Information Sciences 180 (2010) 2230–2238

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A Dominance-based Rough Set Approach to customer behavior in the airline market

James J.H. Liou [a], Gwo-Hshiung Tzeng [b,c,d,*]

[a] Department of Air Transportation, Kainan University, No. 1 Kainan Road, Luchu, Taoyuan County 338, Taiwan
[b] Graduate Institute of Project Management, Kainan University, No. 1 Kainan Road, Luchu, Taoyuan County 338, Taiwan
[c] Department of Business and Entrepreneurial Management, Kainan University, No. 1, Kainan Road, Luchu, Taoyuan County 338, Taiwan
[d] Institute of Management of Technology, Chiao Tung University, Ta-Hsuch Road, Hsinchu 300, Taiwan

ARTICLE INFO

Article history:
Received 9 May 2008
Received in revised form 6 September 2009
Accepted 23 January 2010

Keywords:
Dominance
Rough set
Customer behavior
Data mining
Marketing strategy

ABSTRACT

Market segmentation is a crucial activity in the present business environment. Data mining is a useful tool for identifying customer behavior patterns in large amounts of data. This information can then be used to help with decision-making in areas such as the airline market. In this study, we use the Dominance-based Rough Set Approach (DRSA) to provide a set of rules for determining customer attitudes and loyalties, which can help managers develop strategies to acquire new customers and retain highly valued ones. A set of rules is derived from a large sample of international airline customers, and its predictive ability is evaluated. The results, as compared with those of multiple discriminate analyses, are very encouraging. They prove the usefulness of the proposed method in predicting the behavior of airline customers. This study demonstrates that the DRSA model helps to identify customers, determine their characteristics, and facilitate the development of a marketing strategy.

1. Introduction

In the face of a highly competitive and fast-changing airline market, managers must not only provide high-quality service but also react appropriately to changes in customer needs. However, it would be helpful if, instead of targeting all customers equally or offering the same incentives to all customers, enterprises could target only those customers who meet certain profitability criteria based on their individual needs or purchasing behaviors [5]. Customer behavior is the result of complex interactions between a number of factors, which can include the level of marketing activity, the competitiveness of the environment, brand perception, the influence of new technologies, and individual needs [34]. The characteristics and behaviors of airline customers are even more complex, and customer perception and behavior are affected by many factors, such as safety, service, technology, environment, price and many others. Hence, it is crucial that management determine the most important factors that affect the attitude and loyalty of airline customers. In the past, researchers have generally made use of statistical surveys to determine customer behavior. In such surveys, natural language or linguistic variables (e.g., "although the airline service is satisfactory, the price of the product being offered is high, the individual's decision is not to purchase") are used to describe customer patterns. Unfortunately, this can create an environment of imprecision, uncertainty, and partiality with regard to knowledge. These linguistic variables are then transformed into quantitative values, after which factor, cluster, and discriminant analyses are conducted. However, the semantic imprecision of natural languages leads to problems

[*] Corresponding author. Address: Department of Business Administration, Kainan University, No. 1, Kainan Road, Luchu, Taoyuan County 338, Taiwan. Tel.: +886 3 3412430; fax: +886 3 3412456.
E-mail addresses: ghtzeng@cc.nctu.edu.tw, ghtzeng@mail.knu.edu.tw (G.-H. Tzeng).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.01.025

of computation, especially when the information described in a natural language is beyond the reach of existing bivalent logic and probability theory techniques [41].

Recently, data mining techniques have been adopted to predict customer behavior [6,30]. Data mining is one stage in Knowledge Discovery in Databases (KDD), involving the application of specific algorithms for pattern extraction [21]. Marketing managers can develop strategies to attract new customers and retain highly valued ones based on this mined knowledge. The Dominance-based Rough Set Approach (DRSA), originally developed by Greco et al. [8,9], is a relatively new approach in data mining that is very useful for data reduction in qualitative analysis. The rough set theory, a kind of natural language computation, is particularly useful for dealing with imprecise or vague concepts [23]. Basically, natural language computation is a system in which the objects of computation are simply predicates and propositions drawn from a natural language. A set of decision rules is generated by applying the rough set approach to analyze the classification data. These decision rules are in the form of logic statements of the type "*if* conditions, *then* decision". The set of decision rules represents a preference model for the decision-maker that is expressed in a natural and understandable language. According to Zhu et al. [40], the rough set method does not require additional information about the data; it can work with imprecise values or uncertain data, is capable of discovering important facts hidden in that data, and has the capacity to express them in natural language. The rough set theory has been successfully applied in a variety of fields, including medical diagnosis, engineering reliability, expert systems, empirical studies of material data [15], evaluation of bankruptcy risk [29], machine diagnosis [39], business failure prediction [1,4], network intrusion detection [40], travel demand analysis [7], mining stock price [35], the insurance market [28], and accident prevention [36].

Although the Classical Rough Set Approach (CRSA) is a powerful tool for handling many problems, it is not able to deal with inconsistencies originating from the criteria, e.g., attributes with preference-ordered domains (scale) like product quality, market share, and debt ratio [10]. However, the DRSA has an advantage over the CRSA in that it has access to an information table that displays comprehensive dominance relations. It is able to deal with inconsistencies where decisive classes are not consistent with their criteria. The aim of this study is to mine data regarding airline customer behavior using the DRSA. The derived knowledge can help airlines identify valuable customers, predict future behavior, and enable firms to make proactive, knowledge-driven decisions.

