Expert Systems

Article

DOI: 10.1111/j.1468-0394.2010.00570.x

Pattern filtering and classification for market basket analysis with profit-based measures

Mu-Chen Chen,[1] Chuang-Min Chao[2] and Kuan-Ting Wu[3]

(1) Institute of Traffic and Transportation, National Chiao Tung University, 4F, No. 118, Section 1, Chung Hsiao W. Road, Taipei 100, Taiwan
Email: ittchen@mail.nctu.edu.tw
(2) Department of Business Management, National Taipei University of Technology, Taipei, Taiwan
(3) Institute of Commerce Automation and Management, National Taipei University of Technology, Taipei, Taiwan

Abstract: *Market basket analysis is one of the typical applications in mining association rules. The valuable information discovered from data mining can be used to support decision making. Generally,* support *and* confidence *(objective) measures are used to evaluate the interestingness of association rules. However, in some cases, by using these two measures, the discovered rules may be not profitable and not actionable (not interesting) to enterprises. Therefore, how to discover the patterns by considering both objective measures (e.g. probability) and subjective measures (e.g. profit) is a challenge in data mining, particularly in marketing applications. This paper focuses on pattern evaluation in the process of knowledge discovery by using the concept of profit mining. Data Envelopment Analysis is utilized to calculate the efficiency of discovered association rules with multiple objective and subjective measures. After evaluating the efficiency of association rules, they are categorized into two classes, relatively efficient (interesting) and relatively inefficient (uninteresting). To classify these two classes, Decision Tree (DT)-based classifier is built by using the attributes of association rules. The DT classifier can be used to find out the characteristics of interesting association rules, and to classify the unknown (new) association rules.*

Keywords: data mining, profit mining, association rules, data envelopment analysis, decision tree

1. Introduction

With the continuous growth of information technology, massive amounts of data are collected and stored by enterprises. The adaptive control for the business applications at a global optimization level can be enabled by linking supply chain and intelligent transportation to the enterprise information systems (Hsu & Wallace, 2007). It is very important for enterprises to transform the data into useful information and knowledge for decision making in dynamic markets. Knowledge management is one of the most essential factors for business success in extremely dynamic environments (Xu *et al.*, 2006). In order to overcome the shortcoming of the traditional data analysis tools and techniques that have difficulty dealing with the massive size of databases, data mining techniques recently have been developed. In simple terms, data mining is the task of extracting the interesting patterns or rules from large amounts

of data (Han & Kamber, 2006); in other terms, this is the process of mining the meaningful information in a large database (Tan *et al*., 2006). Market basket analysis (association rule mining) is a representative case in data mining (Han & Kamber, 2006). Market basket analysis can explore for relationships between items from customers' transactions and can be used to help retailers to understand customers' purchasing behaviours. The information about items which are frequently purchased together by customers also can assist decision making, such as marketing, retail store layout, customer segmentation and cross-selling (Srikant *et al*., 1997).

The selection of association rules is based on whether the rule is interesting and useful for users. The measures most frequently used in association rule extraction are *support* and *confidence* (Olafsson *et al*., 2008; Lenca *et al*., 2008). Users can set the *minimum support* (*min sup*) and *minimum confidence* (*min conf*) to filter the discovered rules. However, by only using these two measures, it may generate a large set of rules, most of which may be not valuable (not interesting) (Olafsson *et al*., 2008), or it may not be able to derive any valuable rules. For instance, most customers hardly buy high-priced products, so these products may not satisfy *min sup* and *min conf*, but they are valuable to enterprises (Chen, 2007). As the well-known *Ketel Vodka and Beluga Caviar* problem (Cohen *et al*., 2001) indicates, most customers rarely purchase either of these two products, and they do not easily become frequent itemsets, but their profit may be probably higher than that of *beer and diapers*. We address another example in which the infrequent itemsets are interesting. Tao *et al*. (2003) has pointed out that the association rule of [*wine* → *salmon*, *support* = 1%, *confidence* = 80%] may be more valuable than [*bread* → *milk*, *support* = 3%, *confidence* = 80%] even though the former comes with a lower *support*. The reason is that the items in the first rule could make more profit per unit sale.

There is a significant gap between the statistics-based pattern extraction and the value-based decision making in using data mining (Wang *et al*., 2002). Therefore, how to mine the patterns or rules which consider both objective measures (e.g. probability) and subjective measures (e.g. profit) is a challenge for data mining, and it is an important issue for enterprises.

From the above examples, *Ketel Vodka and Beluga Caviar* and *wine and salmon*, the infrequent itemsets are potentially interesting because they simultaneously consider the domain knowledge such as the value or profit of product, and the traditional statistics-based criteria. By only considering the statistics-based measures and ignoring the subjective domain knowledge, the infrequent products with high value to enterprises are simply viewed as uninteresting. On the other hand, we can set a lower threshold to filter the profitable infrequent itemsets, but a lot of association rules will be generated which results in difficulty choosing the useful rules for decision making. To overcome the above dilemma in traditional association rule mining algorithms (Chen, 2007), the subjective measures can be further used to filter the discovered association rules. Therefore, in this paper we apply the property of Data Envelopment Analysis (DEA) that can measure the relative efficiency of decision-making units (DMUs) to calculate the efficiency of mined rules with multiple criteria. These criteria consist of objective statistics-based measures and subjective domain knowledge. DEA can be used to classify objects with multiple criteria. For instance, Ramanathan (2006) utilized DEA to classify inventory with four criteria including average unit cost, annual dollar usage, critical factor and lead time. Furthermore, Dulá (2008) indicated that DEA can be taken as a data mining tool to recognize geometric outliers applied beyond the common area of efficiency and productivity.

