Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 33 (2007) 1110–1116
Expert Systems with Applications
www.elsevier.com/locate/eswa

Ranking discovered rules from data mining with multiple criteria by data envelopment analysis

Mu-Chen Chen *

Institute of Traffic and Transportation, National Chiao Tung University 4F, No. 118, Section 1, Chung Hsiao W. Road, Taipei 10012, Taiwan, ROC

Abstract

In data mining applications, it is important to develop evaluation methods for selecting quality and profitable rules. This paper utilizes a non-parametric approach, Data Envelopment Analysis (DEA), to estimate and rank the efficiency of association rules with multiple criteria. The interestingness of association rules is conventionally measured based on support and confidence. For specific applications, domain knowledge can be further designed as measures to evaluate the discovered rules. For example, in market basket analysis, the product value and cross-selling profit associated with the association rule can serve as essential measures to rule interestingness. In this paper, these domain measures are also included in the rule ranking procedure for selecting valuable rules for implementation. An example of market basket analysis is applied to illustrate the DEA based methodology for measuring the efficiency of association rules with multiple criteria.

Keywords: Data mining; Association rules; Interestingness; Data envelopment analysis; Multiple criteria

1. Introduction

Data mining techniques have become widespread in business. Moreover, various rules may be obtained using data mining techniques, and only a small number of these rules may be selected for implementation due, at least in part, to limitations of budget and resources. Association rule mining differs from traditional machine learning techniques by permitting decision makers to pick from the many potential models that can be supported by the data (Webb & Zhang, 2005). Generally, association rule mining discovers all rules that meet certain sets of criteria or constraints, such as minimum support and minimum confidence, rather than generating a single model that best matches the data.

Evaluating the interestingness or usefulness of association rules is important in data mining. In many business applications, it is necessary to rank rules from data mining due to the number of quality rules (Tan & Kumar, 2000) and business resource constraint (Choi, Ahn, & Kim, 2005). Selecting the more valuable rules for implementation increases the possibility of success in data mining. For example, in market basket analysis, understanding which products are usually bought together by customers and how the cross-selling promotions are beneficial to sellers both attract marketing analysts. The former makes sellers to provide appropriate products by considering the customers' preferences, and the later allows sellers to gain increased profits by considering the sellers' profits. Customers' preferences can be measured based on support and confidence in association rules. On the other hand, seller profits can be assessed using domain related measures such as sale profit and cross-selling profit associated with the association rules.

Since high value products are relatively uncommonly bought by customers, a rule that is profitable to sellers may not be discovered by setting constraints of minimum support and minimum confidence in the mining process.

* Tel.: +886 2 2349 4967; fax: +886 2 2349 4953.
E-mail address: ittchen@mail.nctu.edu.tw

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.08.007

Cohen et al. (2000) described a good example of this, namely the *Ketel vodka and Beluga caviar* problem. Although, most customers infrequently buy either of these two products, and they rarely appear in frequent itemsets, their profits may be potentially higher than many lower value products that are more frequently bought. Another example regarding the interesting infrequent itemsets is described in Tao, Murtagh, and Farid (2003). The association rule of [*wine* ⇒ *salmon*, 1%, 80%] may be more interesting to analysts than [*bread* ⇒ *milk*, 3%, 80%] despite the first rule having lower support. The items in the first rule typically are associated with more profit per unit sale.

From the examples of *Ketel vodka and Beluga caviar* and *wine and salmon*, infrequent itemsets may be interesting for certain applications provided that domain information is considered (Tao et al., 2003; Webb & Zhang, 2005). However, the traditional association rule mining algorithms (Agrawal, Imielinski, & Swami, 1993; Srikant & Agrawal, 1997) cannot classify such infrequent products to interesting itemsets since the subjective domain knowledge is ignored. A lower threshold can be set to identify the infrequent itemsets with a high value. However, numerous association rules are consequently generated, and it is extremely difficult for analysts to select the useful rules between them.