2. Customer behavior

In most research on customer behavior, customer demographic variables are applied to analyze customer behavior [30]. However, Rayport and Sviokla [27] suggest that a customer's perception of product or service value is comprised of three basic elements: the product or service that a company offers, the context in which a company offers this product or service, and the infrastructure that enables the transaction to take place. In the traditional marketplace, content, context, and infrastructure are bundled together, usually offered by the same company; however, due to new technologies and new selling patterns, it is easy to separate these three elements in the air transport market. The importance of valuable customer behavioral variables—recency, frequency, and monetary (RFM)—has been extensively studied [20,31,32]. Researchers have observed that RFM variables are not only useful for the analysis of customer behavior but can also be effectively used to investigate customer value and niche markets. Hsieh [14] proposed a method that integrated data mining and behavioral scoring models for the management of banking customers. He divided customers into three groups according to their shared behaviors, characteristics and profitability. Marketers infer the profiles of each group of customers and propose management strategies appropriate to the characteristics of each group. Chen et al. [3] integrated customer behavioral variables, demographic variables, and transaction databases to establish a method of mining changes in customer behavior in the retail market. In their study, customer behavior patterns are first identified using association rule mining. After the association rules for customer behavior are discovered, changes in customer behavior are identified by comparing two sets of association rules generated from two datasets from different periods. The changes in patterns thus identified can then be investigated and assessed to provide a basis for formulating marketing strategies. Customer behavior analysis in Internet marketing has already been investigated by many researchers [16,18,19]. In most of such studies, data mining technologies are applied to produce a generalized customer profile of the Internet shopper and to further explore the Web usage pattern of the online consumer. The knowledge obtained through data mining helps foster informed Internet marketing decision-making and allows for the refinement of Web content and infrastructure to improve Internet marketing [18]. Wang and Hong [34], through the use of data mining techniques, developed a Customer Profitability Management (CPM) system for achieving marketing goals by leading customers to progress along pre-determined and desirable paths. Their system emphasizes continuous interplay between active and reactive monitoring procedures, from which any shift in customer behavior can be identified.

Just as in the conventional marketplace, airlines need to build customer loyalty as well as attract new customers. However, the unique characteristics of air transportation have altered the rules in the airline market. In a traditional business environment, where the seller meets a buyer in person, he can understand his behavior (intention to purchase, choice of product, etc.) from facial expressions, body language and verbal communication. The salesperson accumulates this knowledge while dealing face to face with the customer, and then uses this knowledge to increase the customer's satisfaction [16]. However, the average airline customer books his/her ticket via the Internet or through a travel agency. His/her decision may be based on previous experience with that airline's service, word-of mouth, the airline's safety record, convenience, and so on. Also, the products offered by an airline are not physical objects; rather they are performance and reliability. Therefore, the behavior of airline customers is different from that of customers in a traditional market or even Internet shoppers. Therefore, in this study we hope to fill in this gap: to extend the knowledge of customer behavior through data mining. The information thus obtained can be used for better decision-making in the airline market.

3. Basic concepts of the Dominance Rough Set Approach

The rough set theory, firstly introduced by Pawlak [22] in 1982, is a valuable mathematical tool for dealing with vagueness and uncertainty [23]. For a long time, the use of the rough set approach and other data mining techniques was restricted to classification problems where the preference order of the evaluations was not considered. This is due to the fact that this method cannot handle inconsistencies that occur as a result of the violation of the dominance principle [10]. In order to deal with this kind of inconsistency, it was necessary to make a number of methodological changes to the original rough set theory. Greco et al. [8] proposed an extension of the rough set theory based on the dominance principle that would permit it to deal with inconsistency. This method is mainly based on the substitution of the indiscernibility relation for a dominance relation in the rough approximation of decision classes. It is more general than the classic functional or relational model and is more understandable for users because of its natural syntax [10]. The basic concepts of DRSA are described in the following [2,8–10,12,13,17,22–26,33,37,38].

3.1. Data table

For algorithmic reasons, the information regarding the objects is supplied in the form of a data table whose separate rows refer to distinct objects (actions) and whose columns refer to the different attributes or criteria (attributes with preference-ordered domains) considered. Each cell in this table indicates an evaluation (quantitative or qualitative) of the object placed in that row by means of the attribute/criterion in the corresponding column.

Formally, a data table is in the form of a 4-tuple information system $IS = (U,Q,V,f)$, where U is a finite set of objects (universe), $Q = \{q_1, q_2,\dots,q_m\}$ is a finite set of attributes/criteria, V_q is the domain of the attribute/criterion q, $V = \bigcup_{q\in Q}V_q$ and f: $U \times Q \to V$ is a total function such that $f(x,q) \in V_q$ for each $q \in Q$; $x \in U$, called the information function. The set Q is usually divided into set C of condition attributes and set D of decision attributes.

3.2. Rough approximation by means of the dominance relations

Let $\succeq_q$ be an outranking relation to U with reference to criterion $q \in Q$, such that $x \succeq_q y$ means that "x is at least as good as y with respect to criterion q". Suppose that $\succeq_q$ is a complete preorder, i.e., a strongly complete (such that for each $x, y \in U$, at least one of $x \succeq_q y$ and $y \succeq_q x$ is verified and thus x and y are always comparable with respect to criterion q) and transitive binary relation. Moreover, let $\boldsymbol{Cl} = \{Cl_t, t \in T\}$, $T = \{1,\dots,n\}$ be a set of classes of U such that each $x \in U$ belongs to one and only one class $Cl_t \in \boldsymbol{Cl}$. We assume that all $r, s \in T$, such that $r > s$, and each element of Cl_r is preferred to each element Cl_s. In other words, if $\succeq$ is a comprehensive outranking relation on U, then it is supposed that

$$[x \in Cl_r, y \in Cl_s, r > s] \Rightarrow x \succ y,$$

where $x \succ y$ means $x \succeq y$ and not $y \succeq x$.