Decision tree (DT) algorithms have to provide a mechanism to express an attribute splitting condition and its associated results for various attribute types. By using some measures, DT algorithms iteratively select the best attribute to split the objects, and grow the tree until certain criteria are satisfied (Tan *et al*., 2006). By constructing the DT-based classification model, the tree can be applied to classify the unknown data by using data attributes. Because

of the flexibility of DT, it has the capability of using different feature subsets and decision rules at different stages of classification, and the capability of tradeoffs between classification accuracy and time/memory space efficiency (Safavian & Landgrebe, 1991). Perhaps, DT is the most widely used method to build classifiers, and it has been applied to resolve many real-world classification problems (Rokach & Maimon, 2005).

Combining DEA and DT and the resulting applications have been explored in previous studies (Sohn & Moon, 2004; Lee & Park, 2005; Seol *et al*., 2007; Samoilenko & Osei-Bryson, 2008). In these studies, DEA was commonly used to investigate the relative efficiency of DMUs. Instead of using DEA alone, Sohn and Moon (2004) developed a hybrid DEA-DT method to forecast the degree of commercialization efficiency by constructing a DT model, in which the environmental characteristics of new technology are taken as attributes, and the result of DEA is the target variable. Seol *et al*. (2007) used the DT technique to select the inefficient process for improving the service unit's overall efficiency. In Samoilenko and Osei-Bryson (2008), the nature of the relative efficiencies of DMUs is discovered by applying DT. Lee and Park (2005) proposed a profitable customer segmentation system by combining the three methods of DEA, DT and neural networks.

The concept of profit mining was initially coined by Wang *et al*. (2002). They pointed out that the purpose of profit mining is to construct a model to recommend suitable products and proper prices for customers in order to maximize net profit. Hence, this paper emphasizes value-based pattern evaluation in the process of knowledge discovery by using both objective and subjective measures to evaluate the interestingness and usefulness of association rules. The proposed approach can help analysts make decisions with discovered rules, which are profitable to enterprises.

As abovementioned, how to evaluate and choose the valuable or profitable rules or patterns in data mining applications is an important issue. Association rule mining is user-centric because its objective is to discover the useful (or interesting) rules from which new knowledge can be derived (Ceglar & Roddick, 2006). The system of association rule mining needs to facilitate the discovery, heuristically filter, and enable the presentation of the mined rules for subsequent interpretation by users to investigate their usefulness.

Based on the above discussion, this paper, therefore, develops a rule evaluation method based on DEA and DT. The proposed rule evaluation method can help decision makers to interpret and investigate the mined association rules for subsequent applications. In the proposed method, DEA is used to calculate the efficiency of association rules with multiple criteria, and rank the rules by interestingness (efficiency) in order to choose the proper rules for implementation. After evaluating the efficiency of association rules, we use DT to build a classifier to find out the characteristics of rule interestingness. The constructed classifier also can be used to predict whether the unknown (new) association rules are efficient or not.

The remainder of this paper is divided into four sections. Section 2, introduces the objective and subjective measures in mining association rules. The proposed approach for pattern filtering and classification is presented in Section 3. Section 4, the computational results are presented and discussed. Finally, conclusions of this paper are made in Section 5.

2. Objective and subjective measures

Association rule mining is commonly used to analyze the customer transaction database, and to describe the relationships between product items directly from the databases (Agrawal *et al*., 1993). The marketing analyst usually tries to search for a parsimonious representation of the cross-category associations by applying multidimensional scaling techniques or hierarchical clustering (Boztuğ & Reutterer 2008). However, these methods are limited to a relatively small number of categories with symmetric pairwise relationships. From the literature review in

Boztuğ and Reutterer, the research of association rule mining successfully resolves this limitation. Exploratory and explanatory (or predictive) models are two main types of approaches for market basket data analysis (Mild & Reutterer, 2003; Boztuğ & Reutterer, 2008). Boztuğ and Reutterer indicated that exploratory models are limited to discovering notable cross-category interrelationships with respect to observed patterns of simultaneously bought product items or categories. Data mining of rule discovery can handle both very huge numbers of product categories or items and market baskets, but the issue of an aggregate market view still needs to be resolved. Explanatory choice models (Boztuğ & Reutterer, 2008) mainly focus on estimating the effects of marketing-mix variables on category purchase incidences. Most existing explanatory models for market basket analysis are based on either logit or probit models. For explanatory approaches, the set of categories to be included for analyzing cross-category effects on the selected response category is rather restricted. As discussed above, Boztuğ & Reutterer (2008) therefore proposed a two-stage approach for market basket analysis, which integrates characteristics from exploratory and model-based traditions.

Association rules also can find out the frequent itemsets, that is, the sets of product items frequently bought together, and predict the customers' behaviours. Association rules are widely used in traditional business and e-business regions (Choi *et al*., 2005). The former includes item allocation, product assortment, cross-marketing, catalog design, customer segmentation, etc. (Agrawal *et al*., 1993; Srikant *et al*., 1997; Chen & Lin, 2007). The association rules obtained from transaction data can be used to identify which products are frequently bought together by customers. Provided that certain products appear in a market basket, decision makers can infer that certain other products would be bought. Therefore, association rules can be used to allocate products on shelves, recommend products to a customer, and so on. The e-business applications include web personalization (Mobasher *et al*., 2000) and recommender systems (Leung *et al*., 2008). In the WebPersonalizer system developed by Mobasher *et al*. (2000), association rules are used to presenting the relationships among Uniform Resource Identifiers (URIs) based on users' navigational patterns. Mobasher and colleagues applied frequent itemsets discovery (frequent URI sets) of association rules as one of the methods to directly obtain groups of URIs based on users' pageview clusters. Leung *et al*. (2008) developed a cross-level association rule based recommender system, in which user-item and item-item associations are combined to build the preference model. The main characteristic of their proposed recommender system is applying the information of associations between the attributes of a given item and other domain items to address the problem of no recommendations by using collaborative filtering.