In previous studies dealing with the discovery of subjectively interesting association rules, most approaches require manual input or interaction by asking users to explicitly distinguish between interesting and uninteresting rules (Liu, Hsu, Chen, & Ma, 2000). Liu et al. briefly reviewed these existing approaches. The measures of interestingness are specified as constraints in the mining process, and only the rules that satisfied these constraints are retrieved. Klemetinen, Mannila, Ronkainen, Toivonen, and Verkamo (1994) proposed an item constraint, which describes the occurrence of certain items in the conditional (right hand side) and consequent (left hand side) parts. Srikant, Vu, and Agrawal (1997) also proposed a mining algorithm that considered the item and item hierarchy constraints specified by analysts. Moreover, Lakshmanan, Han, and Pang (1998) extended the approach developed by Srikant et al. to consider much more complicated constraints, including domain, class, and SQL-style aggregate constraints. The approach developed by Ng et al. can support constraint based, human-centered exploratory mining of association rules. Goethals and Van den Bussche (2000) also proposed an interactive approach based on querying conditions within the association rule mining process.

Liu et al. (2000) proposed an approach to assist analysts in finding interesting rules from a set of mined association rules by analyzing the rules using the domain information. The mined rules are then ranked according to two subjective interestingness measures, *unexpectedness* and *actionability*. The degree of unexpectedness of rules can be measured by the extent to which they surprise the analyst (Liu & Hsu, 1996; Silberschatz & Tuzhilin, 1996). Meanwhile, the degree of actionability can be measured by the extent to which analysts can use the discovered rules to their advantage. The system developed by Liu et al. (2000) is an interactive and iterative post-processing technique. This system first asks analysts to specify their existing domain knowledge, and then analyzes the discovered rules to identify the potentially interesting ones. However, Liu et al. focused on unexpected rules, which are measured by unexpectedness.

Choi et al. (2005) proposed a group decision making approach based on Analytic Hierarchy Process (AHP) to rank the association rules generated from data mining. This approach would construct a consensus provided that a group of managers work together to select discovered. The rule quality can be improved by considering both objective criteria and subjective preferences of managers. However, this approach encounters a problem of requiring considerable human interaction to find out the weights of criteria by aggregating the opinions of various managers.

Most existing association rule mining algorithms take the measure of large support to find frequent itemsets, and all items are considered to have equal weight (Tao et al., 2003). Therefore, these approaches are unsuitable for discovering the interesting infrequent itemsets as described in the above two examples. Tao et al. developed an approach that used an improved model of weighted support. In the approach of weighted association rule mining, itemsets are no longer simply counted as they appear in a transaction, and the subjective measures (e.g., profit) are also adopted for rule evaluation.

Most of the abovementioned approaches focus on computation efficiency by embedding the subjective constraints in the mining procedure to prune the search space. However, a huge amount of subjective domain knowledge may exist, which can be considered as potential subjective constraints and interestingness measures. It is sophisticated to determine the subjective constraints and interestingness measures before discovering some rules. Provided that the constraints are not adequately stated, the interesting rules may not be discovered after the mining procedure. Additionally, rule interestingness may be a relative measure, but not an absolute one. Generally, decision makers can suitably select interesting rules for implementation after making comparisons between some potential rules.

In data mining, it is substantial to bring together the statistic based rule extraction and profit based action to meet the enterprises' objectives (Wang, Zhou, & Han, 2002). This paper aims at using a non-parametric approach, Data Envelopment Analysis (DEA), to estimate and rank the efficiency (interestingness or usefulness) of association rules with multiple criteria. The interestingness of association rules is measured by multiple criteria involving support, confidence and domain related measures. This paper uses DEA as a post-processing approach. After the rules have been discovered from the association rule mining algorithms, DEA is used to rank those discovered rules based on the specified criteria. The remainder of this paper is organized as follows. Section 2 introduces the concept of association rules. Section 3 then presents the DEA method.

Furthermore, the proposed approach is described in Section 4. Next, an example of market basket analysis is illustrated in Section 5. Finally, Section 6 makes a conclusion to this paper.

2. Association rules

Association rule mining discovers the relationships between items from the set of transactions. These relationships can be expressed by association rules such as [$i_1 \Rightarrow i_2$, i_3 support = 3.5%, confidence = 45%]. This association rule implies that 3.5% of all the transactions under analysis show that items i_1, i_2 and i_3 appear jointly. A confidence of 45% indicates that 45% of the transactions containing i_1 also contain i_2 and i_3. Associations may include any number of items on either side of the rule.