We can define unions of classes relative to a particular dominated or dominating class; these unions of classes are called upward and downward unions of classes and are defined, respectively, as

$$Cl_t^{\geqslant} = \bigcup_{s\geqslant t} Cl_s, \quad Cl_t^{\leqslant} = \bigcup_{s\leqslant t} Cl_s.$$

It is said that object x P-dominates object y with respect to $P \subseteq C$ (denotation xD_py); if $x \succeq_q y$ for all $q \in P$, and $D_P = \cap_{q\in p}\succeq_q$, then the dominance relation D_P is a partial preorder. Given $P \subseteq C$ and $x \in U$, let

$$D_P^+(x) = \{y \in U : yD_Px\},$$
$$D_P^-(x) = \{y \in U : xD_Py\},$$

represent the P-dominating set and the P-dominated set with respect to x, respectively.

The P-lower and P-upper approximations of $Cl_t^{\geqslant}, t \in \{1,\dots,n\}$, with respect to $P \subseteq C$ (denotation $\underline{P}(Cl_t^{\geqslant})$ and $\overline{P}(Cl_t^{\geqslant})$, respectively), are defined as

$$\underline{P}(Cl_t^{\geqslant}) = \{x \in U : D_P^+(x) \subseteq Cl_t^{\geqslant}\},$$
$$\overline{P}(Cl_t^{\geqslant}) = \bigcup_{x\in Cl_t^{\geqslant}} D_P^+(x) = \{x \in U : D_P^-(x) \cap Cl_t^{\geqslant} \neq \varnothing\}.$$

The P-lower approximation of an upward union $Cl_t^{\geqslant}, \underline{P}(Cl_t^{\geqslant})$, is composed of all objects x from the universe such that all objects y, having at least the same evaluations on all of the considered criteria from P, also belong to class Cl_t or better. Thus, one can say that if an object y has at least as good an evaluation on the criteria from P as object x belonging to $\underline{P}(Cl_t^{\geqslant})$, then certainly, y belongs to class Cl_t or better. The P-upper approximation of an upward union $Cl_t^{\geqslant}, \overline{P}(Cl_t^{\geqslant})$ is composed of all objects x

from the universe whose evaluations on the criteria from P are not worse than the evaluation of at least one other object y belonging to Cl_t or better. Analogously, the P-lower and P-upper approximations of $Cl_t^{\leqslant}$, $t \in \{1,\ldots,n\}$, with respect to $P \subseteq C$ (denotation $\underline{P}(Cl_t^{\leqslant})$ and $\overline{P}(Cl_t^{\leqslant})$, respectively) are defined as

$$\underline{P}(Cl_t^{\leqslant}) = \{x \in U : D_P^-(x) \subseteq Cl_t^{\leqslant}\},$$
$$\overline{P}(Cl_t^{\leqslant}) = \bigcup_{x \in Cl_t^{\leqslant}} D_P^-(x) = \{x \in U : D_P^+(x) \cap Cl_t^{\leqslant} \neq \emptyset\}.$$

The P-boundaries (P-doubtable region) of $Cl_t^{\geqslant}$ and $Cl_t^{\leqslant}$ are defined as

$$Bn_P(Cl_t^{\geqslant}) = \overline{P}(Cl_t^{\geqslant}) - \underline{P}(Cl_t^{\geqslant}),$$
$$Bn_P(Cl_t^{\leqslant}) = \overline{P}(Cl_t^{\leqslant}) - \underline{P}(Cl_t^{\leqslant}).$$

We define the accuracy of approximation of $Cl_t^{\geqslant}$ and $Cl_t^{\leqslant}$ for all $t \in \{1,\ldots,n\}$ and for any $P \subseteq C$, respectively, as

$$\alpha_p(Cl_t^{\geqslant}) = \frac{|\underline{P}(Cl_t^{\geqslant})|}{|\overline{P}(Cl_t^{\geqslant})|}, \quad \alpha_p(Cl_t^{\leqslant}) = \frac{|\underline{P}(Cl_t^{\leqslant})|}{|\overline{P}(Cl_t^{\leqslant})|}.$$

The ratio

$$\gamma_p(Cl) = \frac{|U - \left(\bigcup_{t \in \{2,\ldots,n\}} Bn_p(Cl_t^{\geqslant})\right)|}{|U|} = \frac{|U - \left(\bigcup_{t \in \{1,\ldots,n-1\}} Bn_p(Cl_t^{\leqslant})\right)|}{|U|}$$

defines the quality of approximation of the classification $\boldsymbol{Cl}$ by means of the criteria from set $P \subseteq C$, or, briefly, the quality of classification where $|\cdot|$ means the cardinality of a set. This ratio expresses the proportion of all P-correctly classified objects—i.e., all of the non-ambiguous objects to all of the objects in the data table. Every minimal subset $P \subseteq C$ such that $\gamma_P(\boldsymbol{Cl}) = \gamma_C(\boldsymbol{Cl})$ is called a *reduct* of C with respect to $\boldsymbol{Cl}$ and is denoted by $RED_{Cl}(P)$. Again, a data table may have more than one *reduct*. The intersection of all of the *reducts* is known as the *core*, denoted by $CORE_{Cl}$.