It is necessary for decision makers to discover the meaningful and useful patterns or rules from databases in data mining applications. One can identify the interestingness of association rules by using both objective and subjective measures. Geng & Hamilton (2006) summarized nine criteria for rule interestingness, and they can be categorized into three categories, objective, subjective and semantics-based measures. Objective measures are based on the theories of statistics and probability. The user's domain and background knowledge are considered in the subjective measures. Semantics-based measures take into account the semantics of patterns, and also consider the domain knowledge of users. Semantics-based measures are assigned to the category of subjective measures in this paper.

The objective measures in association rules consist of *support*, *confidence* and *lift*; the subjective measures consist of unexpectedness and actionability (Liu *et al*., 2000). *Support* and *confidence* are objective measures most frequently used for association rule mining (Olafsson *et al*., 2008; Lenca *et al*., 2008). However, redundant and uncorrelated association rules may be generated if only *minimum support* and *minimum confidence* are used to filter the rules (Wei *et al*., 2006). Therefore, some previous studies discussed the interestingness of rules combining other measures to evaluate the

usefulness of discovered rules. For example, Chen (2007) combined the product value and cross-selling profit with *support* and *confidence* to select interesting association rules, and utilized DEA to calculate the efficiency of rules. Luo & Wu (2002) used *validity* to replace *confidence* for association rules. The concept of *validity* can diminish the number of rules, but it does not have a great effect on reducing the uncorrelated rules. Therefore, Wei *et al.* (2006) proposed the framework of *support-match* to replace *support-confidence*. Their experimental results revealed that the approach proposed by Wei *et al.* (2006) not only can reduce the redundant and low-association rules, but also generate higher correlation between the antecedent and consequent of rules.

3. The rule filtering and classification approach

3.1. Overview

This section presents the proposed process of pattern filtering and classification, which hybridize association rule mining, DEA and DT. The flowchart of the proposed procedure is schematically illustrated in Figure 1, and the procedure steps are described as follows:

- Step 1: Import transaction database and product database, and set *min sup* and *min conf* for mining association rules.
- Step 2: Generate association rules by using the Apriori algorithm.
- Step 3: Obtain the objective measures and calculate the subjective measures for each association rule.
- Step 4: Compute the efficiency of each association rule by using Ranked Voting DEA. Set an efficiency threshold to categorize the association rules into relatively efficient and relatively inefficient.
- Step 5: Find the characteristics of relatively efficient and relatively inefficient association rules by using DT algorithms.
- Step 6: Select the interesting or profitable association rules for marketing implementation. And, use the DT classifier to classify the new association rules.

3.2. Steps 1 and 2: mining association rules

The Apriori algorithm is a popular algorithm for mining association rules (Agrawal & Srikant, 1994). Let $I = \{i_1, i_2, \ldots, i_m\}$ represent a set of items, and let D be a set of transactions where each transaction $T \in D$ is a set of items such that $T \subset I$. Each transaction has an identifier, namely TID. Every rule can use three measures (*support*, *confidence* and *lift*) to evaluate the characteristics of rule frequency and relationship. An association rule implies the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ has *support* s in the transaction set D if s% of transactions in D contain $X \cup Y$, and the rule $X \Rightarrow Y$ has *confidence* c if c% of transactions in D that contain X also contain Y.

The Apriori algorithm developed by Agrawal *et al.* (1993) and Srikant & Agrawal (1997) is efficient because it restricts the search space and checks only a subset of all association rules, but does not miss important rules. In association rule mining, Apriori is a fundamental algorithm, on which many extended algorithms are based (Ceglar & Roddick, 2006). In the Apriori algorithm, two operators, JOIN and PRUNE, are mainly used to generate association rules (Srikant & Agrawal, 1997; Han & Kamber, 2006). The JOIN operator can generate potential itemset candidates. The PRUNE operator uses

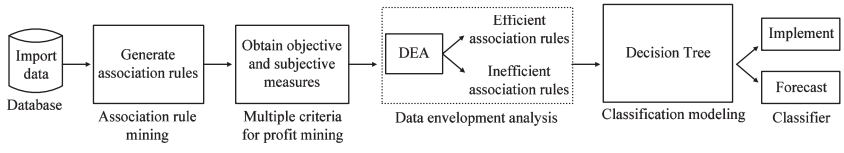

Figure 1: *The flowchart of pattern filtering and classification.*

the minimum support criterion to remove candidates of itemset that are not frequent. The details of the Apriori algorithm can be found in Agrawal *et al.* (1993) and Srikant & Agrawal (1997).

Additionally, *lift* represents the correlation between the occurrence of X and Y. If *lift* is >1.0, then the occurrence of X is positively correlated with Y. If *lift* is <1.0, they are negatively correlated. If *lift* is equal to 1, X and Y are independent. In order to avoid missing the high-value itemsets, *min sup* and *min conf* are set relatively low in the proposed procedure.