The problem of mining association rules is formally stated as follows (Agrawal et al., 1993; Srikant & Agrawal, 1997). Let $I = \{i_1, i_2, \ldots, i_m\}$ denote a set of literals, namely items. Moreover, let **D** represent a set of transactions, where each transaction T is a set of items such that $T \subset$ I. A unique identifier, namely TID, is associated with each transaction. A transaction T is said to contain **X**, a set of some items in I, if $\mathbf{X} \subseteq T$. An association rule implies the form $\mathbf{X} \Rightarrow \mathbf{Y}$, where $\mathbf{X} \subset I$, $\mathbf{Y} \subset I$ and $\mathbf{X} \cap \mathbf{Y} = \emptyset$. The rule $\mathbf{X} \Rightarrow \mathbf{Y}$ holds in the transaction set **D** with *confidence*, c, where c% of transactions in **D** that contain **X** also contain **Y**. The rule has *support*, s, in the transaction set **D** if s% of transactions in **D** contain $\mathbf{X} \cup \mathbf{Y}$. An efficient algorithm is required that restricts the search space and checks only a subset of all association rules, yet does not miss important rules. The Apriori algorithm developed by Agrawal et al. (1993) and Srikant and Agrawal (1997) is such an algorithm. However, the interestingness of rule is only based on support and confidence. The Apriori algorithm is described as follows:

(1) $\mathbf{L}_1$ = find_large_1-itemsets;
(2) **for** ($k = 2$; $\mathbf{L}_{k-1} \neq \emptyset$; k++) **do begin**
(3) $\mathbf{C}_k$ = **apriori_gen** ($\mathbf{L}_{k-1}$); // new candidates
(4) **forall** TID $T \in$ **D do begin**
(5) $\mathbf{C}_T$ = subset ($\mathbf{C}_k$, T); // candidates contained in T
(6) **forall** candidates $C \in \mathbf{C}_T$ **do**
(7) C.*count*++;
(8) **end**
(9) $\mathbf{L}_k = \{C \in \mathbf{C}_k |$ C.*count*/*no_of_data* $\geqslant$ minimum support threshold$\}$
(10) **end**
(11) **Return** $L = \cup_k \mathbf{L}_k$.

In the above Apriori algorithm, the **apriori_gen** procedure generates candidates of itemset and then uses the minimum support criterion to eliminate infrequent itemsets. The **apriori_gen** procedure performs two actions, namely, **join** and **prune**, which are discussed in Han and Kamber (2001). In **join** step, $\mathbf{L}_{k-1}$ is joined with $\mathbf{L}_{k-1}$ to generate potential candidates of itemset. The **prune** step uses the minimum support criterion to remove candidates of itemset that are not frequent. In fact, expanding an itemset reduces its support. A k-itemset can only be frequent if all of its $(k-1)$-subsets are also frequent, consequently **apriori_gen** only generates candidates with this property, a situation easily achievable given the set $\mathbf{L}_{k-1}$.

Association rule mining is a popular technique for *market basket analysis*, which typically aims at finding buying patterns for supermarket, mail-order and other customers. By mining association rules, marketing analysts try to find sets of products that are frequently bought together, so that certain other items can be inferred from a shopping cart containing particular items. Association rules can often be used to design marketing promotions, for example, by appropriately arranging products on a supermarket shelf and by directly suggesting to customers items that may be of interest.

With the constant collection and storage of considerable quantities of business data, association rules are discovered from the domain databases and applied in many areas, such as marketing, logistics and manufacturing (Chen, 2003; Chen, Chiu, & Chang, 2005a; Chen, Huang, Chen, & Wu, 2005b; Chen & Wu, 2005; Chen & Lin, in press; Wang et al., 2004). In the areas of marketing, advertising and sales, corporations have found they can benefit enormously if implicit and previously unknown customer buying and calling patterns can be discovered from large volumes of business data (Chen, Han, & Yu, 1996).

Generally, support and confidence are taken as two measures to evaluate the interestingness of association rules (Agrawal et al., 1993; Srikant & Agrawal, 1997). Association rules are regarded as interesting if their support and confidence are greater than the user-specified minimum support and minimum confidence, respectively. In data mining, it is important but difficult to appropriately determine these two thresholds of interestingness. Data miners usually specify these thresholds in an arbitrary manner. Numerous algorithms for finding association rules have been developed in previous studies (Hipp, Günter, & Nakhaeizadeh, 2000). However, relatively little literature has attempted to employ the application-specific criteria for setting the threshold of association rules.