3.3. Decision rules

The end result of the DRSA is a representation of the information contained in the considered data table in terms of simple "*if*…, *then*…" decision rules. For a given upward union of classes $Cl_t^{\geqslant}$, the decision rules that are induced under the hypothesis that actions belonging to $\underline{P}(Cl_t^{\geqslant})$ are positive (while the others are negative) suggest an assignment to "at least class Cl_t". Analogously, for a given downward union $Cl_s^{\leqslant}$, the rules induced under a hypothesis that actions belonging to $\underline{P}(Cl_s^{\leqslant})$ are positive and that all others are negative suggest an assignment to "at most class Cl_s". On the other hand, the decision rules induced under a hypothesis that actions belonging to the intersection $\overline{P}(Cl_s^{\leqslant}) \cap \overline{P}(Cl_t^{\geqslant})$ are positive, and that all of the others are negative, suggest an assignment to some class between Cl_s and $Cl_t(s < t)$.

The following three types of decision rules can be considered:

1. $D_{\geqslant}$: decision rules that have the following form:

$$\text{if } f\ (x, q_1) \geqslant r_{q1} \text{ and } f(x, q_2) \geqslant r_{q2} \text{ and } \ldots f(x, q_p) \geqslant\ r_{qp}, \text{ then } x\ \in Cl_t^{\geqslant}.$$

These rules are supported only by objects from P-lower approximations of the upward unions of classes $Cl_t^{\geqslant}$.

2. $D_{\leqslant}$: decision rules that have the following form:

$$\text{if } f(x, q_1) \leqslant r_{q1} \text{ and } f(x, q_2) \leqslant r_{q2} \text{ and } f(x, q_p) \leqslant r_{qp}, \text{ then } x \in Cl_t^{\leqslant}.$$

These rules are supported only by objects from the P-lower approximation of the downward unions of classes $Cl_t^{\leqslant}$.

3. $D_{\geqslant\leqslant}$: decision rules that have the following form:

$$\text{if } f(x, q_1) \geqslant r_{q1} \text{ and } f(x, q_2) \geqslant r_{q2} \text{ and } \ldots f(x, q_k) \geqslant r_{qk} \text{ and } f(x, q_{k+1}) \leqslant r_{qk+1} \text{ and } \ldots f(x, q_p) \leqslant r_{qp}, \text{ then } x \in Cl_s \bigcup Cl_{s+1} \bigcup \cdots \bigcup Cl_t.$$

These rules are supported only by objects from the P-boundaries of the unions of classes $Cl_s^{\leqslant}$ and $Cl_t^{\geqslant}$, where $P = \{q_1, q_2, \ldots, q_p\} \subseteq C$, $(r_{q1}, r_{q2}, \ldots, r_{qp}) \in V_{q1} \times V_{q2} \times \cdots \times V_{qp}$ and $t \in \{1,\ldots,n\}$.

The algorithm for induction regarding decision rules is obtained from Greco et al. [11]. It is also noted that the set of decision rules induced from the approximations defined using dominance relations gives a more synthetic representation of knowledge as contained in the decision table than the set of rules induced from the CRSA, which uses indiscernible relations. This is due to the more general syntax of the rules ("$\geqslant$" and "$\leqslant$" are used instead of "=").

Table 1
Specifications of attributes/criteria related to customer behavior.

Attribute/criterion name	Attribute/criterion value	Value set	Preference
Condition attributes			
Gender	Female; male	{F, M}	None
Age	Below 30; 30–40; 41 and above	{1, 2, 3}	None
Occupation	Government employee; employee of company; student; others	{G, P, S, O}	None
Education	High school and below; college; masters and above	{H, C, M}	None
Income	NT$40,000 and below; 40,001–90,000; 90,000 and above	{1, 2, 3}	None
Service quality	Poor; medium; satisfactory	{1, 2, 3}	Gain
Schedule	Bad; good	{1, 2}	Gain
Safety record	Poor; medium; satisfactory	{1, 2, 3}	Gain
Price	Low; medium; high	{1, 2, 3}	Cost
Decision criterion			
Buying behavior	Will not consider purchasing; maybe; will surely purchase	{1, 2, 3}	Gain

4. Empirical study: a case of customer behaviors in the airline market

In order to demonstrate the effectiveness of the DRSA and our proposed approach, we carried out an empirical study that is described in this section. We produced a questionnaire about customer behavior in the airline market with single and multiple-choice answers and used DRSA to explore the classification problem. The results should provide airlines with useful information to help them develop marketing strategies and achieve their marketing goals.

4.1. Preparation of the information table for customer behavior in the airline market

Since the products of airlines are their performance and reliability rather than physical objects, their customer behavior is different from that in the conventional marketplace. Therefore, the first thing to do was to define the attributes/criteria for the information table. After consulting with senior management from travel agencies and airline marketing personnel, we listed 10 criteria to be covered by the first questionnaire. Ninety-two passengers at the Taoyuan International Airport in Taiwan agreed to rate the importance of the decision criteria that affected their purchasing behavior using a five-point Likert-type scale that ranged from 5 (extremely important) to 1 (no effect). Afterwards, the four highest-scoring criteria were extracted and used to construct information systems for airline market mining. Since demographic variables are still important references for airline customer behavior, they were considered as other attributes to be evaluated. Based on the above evaluation process, the results of the first questionnaire indicated that safety, price, quality of service and schedule are the most important criteria that influence the buying behavior of airline customers. Therefore, the second questionnaire contained two parts: (1) demographic information about the participants and (2) criteria for the surveyed airlines. The questionnaire covered eight international airlines serving Taiwan. The attributes/criteria for customer behavior in the airline market are shown in Table 1. The preference regarding schedule, quality of service, safety and buying decisions were set as gains, while the price was set as cost (the lower the better). The personal profile attributes were all set to "no preference". Each respondent was asked to rate his/her level of satisfaction for each criterion and his/her buying decision for the specified airline. 451 respondents completed the questionnaire. The demographics of the 451 respondents are shown in Table 2.

