C11)	CI(C2)	GP(C12)	Age(C7)	MI(C9)	G(C5)		
352	5551	1						40.84	Low CLV
351	2647	2	5					19.47	Low CLV
354	1688		5	3				12.42	Low CLV
353	1376		2	2				10.12	Low CLV
507	1164	2			1	3		8.56	Low CLV
1	3722	3	2					31.04	High CLV
115	3566		2	4				29.74	High CLV
38	1265	3	3					10.55	High CLV
56	628	4					1	5.24	High CLV
39	587	4				3		4.89	High CLV

all rules in practice; in fact, each rule contained 41 numbers averagely. The numbers in rules less than 41 were deleted after detailed reviewing. Thus, a total of 267 rules were got and each rule contained more than 41 numbers. Those rules were used for practical analysis of marketing.

In order to simplify the rules and apply them marketing practice, this study selected 10 out of hundreds of rules that contained most item counts, as shown in Table 5. In rule 352, the key determining factor was their annual income (AI) which was less than 360 thousand (k) New Taiwan Dollar (NTD) (exchange rate which took almost 41% of the customers with low value). In rule 351, it revealed that customers purchased participating whole life (LPL) and their AI was between 360 and 500k NTD, belonging to the low customer value, of about 20%. Thus, it could be summarized from these rules that AI, category of insurance (CI), gross premium (GP), age, and amount of insured were the key influencing factors of low customers' value.

For those who had high CLV customers, their high annual income was the most important factor to them. As in Rule 1, the characters for the customers with high value were 500k < annual income < 1000k NTD and purchasing participating increasing sum assured whole life (PL). 31% of the customers with high value were included in Rule 1. In Rule 115, customers purchased PL products and their gross premiums were over 16k NTD that could be deemed as high customers' value. Furthermore, for those who purchased whole life (WL) product and 500k < annual income < 1000k NTD could be categorized as high customer value. Although annual income and gross premium were both the determining factors influencing customers' value, customers having