3. Data envelopment analysis

In 1978, Data Envelopment Analysis (DEA) was initiated by Charnes, Cooper and Rhodes (CCR), and they demonstrated how to change a fractional linear measure of efficiency into a linear programming model (Charnes, Cooper, & Rhodes, 1978). DEA was defined by Charnes et al. as: a mathematical programming model applied to observational data, which provides a new method of obtaining empirical estimates of extremal relations-such as the production functions and/or efficient production possibility surfaces that are fundamental to modern economics. Researchers have developed several DEA models by theoretically broadening the CCR model (e.g., Charnes,

Cooper, Seiford, & Stutz, 1982; Charnes, Cooper, Golany, Seiford, & Stutz, 1985; Cook & Kress, 1990; Hashimoto, 1997; Obata & Ishii, 2003).

DEA was originally designed to mathematically measure decision making units (DMUs) with multiple inputs and outputs in terms of relative efficiency (i.e., the ratio of total weighted output to total weighted input). However, no obvious production function exists for aggregating the data in its entirety (Adler, Friedman, & Sinuany-Stern, 2002). Cook and Kress (1990) introduced a theoretical extension of DEA to analyze ranked voting data. In the ranked voting system, each candidate (DMU) is regarded as having multiple outputs (ranked votes) and only input with amount unity, i.e., the pure output DEA model (Hashimoto, 1997). In the approach developed by Cook and Kress, preference scores are estimated without initially imposing any fixed weights. The score of each candidate (DMU score) is calculated based on its most favorable weights (Obata & Ishii, 2003). Adler et al. (2002) reviewed some ranking methods in DEA.

The preference score, Z_i, of candidate (DMU) i is the weighted sum of votes with certain weights. The mathematical model of the ranked voting system in DEA is formulated as follows (Cook & Kress, 1990):

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

Subject to:

$$\sum_{j=1}^{k} w_j v_{ij} \leqslant 1, \quad i = 1, 2, \ldots, m; \tag{2}$$

$$w_j - w_{j+1} \geqslant d(j, \varepsilon), \quad j = 1, 2, \ldots, k-1; \tag{3}$$

$$w_j \geqslant d(k, \varepsilon), \tag{4}$$

where w_j denotes the weight of the jth place; v_{ij} represents the number of jth place votes of candidate i ($i = 1, 2, \ldots, m, j = 1, 2, \ldots, k$); and $d(\bullet, \varepsilon)$, known as the discrimination intensity function, is nonnegative and nondecreasing in ε and satisfies $d(\bullet, 0) = 0$. Parameter ε is nonnegative.

The above mathematical model is resolved for each candidate o, $o = 1, 2, \ldots, m$. The resulting objective value is the preference score of candidate o. Constraints (3) ensure that the vote of the higher place may have a greater importance than that of the lower place. Several candidates may achieve a maximum preference score of 1 once the linear programs (1)–(4) are resolved for all candidates. Candidates with preference score 1 are called *efficient candidates*. Without setting the priorities of criteria, Constraints (3) are relaxed.

DEA frequently generates several efficient candidates (Obata & Ishii, 2003). The set of efficient candidates is the top group of DMUs, but no efficient DMU can be distinguished as the winner among this group. It is necessary to further discriminate these efficient candidates. To discriminate efficient candidates, the discriminant method proposed by Obata and Ishii (2003) is adopted in this paper. The discriminant model for efficient candidates is formulated as follows (Obata & Ishii, 2003):

$$\text{Minimize} \quad \sum_{j=1}^{k} w_j \tag{5}$$

Subject to:

$$\sum_{j=1}^{k} w_j v_{oj} = 1; \tag{6}$$

$$\sum_{j=1}^{k} w_j v_{ij} \leqslant 1, \quad \text{for all efficient } i \neq o; \tag{7}$$

$$w_j - w_{j+1} \geqslant d(j, \varepsilon), \quad j = 1, 2, \ldots, k-1; \tag{8}$$

$$w_j \geqslant d(k, \varepsilon). \tag{9}$$

Obata and Ishii's discriminant model, (5)–(9), is also a linear program. The preference score of the second stage (Z'_i) is obtained as a reciprocal of the optimal value, that is $Z'_i = 1/\sum_{j=1}^{k} w_j$. This discriminant model does not employ any information about inefficient candidates, and thus the problem of varying the rank of efficient candidates does not occur (Obata & Ishii, 2003). Similar to Cook and Kress's DEA model, (1)–(4) Constraints (8) are relaxed if the priorities of criteria are not specified.