4.2. Results of the DRSA analysis

The results of the DRSA analysis consisted of four parts: quality of approximation, rule generation, rule validation and the significance of condition attributes/criteria.

4.2.1. Quality of approximation

The accuracy of approximation for the three decision classes is shown in Table 3. The results indicate good accuracy for the different classes. In general, high values for the quality of classification and accuracy mean that the attributes/criteria selected are an approximation of the classification. The "at most 1" class is the "do not consider buying" class. There are 27 objects belonging to that class. The accuracy of approximation for "at most 1" is one. The "at most 2" class includes the "do not consider buying" and "maybe" classes, for which accuracy reaches 0.957. On the other hand, the "at least 3" class refers to the "will surely purchase" class, and its lower and upper approximations are 173 and 185, respectively. The accuracy of the "at least 3" class is 0.935. The "at least 2" class is composed of the "maybe" and "will surely purchase" classes, and its accuracy is raised to one. The overall quality of approximation is calculated as follows: $(451 - 12)/(451) = 0.973$.

4.2.2. Rule generation

We establish a set of rules, the "minimum cover rules" (i.e., where the set does not contain any redundant rules), and these rules are certain, such that there are a total of 25 rules generated from the data. Table 4 shows the minimum cover rules obtained after eliminating those with a "cover strength" (i.e., number of objects, in this case passengers, covered by

Table 2
Background of the respondents.

Distribution	Sample number	Frequency (%)
Gender		
Female	244	54.1
Male	207	45.9
Age		
Less than 30	197	43.7
30–40	156	34.6
40 and above	98	21.7
Occupation		
Government employee	66	14.6
Employee of company	279	61.9
Student	52	11.5
Others	54	12.0
Education		
High school and below	65	14.5
College	332	73.6
Masters and above	54	11.9
Income (NT)		
Less than 40,000	225	49.9
40,000–90,000	204	45.2
90,000 and above	22	4.9

Table 3
Accuracy of classification for customer behavior.

	At most 1	At most 2	At least 2	At least 3
Lower approx.	27	266	424	173
Upper approx.	27	278	424	185
Boundary	0	12	0	12
Accuracy	1	0.957	1	0.935

Table 4
Minimum cover rules with a strength greater than 10.

No.	Conditions	Decision	Strength
1	(Safety $\leqslant$ 1) & (Price $\geqslant$ 2)	$D \leqslant 1$	27
2	(Price $\geqslant$ 3)	$D \leqslant 1$	126
3	(Safety $\leqslant$ 2) & (Occupation = O)	$D \leqslant 2$	32
4	(Safety $\leqslant$ 2) & (Occupation = G)	$D \leqslant 2$	27
5	(Safety $\leqslant$ 2) & (Age = 2) & (Service $\leqslant$ 2)	$D \leqslant 2$	76
6	(Safety $\leqslant$ 1)	$D \leqslant 2$	38
7	(Safety $\leqslant$ 2) & (Education = H) & (Age = 3)	$D \leqslant 2$	18
8	(Occupation = S) & (Safety $\leqslant$ $\leqslant$ 2) & (Service $\leqslant$ 2)	$D \leqslant 2$	19
9	(Age = 3) & (Education = C) & (Service $\leqslant$ 2)	$D \leqslant 2$	23
10	(Safety $\leqslant$ 2) & (Price $\geqslant$ 2)	$D \leqslant 2$	131
11	(Age = 2) & (Safety $\leqslant$ 2) & (Sex = F)	$D \leqslant 2$	53
12	(Safety $\geqslant$ 3) & (Price $\leqslant$ 2)	$D \geqslant 3$	164
13	(Education = M) & (Price $\leqslant$ 1) & (Service $\geqslant$ 3)	$D \geqslant 3$	13
14	(Price $\leqslant$ 2) & (Service $\geqslant$ 3) & (Age = 3)	$D \geqslant 3$	31
15	(Price $\leqslant$ 1) & (Service $\geqslant$ 3) & (Occupation = P) & (Age = 1)	$D \geqslant 3$	16
16	(Age = 2) & (Service $\geqslant$ 3) & (Income = 2) & (Price $\leqslant$ 1)	$D \geqslant 3$	25
17	(Safety $\geqslant$ 2)	$D \geqslant 2$	413
18	(Price $\leqslant$ 1)	$D \geqslant 2$	202

the rule) of less than 10. The reduced rule set contains 18 rules, with two rules corresponding to class 1, nine rules to at most class 2, five rules to class 3, and two rules to at least class 2. We can see from Table 4 that if the price is high, the decision is "will not consider buying" and its cover strength is 126 (rule 2). This means that nearly one-fourth of the respondents will not consider purchasing from a specific airline if its price is high. Rule 5 suggests that the customer's decision will be at most class 2 (maybe) when safety is rated as medium or less, the customer's age is between 30 and 40, and the service quality is rated as medium or less. For students (rule 8), their decisions are at most class 2 when safety and service quality are equal to or less than medium. From rule 11, we can see that if the respondents are female, they are between 30 and 40, and the safety

Table 5
Hit rates for DRSA and discriminant analysis.