3.3. Step 3: obtaining evaluation measures

The objective measures consist of *support*, *confidence* and *lift*. For the rule $X \Rightarrow Y$, they are mathematically expressed as follows (Agrawal & Srikant, 1994; Han & Kamber, 2006):

$$\text{Support}(X \Rightarrow Y) = P(X \cup Y) \tag{1}$$

$$\text{Confidence}(X \Rightarrow Y) = P(Y|X) \tag{2}$$

$$\text{Lift}(X \Rightarrow Y) = \frac{P(X \cup Y)}{P(X)P(Y)} = \frac{P(Y|X)}{P(Y)} \tag{3}$$

The product database is also imported herein to compute the subjective measures. The subjective measures of rule $X \Rightarrow Y$ used in this paper include rule value (Chen, 2007), cross-selling profit (Chen, 2007) and rule profit. As mentioned above, the products with a higher value may potentially belong to interesting infrequent itemsets, so the product value can also be considered as a subjective measure. The rule value is measured by the summation of product prices in a rule. It takes the form of (Chen, 2007)

$$\text{Rule value} = \sum_{i=1}^{n} PX_i + \sum_{j=1}^{m} PY_j \tag{4}$$

where PX_i represents the price of product i in X, PY_j represents the price of product j in Y, and n and m, respectively, represent the numbers of products in X and Y.

The measure of rule value as expressed in equation (4) does not consider the product cost. Each product's profit can be obtained by the difference between price and cost. The rule profit can be also taken as a subjective measure, and it is measured in terms of the total profit in a rule. The total profit of a rule is expressed as

$$\text{Rule profit} = \sum_{i=1}^{n} (PX_i - CX_i) + \sum_{j=1}^{m} (PY_j - CY_j) \tag{5}$$

where CX_i represents the cost of product i in X, and CY_j represents the cost of product j in Y.

Knott *et al.* (2002) raised some cross-selling examples such as an online book business intending to recognize which other books it should direct to its customers, or a photo store with an in-store self-serve kiosk identifying which additional attributes it should promote to its customers, when customers are using the kiosk, and so on. Applying customer transaction databases for cross-selling of new services and products has been an important perspective in customer relationship marketing (Kamakura *et al.*, 2003). Through implementing cross-selling, marketers can offer customers products and services that fit their needs, but that they have not bought so far. The cross-selling profit therefore can be taken as the summation of profits in Y. Hence, it is expressed as (Chen, 2007)

$$\text{Cross selling profit} = \sum_{j=1}^{m} (PY_j - CY_j) \tag{6}$$

3.4. Step 4: generating rule efficiency

This proposed procedure uses DEA to calculate the efficiency of association rules with multiple criteria. After generating the association rules, the ranked voting DEA developed by Cook & Kress (1990) is then applied to generate the efficiency value of each rule. An association rule is taken as a DMU. The measure *lift* is not included in the ranked voting DEA model because it should be >1.0 to guarantee X is positively related to Y. The ranked voting DEA model is formulated as follows (Cook & Kress, 1990).

$$\text{Maximize } z_o = \sum_{j=1}^{k} w_j v_{oj} \tag{7}$$

Subject to:

$$\sum_{j=1}^{k} w_j v_{ij} \leq 1,\ i = 1, 2, \ldots, p \tag{8}$$

$$w_j - w_{j+1} \geq d(j, \varepsilon),\ j = 1, 2, \ldots, k-1 \tag{9}$$

$$w_j \geq d(k, \varepsilon) \tag{10}$$

where z_o denotes the *desirability index* (efficiency value) of oth candidate (rule), w_j denotes the weight of the jth place vote; v_{ij} represents the number of jth place votes of candidate i ($i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, k$, in which p is the number of candidates, and k is the number of place votes); and $d(\bullet, \varepsilon)$, known as the *discrimination intensity function*, is nonnegative and nondecreasing in ε and satisfies $d(\bullet, 0) = 0$. Parameter ε, called *discriminating factor*, is nonnegative.

The above mathematical model is solved for each candidate o. The objective is to maximize the *desirability index* of candidate o. The best attainable performance level is set to 1 as shown in Constraint set (8). Constraint set (9) ensures that the vote of the higher place may have a greater importance than that of the lower place. Without setting the priorities of criteria, Constraint set (9) is relaxed. In the original DEA, candidates with *desirability index* (preference score) of 1.0 are called *efficient candidates*. Readers are referred to Cook & Kress (1990) for further details of this ranked voting DEA model. In this paper, we label the rule with an efficiency value higher than or equal to a threshold as relatively efficient; otherwise relatively inefficient.

3.5. *Steps 5 and 6: building DT classifier and implementation*

With the above DEA model, the efficiency of each association rule can be calculated by using the evaluation measures. By setting an efficiency threshold, the association rules can be categorized into relatively efficient and relatively inefficient. DT is hence applied to construct the classifier, which can describe the characteristics of efficient and inefficient rules. As well, the DT classifier can serve as a model to predict new association rules.

Selecting the best attribute for split is essential to DT algorithms. Goodness of split is estimated by impurity in which entropy, Gini index and χ^2 statistics are usually used as the measures (Tan *et al.*, 2006). With these impurity measures, Entropy Reduction, Gini Reduction and a χ^2-test are used to develop different DT algorithms. Entropy, Gini index and χ^2 statistics are mathematically expressed as the following equations (Tan *et al.*, 2006):

$$\text{Entropy}(t) = -\sum_{i=0}^{c-1} P(i|t) \log_2 P(i|t) \tag{11}$$

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [P(i|t)]^2 \tag{12}$$

$$\text{Classification error}(t) = 1 - \max_i [P(i|t)] \tag{13}$$

where c is the number of classes, and $P(i|t)$ represents the percentage of records belonging to class i in node t.

After obtaining the DT classifier, we can find out the characteristics of relatively efficient and relatively inefficient association rules by exploring the DT classifier for each path starting from root node to leaf node (Han & Kamber, 2006). Classifying a test object is straightforward once a DT is built. Several previous studies (e.g. Polat *et al.* (2009), Uhmn *et al.* (2009)) have compared the performance of some classifiers, and demonstrated that DT may not be the best classifier compared with other classifiers. However, DT, particularly a small-sized tree, is easy to explain (Tan *et al.*, 2006) since it can generate interpretive rules.