4. Proposed post-processing approach

The interestingness of a rule can be use to filter a large number of rules and report only those which may be useful to decision makers (Mitra, Pal, & Mitra, 2002). The thresholds of support and confidence are selected by only considering the database perspective. However, the interestingness of an association rule is commonly application-dependent (Srikant et al., 1997). The domain information in application areas can potentially provide useful criteria for picking important rules, and can be adopted to improve the rule selection procedure.

Association rule mining contains no theoretically optimal thresholds to filter rules and patterns. The idea of efficiency in DEA is a comparative concept (Serrano-Cinca, Fuertes-Calle'n, & Mar-Molinero, 2005). As discussed above, the idea of interestingness resembles the comparative concept of efficiency, and no absolute measure exists. In DEA, the set of efficient candidates is taken as the leading set of candidates. All candidates (association rules) can be ranked in decreasing order of preference score, and interesting rules then can be selected by setting a threshold of preference score or selecting rules with first N highest preference scores. This paper further adopts Obata and Ishii's discriminant model, (5)–(9), is further adopted to discriminate the efficient candidates (i.e., the interesting association rules), which are generated in Cook and Kress's DEA model, (1)–(4). The discriminant model can be used to identify one winner (the most favorable association rule) if such a winner exists.

With the above discussion, the proposed post-processing approach for finding the interesting association rule is schematically illustrated in the flow chart in Fig. 1. The proposed approach is described as follows:

Step 1. Input data for association rule mining.
Step 2. Mine association rules by using the Apriori algorithm with minimum support and minimum confidence.
Step 3. Determine subjective interestingness measures by further considering the domain related knowledge.
Step 4. Calculate the preference scores of association rules discovered in Step 2 by using Cook and Kress's DEA model, (1)–(4).
Step 5. Discriminate the efficient association rules found in Step 3 by using Obata and Ishii's discriminant model, (5)–(9).
Step 6. Select rules for implementation by considering the reference scores generated in Step 5 and domain related knowledge.

In Step 2 of the proposed post-processing approach, the Apriori algorithm (Agrawal et al., 1993; Srikant & Agrawal, 1997) is initially used to discover the association rules only with the thresholds of support and confidence. These two thresholds can be set relatively lower since this paper intends to find some infrequent itemsets, which may be of interest to analysts by considering subjective domain related knowledge. The interestingness of association rules generated in Step 2 can be further judged by the following DEA approach in Steps 4 and 5.

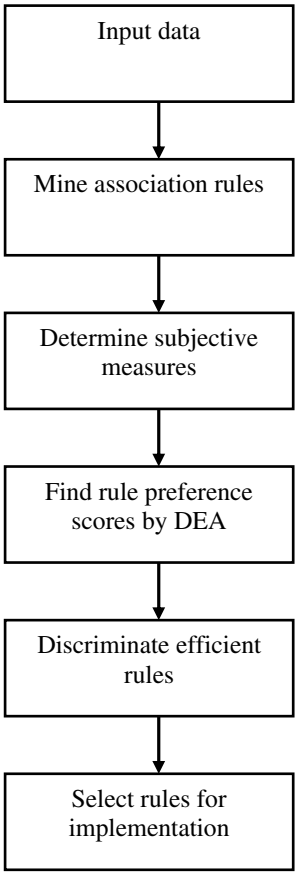

Fig. 1. Flow chart of the proposed post-processing approach.

The traditional Apriori algorithm cannot classify the infrequent items to interesting itemsets since the subjective domain knowledge is ignored. A huge amount of subjective domain knowledge may exist, which can be considered as potential subjective constraints and measures for evaluating association rules. Following the discovery and reporting of some rules, a data miner can select the subjective interestingness measures in Step 3. In market basket analysis, understanding which products are usually bought together by customers and which products are beneficial to sellers are both interesting subjects for marketing analysts. The former can be measured in terms of support and confidence in association rules. In this paper, the subjective measures of sellers' profits are evaluated in terms of itemset value and cross-selling profit corresponding to the association rules. For association rules like $\mathbf{X} \Rightarrow \mathbf{Y}$, four criteria are jointly used for rule evaluation as follows:

Support: The support, s, is the percentage of transactions that contain $\mathbf{X} \cup \mathbf{Y}$ (Agrawal et al., 1993). It takes the form

$$s = P(\mathbf{X} \cup \mathbf{Y}) \tag{10}$$

Confidence: The confidence, c, is the ratio of the percentage of transactions that contain $\mathbf{X} \cup \mathbf{Y}$ to the percentage of transactions that contain $\mathbf{X}$ (Agrawal et al., 1993). It takes the form

$$c = \frac{P(\mathbf{X} \cup \mathbf{Y})}{P(\mathbf{X})} = P(\mathbf{Y}|\mathbf{X}) \tag{11}$$