	Class 1	Class 2	Class 3
Class 1	25(27)	0(0)	0(0)
Class 2	0(11)	237(219)	3(16)
Class 3	0(0)	10(14)	167(164)
Ambiguous decision	9(0)		
Overall classification error	0.044(0.092)		
Correct hit rate	95.6%(91.8%)		

*Parenthesis () indicates the results of discriminant analysis.

is rated medium or less, then their decision will be at most class 2, and the cover strength is 53. Rule 12 indicates that if the airline's safety is satisfactory and the price is rated as medium or less, customers will surely purchase and the cover strength reaches 164. If the passenger's income is between NT$ 40,000–90,000, his/her age is between 30 and 40 (which is considered middle age in this society), safety is rated satisfactory and the price is low, then he/she will definitely be a loyal customer. If safety is greater than or equal to medium and the price is low, then the respondent's decision will be at least class 2, and the cover strength numbers are 413 and 202, respectively. The results indicate that a good safety record is attractive to 90% of the customers, while low price is attractive to about 50%. The set of minimum cover rules indicates that for those who are under 30 years of age, whether employed with private companies or part of student groups, price and service are more important. For higher-income, more highly educated or government-employee female respondents, safety is important. It is also worth noting that of the condition criteria, the schedule is the one that seems to have little effect. This may be because there were so few higher-income respondents in the surveyed population. The higher-income group is usually more concerned with time, but in our demographic data, there were only 22 respondents in the high-income group.

4.2.3. Rule validation

We check the feasibility of the decision rules generated in this study through the application of the 10-fold cross-validation technique. First, from a randomly chosen 90% of the data, we generate decision rules. The remaining 10% of the data are used to validate the hit rate of the generated decision rules—i.e., the percentage of correct predictions for each class. This procedure is repeated 10 times; the hit rate is shown in Table 5.

As shown in Table 5, the overall classification error is only 4.4%, with 431 objects decided correctly and only 19 objects decided incorrectly. The nine ambiguous objects indicate that the object classification occurs 10 times during the validation processes. This ambiguity may be due to the hesitation of the respondents. The effectiveness of the DRSA is shown by the results of discriminant analysis in Table 5. The hit rate has increased from 91.8% with discriminant analysis to 95.6% for DRSA.

4.2.4. Significance of the condition attributes/criteria

Since there is no *reduct* generated from the information system, all of the condition attributes/criteria are essentially utilized to distinguish the classes. However, the significance of the condition attributes/criteria can be measured by their presence in the derived rules. When a condition attribute/criterion shows up more frequently in the cover rules, and the strength of these rules is higher, this is more likely to be a key factor in customer decisions. The frequency with which the attributes/criteria and cover strength appear in the derived minimum cover rules (Table 4)—number of objects matching the rule—can be used as an indication of the importance of the attributes/criteria. Rules 17 and 18 indicate that safety and price are the most significant criteria for customer decisions, with cover strengths of 413 and 202, respectively. The implication is that most of the respondents make purchase decisions mainly based on the airline's safety record and ticket price.

Another approach to exploring the importance of attributes/criteria is to set the strength level higher than a certain threshold value. Each rule in the derived rule set is thus ensured to be above a certain percentage in terms of cover strength. If we choose to generate all possible rules where the relative strength of each class is greater than 40%, then the total number

Table 6
Possible rules for strengths greater than 40% for each class.

No.	Condition	Decision	Strength
1	(Service $\leqslant$ 1) & (Price $\geqslant$ 2)	$D \leqslant 1$	10
2	(Safety $\leqslant$ 1) & (Price $\geqslant$ 2)	$D \leqslant 1$	27
3	(Price $\geqslant$ 3)	$D \leqslant 2$	126
4	(Safety $\leqslant$ 2) & (Price $\geqslant$ 2)	$D \leqslant 2$	131
5	(Education = C) & (Service $\leqslant$ 2) & (Safety $\leqslant$ 2)	$D \leqslant 2$	152
6	(Safety $\geqslant$ 2)	$D \geqslant 2$	413
7	(Price $\leqslant$ 1)	$D \geqslant 2$	202
8	(Service $\geqslant$ 3) & (Price $\leqslant$ 2)	$D \geqslant 2$	172
9	(Safety $\geqslant$ 3) & (Price $\leqslant$ 2)	$D \geqslant 3$	164

of rules will be reduced to nine (Table 6). In the resulting set of rules, safety and price are still the most crucial condition criteria. It is clear that the higher the satisfaction level in terms of safety, the higher the desire to purchase will be, and that a low price is still a good way to attract people's attention.

5. Discussion

Using the same data set and dummy-variable technique, we also apply discriminant analysis. The results are shown in Table 5. Clearly the DRSA shows better prediction ability than does discriminant analysis. The hit rate increases from 91.8% for discriminant analysis to 95.6% for DRSA. Furthermore, in contrast to discriminant analysis, the rough set theory requires no underlying statistical assumptions. In particular, the DSRA can handle both attributes/criteria with and without preference order. In this study, we use this advantage to analyze survey data (whether with quantitative or qualitative attributes, and with or without preference order) and to better understand customer behavior in the airline market. Although rough set theory has been successfully applied to a variety of areas, it is still not often applied in the customer behavior field. In this study, we also consider inconsistencies within criteria. The empirical results show that it is appropriate to apply DSRA to the mining of knowledge regarding customer behavior for the air transport market. Although the selected attributes/criteria may be strongly influenced by the local environment and culture, the proposed factor structure can be easily adapted and extended to the conditions and culture in other markets in any particular environment.