4. Computational results

This section presents the computational results by using the proposed pattern filtering and classification approach developed in the previous section with a set of experiments.

Table 1: *Summary of data columns*

Column name	*Description*	*Randomly generate*
TRX_CODE	Transaction ID	No
ITEM_CODE	Product item ID	No
ITEM_PRICE	Product selling price	Yes
ITEM_COST	Product cost	Yes
ITEM_TOTAL_QTY	Product sale quantity	Yes

4.1. Data

In this paper, data are obtained from the SuperMarket database in XAffinity® software (http://www.xore.com, Exclusive Ore® Inc.). The database includes tables of product items, product categories and transaction data in a supermarket. The columns used in this paper are listed in Table 1. Transaction ID, TRX_CODE and product item ID, ITEM_CODE, are retrieved from the SuperMarket database for generating association rules with *Apriori*. There are 20,245 product items with 98,742 transaction records. To obtain the subjective measures (rule value, rule profit and cross-selling profit), product selling price, cost and sale quantity are additionally required, and they are generated in a random manner herein.

4.2. Examples of association rules

After retrieving and pre-processing the necessary data, association rules are mined by using the data mining tool SAS® Enterprise Miner™ (http://www.sas.com, SAS Institute Inc.). Because there are massive product items and transaction records, and since we intend to find the potentially profitable patterns, the thresholds of *min sup* and *min conf* are set relatively low, 0.5% and 10%, respectively. The maximum length of frequent itemset is set to 4. With such a setting, 152 association rules are generated.

Taking the following association rules for illustration,

Rule 9: Item = 1017700 ⇒ Item = 1009636,

Objective measures: [*support* = 2.38%, *confidence* = 34.04, *lift* = 3.33],

Subjective measures: [*rule value* = US$850, *rule profit* = US$308, *cross-selling profit* = US$73]

Rule 122: Item = 1028068 ⇒ Item = 1017637,

Objective measures: [*support* = 0.54%, *confidence* = 27.85, *lift* = 1.76],

Subjective measures: [*rule value* = US$1,690, *rule profit* = US$620, *cross-selling profit* = US$271]

Observing the above two association rules, Rule 9 will be selected by using the traditional association rule mining with only the objective measures, support, confidence and lift. Although, Rule 122 has lower support and confidence, it is certainly more profitable to enterprises in terms of subjective measures. Rule 122 has higher rule value, rule profit and cross-selling profit. In such a situation, decision makers may have a dilemma to select either Rule 9 or Rule 122 for marketing implementation.

4.3. The DEA analysis

The discovered 152 rules (DMUs) with both objective and subjective measures were used as input to the ranked voting DEA model for calculating the efficiency values. Without setting the importance of measure, that is, relaxing Constraint set (9), Rules 9 and 122 have efficiency values of 0.82 and 1.00, respectively. Considering the efficiency, Rule 122 is more interesting than Rule 9. As mentioned above, taking the profit-based measures (subjective measures) into account, decision makers may discover the potentially profitable association rules for implementations such as bundle sale, shelf space management, promotion, etc.

By setting the priority for objective and subjective measures, three analysis scenarios are defined in performing the DEA model coded in the optimization modelling tool LINGO (http://www.lindo.com, Lindo Systems). In Scenario A, the weight priority is not set between objective and subjective measures. Scenario B sets objective measures as more important than the subjective ones. On the contrary, Scenario C sets subjective measures as more important than the objective ones. The 20 most efficient association rules of Scenarios A, B and C are summarized in Table 2.

As mentioned above, decision makers may have a dilemma to select either Rule 9 or Rule 122 for marketing implementation since Rule 9 is more interesting in terms of objective measures, but Rule 122 is more profitable in terms of subjective measures. By using the DEA model in Cook & Kress (1990), the efficiency values of Rule 122 are 1.0000 in Scenarios A, B and C, but the efficiency values of Rule 9 are respectively 0.8223, 0.5030 and 0.8223 in Scenarios A, B and C. Based on the efficiency value, decision makers can select Rule 122 for designing the cross-selling marketing; if customers buy product item 1028068, product item 1017637 will be recommended to them.

Table 2: *The 20 most efficient association rules*

Scenario A		*Scenario B*		*Scenario C*	
Rule no.	*Efficiency*	*Rule no.*	*Efficiency*	*Rule no.*	*Efficiency*
1	1.0000	76	1.0000	1	1.0000
2	1.0000	85	1.0000	2	1.0000
4	1.0000	101	1.0000	4	1.0000
11	1.0000	119	1.0000	11	1.0000
21	1.0000	122	1.0000	21	1.0000
31	1.0000	57	0.9999	31	1.0000
35	1.0000	144	0.9978	35	1.0000
42	1.0000	115	0.9942	42	1.0000
57	1.0000	113	0.9883	57	1.0000
63	1.0000	31	0.9827	63	1.0000
76	1.0000	90	0.9760	76	1.0000
85	1.0000	150	0.9743	85	1.0000
101	1.0000	63	0.9638	101	1.0000
119	1.0000	55	0.9613	119	1.0000
122	1.0000	21	0.9584	122	1.0000
144	0.9978	25	0.9584	144	0.9978
115	0.9947	37	0.9584	115	0.9946
25	0.9932	104	0.9584	25	0.9932
113	0.9894	142	0.9436	113	0.9894
12	0.9889	87	0.9346	12	0.9888

4.4. Experimental results of DTs

DT can be used to find out the characteristics of relatively efficient association rules and relatively inefficient ones as well as serving as a classifier to predict the new association rules. Because there exist few association rules with efficiency 1.0, we set a threshold of efficiency to categorize association rules into two classes, relatively efficient and relatively inefficient. Three DT algorithms respectively based on χ^2-test, Entropy Reduction and Gini Reduction in SAS® Enterprise Miner™ are applied to classify the discovered association rules.