Itemset value: The itemset value, v, is the sum of values of all items (v_g's) in the itemset, $\mathbf{X} \cup \mathbf{Y}$, and can be calculated by

$$v = \sum_{g \in \mathbf{X} \cup \mathbf{Y}} v_g \tag{12}$$

Cross-selling profit: For the rule $\mathbf{X} \Rightarrow \mathbf{Y}$, the cross-selling is described as recommending that customers purchase $\mathbf{Y}$, if they have bought $\mathbf{X}$. Therefore, the cross-selling profit is the sum of the profits of all items (p_h's) in $\mathbf{Y}$, and can be calculated by

$$p = \sum_{h \in \mathbf{Y}} p_h \tag{13}$$

In the above four measures, support and confidence are generated by the Apriori algorithm. Meanwhile, the other two measures of itemset value and cross-selling profit are subjectively selected by analysts. In Step 4, these four measures are used as criteria in Cook and Kress's DEA model, (1)–(4), to calculate the preference scores for all association rules generated in Step 2. Provided that there exist more than one efficient association rules with preference score 1, the algorithm proceeds to Step 5 to further discriminate

these rules by using Obata and Ishii's discriminant model, (5)–(9). Finally, the rankings of all rules can be obtained, and analysts accordingly select useful rules for implementation.

5. Illustrative example

An example of market basket data is used to illustrate the proposed approach presented in Section 4. Association rules first are discovered by the Apriori algorithm, in which minimum support and minimum confidence are set to 1.0% and 10.0%, respectively. Forty-six rules then are identified in Step 2. The itemset values and cross-selling profits for these 46 rules are then calculated, as summarized in Table 1.

Table 1
Summary of results of Cook and Kress's model

Rule no.	Support (%)	Confidence (%)	Itemset value	Cross-selling profit	Preference score (Z_i)
1	**3.87**	**40.09**	**337.00**	**25.66**	**1.00**
2	1.42	18.17	501.00	11.63	0.78
3	2.83	17.64	345.00	11.29	0.84
4	2.34	30.83	163.00	19.73	0.71
5	2.63	23.90	325.00	15.30	0.78
6	**1.19**	**55.65**	**436.00**	**35.61**	**1.00**
7	**1.19**	**47.42**	**598.00**	**30.35**	**1.00**
8	1.19	15.70	436.00	52.91	0.69
9	1.19	10.82	598.00	36.45	0.85
10	1.19	12.32	436.00	20.08	0.67
11	1.19	12.32	598.00	40.04	0.85
12	**3.87**	**38.08**	**337.00**	**103.97**	**1.00**
13	1.18	15.09	710.00	41.19	0.99
14	**2.44**	**15.22**	**554.00**	**41.56**	**1.00**
15	2.14	28.21	372.00	77.02	0.78
16	2.51	22.81	534.00	62.26	0.99
17	**1.19**	**50.92**	**436.00**	**139.02**	**1.00**
18	**1.19**	**45.25**	**598.00**	**123.52**	**1.00**
19	1.19	11.70	436.00	43.54	0.67
20	1.19	11.70	598.00	62.50	0.88
21	1.42	13.99	501.00	61.16	0.79
22	**1.18**	**12.23**	**710.00**	**53.45**	**1.00**
23	**1.50**	**13.64**	**698.00**	**59.59**	**1.00**
24	2.83	27.82	345.00	78.17	0.84
25	2.44	25.27	554.00	71.00	1.00
26	1.25	15.97	718.00	44.87	1.00
27	1.22	34.89	339.00	98.04	0.75
28	1.30	35.12	435.00	98.68	0.81
29	1.42	33.81	534.00	95.01	0.90
30	1.91	25.26	380.00	70.97	0.75
31	**1.43**	**37.14**	**618.00**	**104.35**	**1.00**
32	2.38	21.63	542.00	60.78	0.98
33	1.18	30.24	366.00	84.98	0.70
34	1.23	29.36	626.00	82.51	0.96
35	1.58	22.65	354.00	63.64	0.67
36	2.34	22.99	163.00	22.76	0.60
37	2.14	22.14	372.00	21.92	0.75
38	1.91	11.94	380.00	11.82	0.72
39	2.03	18.42	360.00	18.23	0.72
40	1.19	30.73	436.00	30.43	0.75
41	2.63	25.87	325.00	67.52	0.78
42	2.51	25.98	534.00	67.81	0.99
43	**1.50**	**19.16**	**698.00**	**50.02**	**1.00**
44	2.38	14.85	542.00	38.75	0.98
45	2.03	26.73	360.00	69.78	0.75
46	1.19	30.73	598.00	80.22	0.93