A condition attribute/criterion can be dropped if its removal will have no impact on the quality of the approximation. However, in our empirical study, no *reduct* or *core* is generated. This means that all of the selected attributes/criteria are important factors for the quality of the approximation. This result may be partly due to the fact that all criteria were derived from the first questionnaire, which included only important criteria. Table 3 shows that the accuracy of the classification for customer behavior is as high as 0.973, which indicates the narrow boundary between the lower approximation and the upper approximation. Thus, the granules derived from the refined data can be understood to properly represent customer behavior. The high accuracy rate also shows that the objects in that class have higher dependency among all condition attributes/criteria for those objects. The total number of minimum cover rules is 25, but there are seven rules whose strength is less than 10. This indicates that most of the data derived from this study can be classified into 18 rules with only a few unique data. In contrast to the classical rough set theory, which tends to generate too many rules with only a little cover strength, with the DSRA we seem to be able to obtain better decision rules with stronger cover strength. This could be due to the fact that the DSRA considers both attributes/criteria and their inconsistencies. When there are too many unique rules generated, it is difficult to understand the contents of the data sets that relate to the decision rules in the condition part.

In this study of customer behavior in the air transport market in Taiwan, we find that two criteria dominate customer decision-making: safety and price. Although price and quality of service are usually of major concern in the air transport market, the airline's safety record is the most critical factor affecting customer decision-making. This is not difficult to understand, as the less than desirable safety record of Asian airlines in general has left most air travelers in Taiwan worried about safety. Therefore, the key to a successful strategy is to rebuild the confidence of air transport customers. The decision rules indicate that if the airline performs well in terms of safety, the customer will consider that airline even if ticket price and quality of service are less competitive. The schedule has little influence on the customer's decision-making. In order to isolate representative rules and an effective factor structure, we conducted a 10-fold cross-validation and possible rule generation. The results from both processes show the validity of the decision rules. There is very good agreement, with only a 4.36% overall classification error. The estimation results show that the accuracy of the approximation, the quality of the approximation, and the hit rates are satisfactory. The category of condition attributes/criteria for targeted customers can be increased in the future to help the decision-maker make a more precise judgment. More categories of condition attributes/criteria generate more refined decision rules that can improve the quality of the decision-maker's strategy.

6. Conclusions

This study illustrates the usefulness of the DRSA approach as an operational tool for the prediction of customer behavior in the air transport market. The proposed prediction model takes the form of decision rules. Since the derived rules are supported by real examples, they describe only the most relevant attributes/criteria. The classical rough set theory handles attributes without preferences, a technique that does not always accurately represent the real world. The DRSA is constructed by extending the classical rough set theory to include qualitative reasoning for preference-based customer behavior analysis. This is done by replacing the indiscernibility relation with the dominance relation. In this way we can represent the conflicting preference relations that affect customer behavior more objectively without introducing the equivalence class concept of the classical rough set theory. Compared with those of the traditional statistical method, the results indicate that the DRSA has better prediction ability. Moreover, the derived decision rules are in natural language form, which makes their meaning easier to understand than with traditional methods.

Acknowledgements

This paper was supported by National Science Committee of Taiwan under Grant No. NSC 98-2410-H-424-005.

References

[1] M.J. Beynon, M.J. Peel, Variable precision rough set theory and data discretisation: an application to corporate failure prediction, Omega 29 (2001) 561–576.

[2] J. Blaszczynski, S. Greco, R. Slowinski, Multi-criteria classification – a new scheme for application of dominance-based decision rules, European Journal of Operational Research 181 (2007) 1030–1044.

[3] M.C. Chen, A.L. Chiu, H.W. Chang, Mining changes in customer behavior in retail marketing, Expert Systems with Applications 28 (2005) 773–781.

[4] A.I. Dimitras, R. Slowinski, R. Susmaga, C. Zopounidis, Business failure prediction using rough sets, European Journal of Operational Research 114 (1999) 263–280.

[5] J. Dyche, J. Dych, The CRM Handbook: A Business Guide to Customer Relationship Management, Addison-Wesley, MA, 2001.

[6] P. Giudici, G. Passerone, Data mining of association structures to model customer behavior, Computational Statistics and Data Analysis 38 (2002) 533–541.

[7] C. Goh, R. Law, Incorporating the rough sets theory into travel demand analysis, Tourism Management 24 (2003) 511–517.

[8] S. Greco, B. Matarazzo, R. Slowinski, A new rough set approach to evaluation of bankruptcy risk, in: C. Zopounidis (Ed.), Rough Fuzzy and Fuzzy Rough Sets, Kluwer, Dordrecht, 1998, pp. 121–136.

[9] S. Greco, B. Matarazzo, R. Slowinski, Extension of the rough set approach to multicriteria decision support, INFOR 38 (2000) 161–195.

[10] S. Greco, B. Matarazzo, R. Slowinski, Rough sets theory for multicriteria decision analysis, European Journal of Operational Research 129 (2001) 1–47.