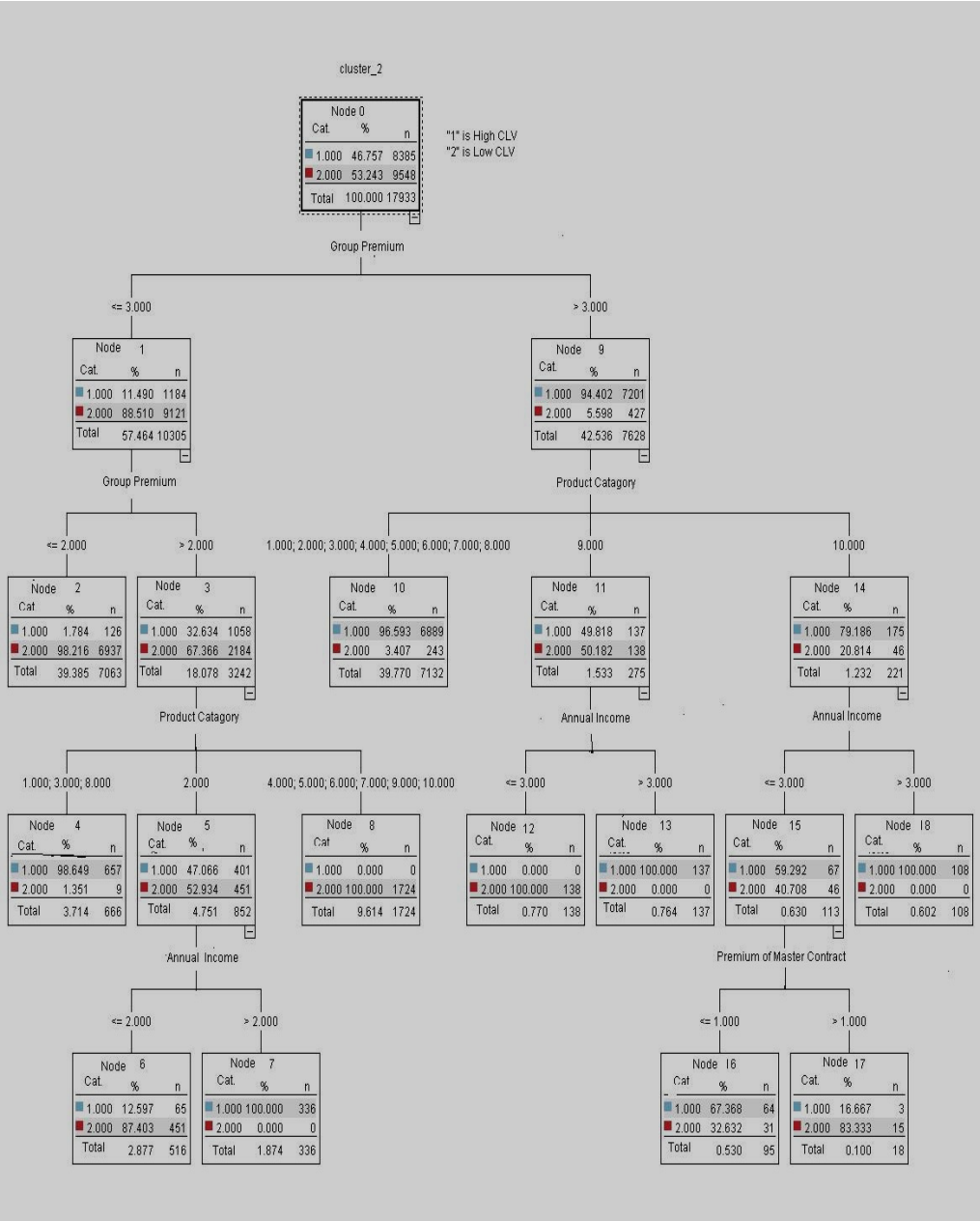

Figure 2. Insurance of customers' value by C5.0 decision tree.

lower gross premium might be classified as low customers' value; on the other hand, it tended to high customers' value if the values of two determining factors were larger.

Decision tree

Clementine12.0 and C5.0 algorithm were applied to process the DT analysis. First, K-means was used to divide CLV into two groups as the dependent variables which would incorporate nine independent variables (four demographic data and five purchase records) into consideration of analysis simultaneously.

25,584 pieces of customer data were partitioned at the ratio of 70 : 30 for DT analysis which yielded four levels and 11 nodes; meanwhile, some important and significant variables were identified as gross premium, annual income, category of insurance, and premium of master contract (PMC), shown in Figure 2.

Six rules were harvested after analysis that could describe high CLV characteristics specifically, as shown in Table 6. In DT analysis results, node 7 indicated the following facts: the customers' gross premiums were

Table 6. Rules for decision tree.

Cluster	Ordinary of rules	Characteristics description	Rule number	Accuracy rate
High CLV	1 (node 4)	If GP <= 3 and GP> 2 and CI in [1, 3, 8]	666	98.6%
	2 (node 7)	if GP <= 3 and GP> 2 and CI in [2] and AI > 2	336	100%
	3 (node 10)	If GP > 3and CI in [1, 2, 3, 4, 5, 6, 7, 8]	6889	96.6%
	4 (node 13)	if GP > 3 and CI in [9] and AI > 3	137	100%
	5 (node16)	if GP > 3 and CI in [10] and AI<= 3 and PMC <= 1	64	67.4%
	6 (node18)	if GP > 3and CI in [10] and AI > 3	108	100%
Low CLV	1 (node 2)	if GP <= 3 and GP <= 2	6937	98.2%
	2 (node 6)	if GP <= 3 and GP > 2 and CI in [2] and AI <= 2	451	87.4%
	3 (node 8)	if GP <= 3 and GP> 2 and CI in [4, 5, 6, 7, 9, 10]	1724	100%
	4 (node 12)	if GP > 3and CI in [9] and AI <= 3	138	100%
	5 (node 17)	if GP > 3and CI in [10] and AI <= 3 and PMC > 1	15	83.3%

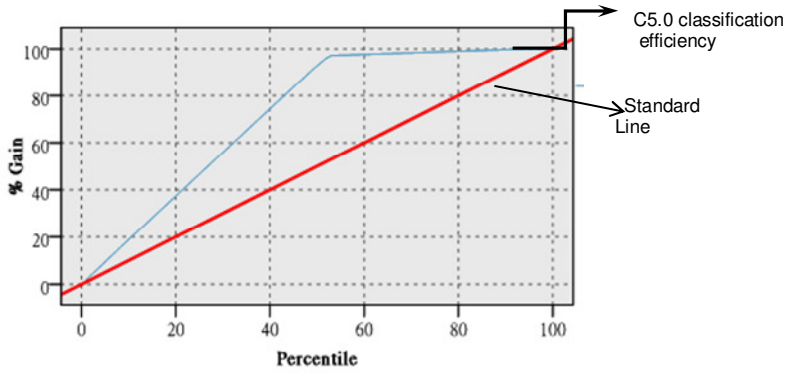

Figure 3. Gain chart in C5.0.

between 7.5 and 12k NTD, their category of insurances was PLs and amount of insurance was over 500k NTD. Meanwhile, node 17 summarized the characteristics of low CLV as gross premium over 12k NTD; choosing LESA, annual of income was under 1 million, and premium of master policy was over 2.5k NTD. All rules are shown in Table 6.

Conclusion

The influencing factors of insurance of customers' values were discussed by using RST and DT in this study. By extracting variables, both of them have concrete results in general. The primary influencing factors were gross premium, annual income, and category of insurance. The minor variables, that is, gender, amount of insured, and age of RST as well as premium of master contract in DT, were relatively small and less influential in serving as the auxiliary factors for determining marketing strategies.