Table 2
Summary of results of Obata and Ishii's model

Rule no.	Support (%)	Confidence (%)	Itemset value	Cross-selling profit	Preference score (Z'_i)
26	1.25	15.97	718.00	44.87	718.00
22	1.18	12.23	710.00	53.45	393.23
18	1.19	45.25	598.00	123.52	306.12
17	1.19	50.92	436.00	139.02	164.95
7	1.19	47.42	598.00	30.35	2.04
23	1.50	13.64	698.00	59.59	1.17
6	1.19	55.65	436.00	35.61	0.79
43	1.50	19.16	698.00	50.02	0.26
31	1.43	37.14	618.00	104.35	0.16
12	3.87	38.08	337.00	103.97	0.12
1	3.87	40.09	337.00	25.66	0.10

Preference scores (Z_i) for each rule listed in Table 1 are calculated by Cook and Kress's DEA model. In this paper, $d(\bullet,\varepsilon) = 0$ is adopted. Additionally, the priorities of criteria are not specified. Table 1 reveals 11 efficient association rules with preference score 1. These 11 efficient rules are further analyzed using Obata and Ishii's discriminant model. According to the preference scores (Z'_i) obtained by solving Obata and Ishii's discriminant model, these 11 rules are ranked in decreasing order in Table 2.

The interestingness of a rule is measured by two types of preference scores (Z_i and Z'_i), which are essential measures for filtering a number of rules and report only those which are most interesting to decision makers. The criteria of support and confidence only consider the database perspective. According to these two criteria, marketing analysts are most likely to pick rules 1 and/or 12 to design their promotion campaigns. However, with the above discussion, the interestingness of an association rule is generally application-dependent, and the domain information in application areas can potentially provide useful criteria for picking important rules.

For example, this paper also considers two subjective measures of itemset value and cross-selling profit. Observing the results of Cook and Kress's DEA model and Obata and Ishii's discriminant model based on four criteria, analysts likely select rules 26, 22 and/or 18 to design the marketing activities. From Tables 1 and 2, rules 1 and 12 are also identified as efficient candidates (interesting rules), while these two rules have relatively lower preference scores in Obata and Ishii's discriminant model.

6. Conclusion

Association rule discovery is one of the popular techniques recently developed in the area of data mining. Evaluating the interestingness or usefulness of association rules is an essential task in data mining applications. In market basket analysis, marketing analysts are no longer satisfied by a set of rules or patterns discovered by a data

mining algorithm. Instead, marketing analysts wish to develop rules or patterns that are ranked with respect to certain criteria. The complexity of rule evaluation and selection is difficult for analysts. The traditional approaches usually ignore the subjective domain knowledge in selecting useful rules. To meet the requirements of marketing analysts, Data Envelopment Analysis (DEA) is used in this paper to evaluate the efficiency (interestingness or usefulness) of association rules with multiple criteria, including subjective domain related measures. The proposed approach provides more insights into the rules discovered and can assist rule evaluation and selection.

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

Adler, N., Friedman, L., & Sinuany-Stern, Z. (2002). Review of ranking methods in the data envelopment analysis context. *European Journal of Operational Research, 140*(2), 249–265.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD Conference on Management of Data*, 254–259.

Charnes, A., Cooper, W. W., Golany, B., Seiford, L., & Stutz, J. (1985). Foundations of data envelopment analysis for Pareto–Koopmans efficient empirical production functions. *Journal of Econometrics, 30*(1–2), 91–107.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*(6), 429–444.

Charnes, A., Cooper, W. W., Seiford, L., & Stutz, J. (1982). A multiplicative model for efficiency analysis. *Socio-Economic Planning Sciences, 6*(3), 223–224.

Chen, M.-C. (2003). Configuration of cellular manufacturing systems using association rule induction. *International Journal of Production Research, 41*(2), 381–395.

Chen, M.-C., Chiu, A.-L., & Chang, H.-H. (2005a). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications, 28*(4), 773–781.