Three thresholds, 1.0, 0.9 and 0.8, are, respectively, set for analysis. The number of relatively efficient rules and the number of inefficient rules with three various thresholds for Scenarios A, B and C are listed in Table 3. For building the DT classifiers, 152 rules are divided into a training dataset (70%) and test dataset (30%) by using stratified sampling according to the number of relatively efficient and the number of relatively inefficient rules.

4.4.1. Results of Scenarios A and C For the cases of three thresholds in Scenario A, the classification results by three DT algorithms are summarized in Table 4. From this table, the three DT algorithms perform equally. For the case of efficiency threshold set to 0.9, all three DT algorithms generate the same classifier as shown in Figure 2.

As abovementioned, the DT model as shown in Figure 2 can assist decision makers in interpreting and investigating the discovered association rules such that they can understand why the rules are classified as efficient or inefficient. As shown in Figure 2, there are seven leaf

Table 3: *The numbers of relatively efficient and relatively inefficient rules*

Threshold	*Scenario A*		*Scenario B*		*Scenario C*	
	Efficient no.	*Inefficient no.*	*Efficient no.*	*Inefficient no.*	*Efficient no.*	*Inefficient no.*
1.0	15	137	5	147	15	137
0.9	42	110	26	126	42	110
0.8	80	72	40	112	80	72

(terminal) nodes representing seven DT rules. Observing the first node in the tree, decision makers can identify that Rule Value is the most critical criterion to select relatively efficient association rules for implementation. The first node branches into two child nodes (Rule Value ≥ 1480 and Rule Value < 1480). In the right child node, there are 11 and 8 efficient association rules in the training dataset and test dataset, respectively. By traversing the tree path, DT rules can be easily found. Taking the following two DT rules as examples:

IF Rule Value ≥ 1480, **THEN** the association rule is labelled as relatively efficient.

IF Rule Value < 1480 **AND** Rule Cross Selling Profit < 289.5 **AND** Support < 0.014875 **AND** Confidence < 0.524171, **THEN** the association rule is labelled as relatively inefficient.

For the second DT rule as shown above, there are 0 (0%) efficient association rules and 69 (100%) inefficient association rules in the training dataset, and one (3.1%) efficient association rule and 31 (96.9%) inefficient association rules in the test dataset. Other DT rules can be easily interpreted in a similar manner.

By chance, the results of Scenario C are the same as that of Scenario A.

Table 4: *Classification results of Scenario A*

	χ^2-*Test*	*Entropy reduction*	*Gini reduction*
Efficiency threshold = 1.0			
Leaf node no.	3	3	7
Training accuracy	0.9340	0.9340	0.9528
Test accuracy	0.8913	0.8913	0.9130
Efficiency threshold = 0.9			
Leaf node no.	7	7	7
Training accuracy	1.0000	1.0000	1.0000
Test accuracy	0.9783	0.9783	0.9783
Efficiency threshold = 0.8			
Leaf node no.	5	5	5
Training accuracy	0.9245	0.9623	0.9623
Test accuracy	0.8478	0.9565	0.9565

4.4.2. Results of Scenario B For the cases of three thresholds in Scenario B, the classification

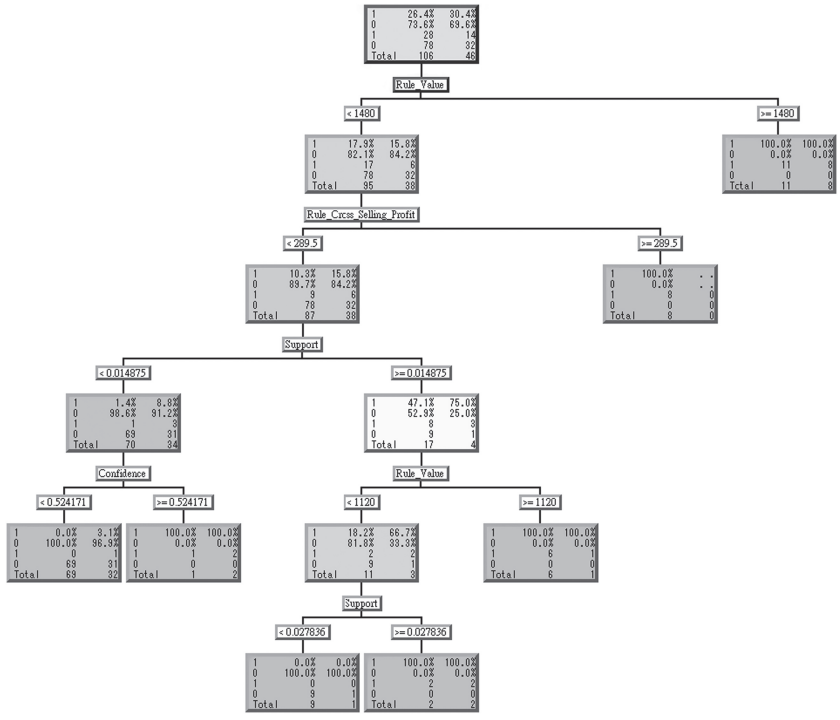

Figure 2: *The decision tree classifier of Scenario A with efficiency threshold set to 0.9.*