In terms of precision of prediction, the RST accuracy rate was about 92.9% and the rate for nodes in DT was over 83%, except node 16. The evaluation results of integral performance by using Gant Chart were quite well (Figure 3). The data analysis performances of RST and DT were excellent, but, there are still some advantages and disadvantages:

1. RST could generate 621 rules in result output. It was flexible for users to choose the needed factors based on their demand by comparing each rule content and rate.
2. DT was a model and some DT-related commercialized software could be found. It generated final results and could be used for direct applications. But, it did not reveal the analysis course and users could not have the detail data information.
3. DT had more than one algorithm to use. Different algorithms might generate different results and users could not justify which one was the best. It depends on the data content to choose a proper algorithm.

This study provided users with application advice of using more than one analysis tools, to have perfect results; and to avoid using single tool, all to aoid having incomplete analysis. Also, using RST and DT simultaneously might have complementary effects in analyzing customers'

characteristics.

REFERENCE

Azbi R, Vanderpooten D (2002). Construction of rule-based assignment models, Eur. J. Oper. Res. 138(2):274-293.

Berger PD, Nasr NI (1998). Customer lifetime value: marketing models and applications, J. Interactive Mark. 12(1):17 -30.

Beynon MJ, Peel MJ (2001). Variable precision rough set theory and data discrimination: An application to corporate failure prediction, MEGA: Int. J. Manage. Sci. 29(6):561-576.

Blattberg RC, Deighton J (1996). Manage marketing by the customer equity test, Harv. Bus. Rev. 74(Jul-Aug):136-144.

Breiman L, Freidman J, Olshen R, Stone C. (1984). Classification and regression trees, California: Wadsworth International.

Chan CC (1998). A rough set approach to attribute generalization in data mining, J. Inf. Sci. 107(1-4):169-176.

Dimitras AL, Slowinski R, Susmaga R, Zopounidis C (1999). Business failure prediction using rough sets, Eur. J. Oper. Res. 114(2):263-280.

Dombi J, Zsiros A (2005). Learning multicriteria classification models from examples: decision rules in continuous space, Eur. J. Oper. Res. 160(3):663-675.

Donald J (1989). Determining a Customer's Lifetime Value, Direct Mark. 51(11):60-62.

Donkers B, Verhoef PC, de Jong MG (2007). Modeling CLV: A test of competing models in the insurance industry, Quant. Mark. Econ. 5(2):163-190.

Dwyer FR (1997). Customer Lifetime Valuation to Support Marketing Decision Making, J. Direct Mark. 11(4): 6 - 13.

Fader PS, Hardie BGS, and Lee K-L (2005). Counting your customers the easy way: An alternative to the Pareto/NBD model, Mark. Sci. 24(2):275-284.

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996). From Data Mining to Knowledge Discovery: An Overview In Advances in Knowledge, Discovery and Data Mining. MIT Press, Cambridge, MA.

Fisher DH, Mckusick KB (1989). An empirical comparison of ID3 and backpropagation. In Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, Detroit: pp.788-793.

Frey L, Fisher D (2003). Identifying Markov blankets with decision tree induction, In: Proceeding of the Third IEEE Int. Conf. Data Mining pp.59-66.

Gupta S, Hanssens D, Hardie B, Kahn W, Kumar V, Lin N, Ravishanker N, Sriram S. (2006). Modeling Customer Lifetime Value. J. Serv. Res. 9(2):139-155.

Gupta S, Lehmann DR (2003). Customers as assets. J. Interactive Mark. 17(1):9-24.

Han J, Kamber M, Tung AKH (2001). Spatial clustering methods in data mining: A survey, H. Miller and J. Han (eds.), Geographic Data Mining and Knowledge Discovery, London: Taylor and Francis pp.188-217.

Hautaniemi S, Kharait S, Iwabu A, Wells A, Lauffenburger DA (2005). Modeling of signal-response cascades using decision tree analysis. Bioinformatics 21(9): 2027-2035.

Hu W, Wu O, Chen Z, Fu Z, Maybank S (2007). Recognition of pornographic web pages by classifying texts and images, IEEE Trans. Pattern Anal. Mach. Intel. 29(6):1019-1034.

Hu Y-C, Chen R-S, Tzeng G-H (2003). Finding fuzzy classification rules using data mining techniques, Pattern Recogn. Lett. 24(1-3): 509-519.

Huang X, John RJ (1997). A machine-learning approach to automated knowledge- base building for remote sensing image analysis with GIS data. Eng. Remote Sensing 63(10):1185-1194.

Jackson AG, Leclair SR, Ohmer MC, Ziarko W, Al-kamhwi H (1996). Rough sets applied to materials data. Acta Material 44(11): 4475-4484.