Chen, M.-C., Huang, C.-L., Chen, K.-Y., & Wu, H.-P. (2005b). Aggregation of orders in distribution centers using data mining. *Expert Systems with Applications, 28*(3), 453–460.

Chen, M.-C., & Lin, C.-P. (in press). A data mining approach to product assortment and shelf space allocation, *Expert Systems with Applications*, doi:10.1016/j.eswa.2006.02.001.

Chen, M.-C., & Wu, H.-P. (2005). An association-based clustering approach to order batching considering customer demand patterns. *Omega-International Journal of Management Science, 33*(4), 333–343.

Chen, M.-S., Han, J., & Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 866–883.

Choi, D. H., Ahn, B. S., & Kim, S. H. (2005). Prioritization of association rules in data mining: multiple criteria decision approach. *Expert Systems with Applications, 29*(4), 876–878.

Cohen, E., Datar, M., Fujiwara, S., Gionis, A., Indyk, R., Motwani, P., Ullman, J., & Yang, C. (2000). Finding interesting associations without support pruning. In *Proceedings of the 16th international conference on data engineering*, pp. 489–500.

Cook, W. D., & Kress, M. (1990). A data envelopment model for aggregating preference rankings. *Management Science, 36*(11), 1302–1310.

Goethals, B., & Van den Bussche, J. (2000). On supporting interactive association rule mining. *Lecture Notes in Computer Science, 1874*, 307–316.

Han, J., & Kamber, M. (2001). *Data mining: concepts and techniques*. Morgan Kaufmann: San Francisco.

Hashimoto, A. (1997). A ranked voting system using a DEA/AR exclusion model: a note. *European Journal of Operational Research, 97*, 600–604.

Hipp, J., Günter, U., & Nakhaeizadeh, G. (2000). Algorithms for association rule mining-a general survey and comparison. *ACM SIGKDD Explorations Newsletter, 2*(1), 58–64.

Klemetinen M, Mannila H, Ronkainen P, Toivonen H, & Verkamo AI (1994). Finding interesting rules from large sets of discovered association rules. In *Proceedings of the third international conference on information and knowledge management*, pp. 401–407.

Liu, B., & Hsu, W. (1996). Post-analysis of learned rules. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, AAAI-96, 828–834.

Liu, B., Hsu, W., Chen, S., & Ma, Y. (2000). Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems, 15*(5), 47–55.

Mitra, S., Pal, S. K., & Mitra, P. (2002). Data mining in soft computing framework: a survey. *IEEE Transactions on Neural Networks, 13*(1), 3–14.

Ng. R. T., Lakshmanan, L. V. S., Han, J., & Pang, A. (1998). Exploratory mining and pruning optimizations of constrained association rules. In *Proceedings of the ACM SIGMOD international conference on management of data, SIGMOD-98*, pp. 13–24.

Obata, T., & Ishii, H. (2003). A method for discriminating efficient candidates with ranked voting data. *European Journal of Operational Research, 151*(1), 233–237.

Serrano-Cinca, C., Fuertes-Calle'n, Y., & Mar-Molinero, C. (2005). Measuring DEA efficiency in internet companies. *Decision Support Systems, 38*, 557–573.

Silberschatz, A., & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 970–974.

Srikant, R., & Agrawal, R. (1997). Mining generalized association rules. *Future Generation Computer Systems, 13*, 161–180.

Srikant, R., Vu, Q., & Agrawal, R. (1997). Mining association rules with item constraints. In *Proceedings of the third international conference on knowledge discovery and data mining, KDD-97*, pp. 67–73.

Tan, P. N., & Kumar, V. (2000). Interestingness measures for association patterns: A perspective, KDD 2000 workshop on postprocessing in machine learning and data mining, Boston, MA, August.

Tao, F., Murtagh, F., & Farid, M. (2003). Weighted association rule mining using weighted support and significance framework. In *Proceedings of the ACM SIGMOD international conference on management of data, Sigmod-03*, pp. 661–666.

Wang, K., Zhou, S., & Han, J. (2002). Profit mining: from patterns to actions. *Proceedings of International Conference on Extending Data Base Technology*, 70–87.

Wang, Y.-F., Chuang, Y.-L., Hsu, M.-H., & Keh, H.-C. (2004). A personalized recommender system for the cosmetic business. *Expert Systems with Applications, 26*(3), 427–434.

Webb, G. I., & Zhang, S. (2005). K-optimal rule discovery. *Data Mining and Knowledge Discovery, 10*(1), 39–79.