Table 5: *Classification results of Scenario B*

	Chi-square Test	*Entropy reduction*	*Gini reduction*
Efficiency threshold = 1.0			
Leaf node no.	4	3	4
Training accuracy	1.0000	0.9811	1.0000
Test accuracy	0.9783	0.9783	0.9783
Efficiency threshold = 0.9			
Leaf node no.	3	3	3
Training accuracy	1.0000	1.0000	1.0000
Test accuracy	1.0000	1.0000	1.0000
Efficiency threshold = 0.8			
Leaf node no.	3	3	3
Training accuracy	1.0000	1.0000	1.0000
Test accuracy	0.9783	0.9783	0.9783

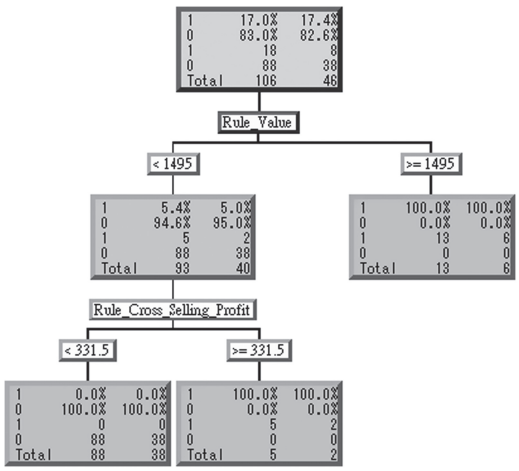

Figure 3: *The decision tree of Scenario B with efficiency threshold set to 0.9.*

results by DT are summarized in Table 5. From this table, the three DT algorithms perform equally well. For the case of efficiency threshold set to 0.9, all three DT algorithms generate the same classifier as shown in Figure 3. The three DT rules are presented as follows:

IF Rule Value ≥ 1495, **THEN** the association rule is labelled as relatively efficient.

IF Rule Value < 1495 **AND** Rule Cross Selling Profit < 331.5, **THEN** the association rule is labelled as relatively inefficient.

IF Rule Value < 1495 **AND** Rule Cross Selling Profit $>= 331.5$, **THEN** the association rule is labelled as relatively efficient.

5. Conclusions

Data mining techniques are expected to extract useful rules or patterns from a massive database. However, a large set of patterns as well may be extracted by performing the data mining functions. Because not all the patterns generated are useful or interesting to decision makers, pattern evaluation therefore plays an essential role in the Knowledge Discovery in Databases (KDD) procedure. Decision makers can additionally set criteria based on their domain and background knowledge to select patterns for marketing implementation. Such subjective criteria cannot be measured statistically or directly from the database. In this paper, we propose a pattern filtering and classification approach by simultaneously taking both objective and subjective measures into consideration.

The DEA model can filter a large set of discovered association rules for making better marketing decisions. Additionally, decision trees can help decision makers interpret the discovered association rules, and build classifiers to predict relatively efficient or inefficient rules. The proposed approach provides an alternative for enhancing the pattern evaluation in KDD process.

This paper focuses on evaluating discovered association rules with profit-based measures, mainly in terms of price, to tackle the issue raised in the *Ketel Vodka and Beluga Caviar* problem (Cohen *et al.*, 2001), so future works can additionally take other marketing-mix variables into consideration.

Acknowledgements

The authors would like to thank the National Science Council of the Republic of China, Taiwan for financially supporting this research under Contract No. NSC 95-2416-H-009-034-MY3.

References

AGRAWAL, R., T. IMIELINSKI and A. SWAMI (1993) Mining association rules between sets of items in large databases, in *Proceedings of the ACM SIGMOD Conference on Management of Data*, 26–28 May 1993, Washington, DC, USA, 254–259.

AGRAWAL, R. and R. SRIKANT (1994) Fast algorithms for mining association rules, in *Proceedings of 20th International Conference on Very Large Data Bases*, 12–15 September 1994, Santiago de Chile, Chile, 487–499.

BOZTUĞ, Y. and T. REUTTERER (2008) A combined approach for segment-specific market basket analysis, *European Journal of Operational Research*, **187**, 294–312.

CEGLAR, A. and J.F. RODDICK (2006) Association mining, *ACM Computing Surveys*, **38**, 1–42.

CHEN, M.-C. (2007) Ranking discovered rules from data mining with multiple criteria by data envelopment analysis, *Expert Systems with Applications*, **33**, 1110–1116.

CHEN, M.-C. and C.-P. LIN (2007) A data mining approach to product assortment and shelf space allocation, *Expert Systems with Applications*, **32**, 976–986.

CHOI, D.H., B.S. AHN and S.H. KIM (2005) Prioritization of association rules in data mining- multiple criteria decision approach, *Expert Systems with Applications*, **29**, 867–878.

COHEN, E., M. DATAR, S. FUJIWARA, A. GIONIS, P. INDYK, R. MOTWANI, J.D. ULLMAN and C. YANG (2001) Finding interesting associations without support pruning, *IEEE Transactions on Knowledge and Data Engineering*, **13**, 64–78.

COOK, W.D. and M. KRESS (1990) A data envelopment model for aggregating preference rankings, *Management Science*, **36**, 1302–1310.

DULÁ, J.H. (2008) A computational study of DEA with massive data sets, *Computers & Operations Research*, **35**, 1191–1203.

GENG, L. and H.J. HAMILTON (2006) Interestingness measures for data mining: a survey, *ACM Computing Surveys*, **38**, 1–32.

HAN, J. and M. KAMBER (2006) *Data Mining: Concepts and Techniques*, 2nd edn, San Francisco, CA: Morgan Kaufmann Publishers.