Jackson D (1989). Determining a customer's lifetime value. Direct Mark. 51(11):60.

Kass GV (1980). An exploratory technique for investigating large quantities of categorical data. Appl. Stat. 29(2):119 -127.

Kumar V, Lemon KN, Parasuraman A (2006). Managing customers for value: an overview and Research agenda. J Serv. Res. 9(2):87-94.

Li R, Wang ZO (2004). Mining classification rules using rough sets and neural networks, Eur. J. Oper. Res. 157(2):439-448.

Lim T-S, Loh W-Y, Shih Y-S (2000). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. Machine Learn. 40(3):203-229

MacQueen J (1967). Some Methods for Classification and Analysis of Multivariate Observations, In: Proceedings of the fifth Berkeley symposium on Mathematicsm Statistics, and Probability, (eds), LeCam LM and Neyman J. Berkeley: U. California Press pp.281.

Mooney RJ, Shavlik JW, Towell G, Gove A (1989). An experimental Comparison of symbolic and connectionist learning algorithms, In Proceedings of the Eleventh International Joint Conf. on Artificial Intelligence, Detroit: pp.775-780, MI Reprinted in Readings in Machine Learning (1990).

Moshkovich HM, Mechitov AI, Olson DL (2002). Rule induction in the data mining: effect of ordinal scales. Expert Syst. Appl. 22(4):303-311.

Pawlak Z (1982). Rough sets. Int. J. Comput. Inf. Sci. 11(5):341-356.

Pawlak Z (2002). Rough sets, decision algorithms and Bayes' theorem, Eur. J. Oper. Res. 136(1):181-189.

Punj G, Stewart DW (1983). Cluster Analysis in Marketing Research: Review and Suggestions for Application. J. Mark. Res. 20(2):134-148.

Quinlan JR (1986). Induction of decision tree. Mach. Learn. 1:81-106.

Quinlan JR (1993). C4. 5: programs for machine learning, San Mateo, CA: Morgan Kaufmann.

Ranilla J, Luaces O, Bahamonde A (2003). A heuristic for learning decision trees and pruning them into classification rules. AI Commun. 16:71-87.

Ray S, Turi RH (1999). Determination of Number of Clusters in K-means Clustering and Application in Colour Image Segmentation, In: The 4th International Conf. on Advances in Pattern Recognition and Digital Techniques, Calcutta, India, 27-29, Narosa Publishing House, New Delhi, India pp.137-143.

Reinartz WJ, Kumar V (2000). On the profitability of long-life customers in a nonconttactual setting: an empirical investigation and implications for marketing. J. Mark. 64(4):17-35.

Salazar MT, Harrison T, Ansell J (2007). An approach for the identification of cross-sell and up-sell opportunities using a financial services customer database, J. Finan. Serv. Mark. 12(2):115-131.

Shiue YR, Guh RS (2006). The optimization of attribute selection in decision tree-based production control systems, Int. J. Adv. Manuf. Technol. 28(7-8):737-746.

Shyng J-Y, Tzeng G-H, and Wang F-K (2007). Rough set theory in analyzing the attributes of combination values for the insurance market, Expert Syst. Appl. 32(1): 56-64.

Shyng J-Y, Shieh H-M, Tzeng G-H, Hsieh S-H (2009). Using FSBT technique with Rough Set Theory for personal investment4 portfolio analysis, Eur. J. Oper. Res. 201(2):601-607.

Srivastava, Rajendra K, Shervani, Tasadduq A, Liam F (1998). Market-based assets and shareholder value: A framework for analysis, J. Mark. 62(1):2-18.

Tirenni G, Kaiser C, Herrmann A (2007). Applying decision trees for value-based customer relations management: Predicting airline customers' future values. J. Database Mark. Customer Strat. Manage. 14(2):130-142.

Tsumoto S (2004). Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model. Inf. Sci. 162(2): 65-80.

Verhoef PC, Donkers B (2001). Predicting customer potential value an application in the insurance industry, Decis. Support Syst. 32(2):189-199.

Weiss SM, Dapouleas I (1989). An empirical comparison of pattern recognition, neural nets, and machine learning classification methods, In Proceedings of the Eleventh International Joint Conf. on Artif. Intell. Detroit: pp.781-787.

Witlox F, Tindermans H (2004). The application of rough sets analysis in activity-based modeling Opportunities and constraints, Expert Syst. Appl. 27(2):171-180.

Wu C-Hg, Kao S-C, Su Y-Y, and Wu C-C (2005). Targeting customers via discovery knowledge for the insurance industry, Expert Syst. Appl. 29(2):291-299.

Zhai LY, Khoo LP, Fok SC (2002). Feature extraction using rough set theory and genetic algorithms an application for the simplification of product quality evaluation. Comput. Ind. Eng. 43(4): 661-676.