[11] S. Greco, B. Matarazzo, R. Slowinski, J. Stefanowski, An algorithm for induction of decision rules consistent with dominance principle, in: W. Ziarko, Y. Yao (Eds.), Rough Sets and Current Trends in Computing, LNAI 2005, Springer-Verlag, Berlin, 2001, pp. 304–313.

[12] S. Greco, B. Matarazzo, R. Slowinski, Rough sets methodology for sorting problems in presence of multiple attributes and criteria, European Journal of Operational Research 138 (2002) 247–259.

[13] S. Greco, B. Matarazzo, R. Slowinski, Rough set analysis of preference-ordered data, in: J.J. Alpigini, J.F. Peters, A. Skowron, N. Zhong (Eds.), Rough Sets and Current Trends in Computing, Springer-Verlag, Berlin, 2002, pp. 44–59.

[14] N.C. Hsieh, An integrated data mining and behavioral scoring model for analyzing bank customers, Expert Systems with Applications 27 (2004) 623–633.

[15] A.G. Jackson, S.R. Leclair, M.C. Ohmer, W. Ziarko, H. Al-kamhwi, Rough sets applied to material data, ACTA Material 44 (1996) 4475–4484.

[16] M. Jenamani, P. Mohapatra, S. Ghose, A stochastic model of e-customer behavior, Electronic Commerce Research and Applications 2 (2003) 81–94.

[17] W. Kotlowski, K. Dembczynski, S. Greco, R. Slowinski, Stochastic dominance-based rough set model for ordinal classification, Information Sciences 178 (2008) 4019–4037.

[18] I. Kwan, J. Fong, H.K. Wong, An e-customer behavior model with online analytical mining for internet marketing planning, Decision Support Systems 41 (2005) 189–204.

[19] H. Lu, C.C. Lin, Predicting customer behavior in the market-space: a study of Rayport and Sviokla's framework, Information and Management 40 (2002) 1–10.

[20] J.R. Miglautsch, Thoughts on RFM scoring, International Society for Strategic Marketing, Journal of Database Marketing 8 (2000) 67–72.

[21] S. Mitra, P. Mitra, S.K. Pal, Evolutionary modular design of rough knowledge-based network using fuzzy attributes, Neurocomputing 36 (2001) 45–66.

[22] Z. Pawlak, Rough sets, International Journal of Computer and Information Science 11 (1982) 341–356.

[23] Z. Pawlak, Rough sets, in: T.Y. Lin, N. Cercone (Eds.), Rough Sets and Data Mining: Analysis for Imprecise Data, Kluwer Academic Publishers, Norwell, MA, 1997.

[24] Z. Pawlak, A. Skowron, Rudiments of rough sets, Information Sciences 177 (2007) 3–27.

[25] Z. Pawlak, A. Skowron, Rough sets and Boolean reasoning, Information Sciences 177 (2007) 41–73.

[26] Z. Pawlak, A. Skowron, Rough sets: some extensions, Information Sciences 177 (2007) 28–40.

[27] J.F. Rayport, J.J. Sviokla, Managing in the market-space, Harvard Business Review 72 (1994) 141–150.

[28] J.Y. Shyng, F.K. Wang, G.H. Tzeng, K.S. Wu, Rough set theory in analyzing the attributes of combination values for the insurance market, Expert Systems with Applications 32 (2007) 56–64.

[29] R. Slowinski, C. Zopounidis, Application of the rough set approach to evaluation of bankruptcy risk, International Journal of Intelligent Systems in Accounting, Finance and Management 4 (1995) 27–41.

[30] H.S. Song, J.K. Kim, S.H. Kim, Mining the change of customer behavior in an internet shopping mall, Expert Systems with Applications 21 (2001) 157–168.

[31] H. Suh, K.C. Noh, C.K. Suh, Customer list segmentation using the combined response model, Expert Systems with Applications 17 (1999) 89–97.

[32] C.Y. Tsai, C.C. Chiu, A purchase-based market segmentation methodology, Expert Systems with Applications 27 (2004) 265–276.

[33] T.L. (Bill) Tsenga, C.C. Huang, Rough set-based approach to feature selection in customer relationship management, Omega 35 (2007) 365–383.

[34] H.F. Wang, W.K. Hong, Managing customer profitability in a competitive market by continuous data mining, Industrial Marketing Management 35 (2006) 715–723.

[35] Y.F. Wang, Mining stock price using fuzzy rough set system, Expert Systems with Applications 24 (2003) 13–23.

[36] J.C. Wong, Y.S. Chung, Rough set approach for accident chains exploration, Accident Analysis and Prevention 39 (2007) 629–637.

[37] Y. Yao, Y. Zhao, Attribute reduction in decision-theoretic rough set models, Information Science 178 (2008) 3356–3373.

[38] Y. Yao, Y. Zhao, Discernibility matrix simplification for constructing attribute reducts, Information Sciences 179 (2009) 867–882.

[39] L.Y. Zhai, L.P. Khoo, S.C. Fok, Feature extraction using rough set theory and genetic algorithms an application for the simplification of product quality evaluation, Computers and Industrial Engineering 43 (2002) 661–676.

[40] D. Zhu, G. Premkumar, X. Zhang, C.H. Chu, Data mining for network intrusion detection: a comparison of alternative methods, Decision Science Journal 32 (2001) 635–660.

[41] L.A. Zadeh, A new frontier in computation—computation with information described in natural language, keynote speaker's lecture notes, IEEE Conference on Systems, Man, and Cybernetics, Taipei, Taiwan, 2006.