HSU, C. and W.A. WALLACE (2007) An industrial network flow information integration model for supply chain management and intelligent transportation, *Enterprise Information Systems*, **1**, 327–351.

KAMAKURA, W.A., M. WEDEL, F. DE ROSA and J.A. MAZZON (2003) Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and prediction, *International Journal of Research in Marketing*, **20**, 45–65.

KNOTT, A., A. HAYES and S.A. NESLIN (2002) Next-product-to-buy models for cross-selling applications, *Journal of Interactive Marketing*, **16**, 59–75.

LEE, J.H. and S.C. PARK (2005) Intelligent profitable customers segmentation system based on business intelligence tools, *Expert Systems with Applications*, **29**, 145–152.

LENCA, P., P. MEYER, B. VAILLANT and S. LALLICH (2008) On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid, *European Journal of Operational Research*, **184**, 610–626.

LEUNG, C.W.-K., S.C.-F. CHAN and F.-L. CHUNG (2008) An empirical study of a cross-level association rule mining approach to cold-start recommendations, *Knowledge-Based Systems*, **21**, 515–529.

LIU, B., W. HSU, S. CHEN and Y. MA (2000) Analyzing the subjective interestingness of association rules, *Intelligent Systems and Their Applications, IEEE*, **15**, 47–55.

LUO, K. and J. WU (2002) How to get valid association rules, *Mini-Micro System*, **23**, 711–713.

MILD, A. and T. REUTTERER (2003) An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data, *Journal of Retailing and Consumer Services*, **10**, 123–133.

MOBASHER, B., R. COOLEY and J. SRIVASTAVA (2000) Automatic personalization based on web usage mining, *Communications of the ACM*, **43**, 142–151.

OLAFSSON, S., X. LI and S. WU (2008) Operations research and data mining, *European Journal of Operational Research*, **187**, 1429–1448.

POLAT, K., S. KARA, A. GÜVEN and S. GÜNEŞ (2009) Comparison of different classifier algorithms for diagnosing macular and optic nerve diseases, *Expert Systems*, **26**, 22–34.

RAMANATHAN, R. (2006) ABC inventory classification with multiple-criteria using weighted linear optimization, *Computers & Operations Research*, **33**, 695–700.

ROKACH, L. and O MAIMON (2005) Top down induction of decision trees classifiers: a survey, *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, **35**, 476–487.

SAFAVIAN, S.R. and D. LANDGREBE (1991) A survey of decision tree classifier technology, *IEEE Transactions on Systems, Man and Cybernetics*, **21**, 660–674.

SAMOILENKO, S. and K.M. OSEI-BRYSON (2008) Increasing the discriminatory power of DEA in the presence of the sample heterogeneity with cluster analysis and decision trees, *Expert Systems with Applications*, **34**, 1568–1581.

SEOL, H., J. CHOI, G. PARK and Y. PARK (2007) A framework for benchmarking service process using data envelopment analysis and decision tree, *Expert Systems with Applications*, **32**, 432–440.
SOHN, S.T. and T.H. MOON (2004) Decision tree based on data envelopment analysis for effective technology commercialization, *Expert Systems with Applications*, **26**, 279–284.
SRIKANT, R. and R. AGRAWAL (1997) Mining generalized association rules, *Future Generation Computer Systems*, **13**, 161–180.
SRIKANT, R., Q. VU and R. AGRAWAL (1997) Mining association rules with item constraints, in *Proceedings 3rd International Conference Knowledge Discovery and Data Mining (KDD 97)*, 14–17 August 1997, Newport Beach, CA, USA, 67–73.
TAN, P.N., M. STEINBACH and V. KUMAR (2006) *Introduction to Data Mining*, Boston, MA: Addison Wesley.
TAO, F., F. MURTAGH and M. FARID (2003) Weighted association rule mining using weighted support and significance framework, in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 24–27 August 2003, Washington, DC, USA, 661–666.
UHMN, S., D.-H. KIM, Y.-W. KO, S. CHO, J. CHEONG and J. KIM (2009) A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis, *Expert Systems*, **26**, 60–69.
WANG, K., S. ZHOU and J. HAN (2002) Profit mining: from patterns to actions, in *Proceedings of the 8th International Conference on Extending Database Technology (EDBT 2002)*, 25–27 March 2002, Prague, Czech Republic, 70–87.
WEI, J.M., W.G. YI and M.Y. WANG (2006) Novel measurement for mining effective association rules, *Knowledge-Based Systems*, **19**, 739–743.
XU, L., C. WANG, X. LUO and Z. SHI (2006) Integrating knowledge management and ERP in enterprise information systems, *Systems Research and Behavioral Science*, **23**, 147–156.

The authors

Mu-Chen Chen

Mu-Chen Chen is a professor of the Institute of Traffic and Transportation in National Chiao Tung University, Taipei, Taiwan. He received his PhD and MSc degrees both in Industrial Engineering and Management from National Chiao Tung University, and his BS degree in Industrial Engineering from Chung Yuan Christian University. His teaching and research interests include Data Mining, Logistics and Supply Chain Management and Meta-heuristics.

Chuang-Min Chao

Chuang-Min Chao is an assistant professor of the Department of Business Management, National Taipei University of Technology, Taipei, Taiwan. She got her PhD degree in Economics from the Department of Economics, University of Rochester, NY, USA. Her research topics are related to the efficiency measurement using Data Envelopment Analysis or Stochastic Frontier Analysis, corporate governance, as well as the investors' behaviour in the stock market.

Kuan-Ting Wu

Kuan-Ting Wu received his MBA from the Institute of Commerce Automation and Management, National Taipei University of Technology, Taipei, Taiwan. His research interests include Data Mining and Customer Relationship Management.