23

On Ranking Discovered Rules of Data Mining by Data Envelopment Analysis: Some New Models with Applications

Mehdi Toloo[1] and Soroosh Nalchigar[2]
[1]*Department of Mathematics, Islamic Azad University of Central Tehran Branch, Tehran*
[2]*University of Pierre and Marie Curie, Paris*
[1]*Iran*
[2]*France*

1. Introduction

The convergence of computing and communication has resulted in a society that feeds on information. There is exponentially increasing huge amount of information locked up in databases—information that is potentially important but has not yet been discovered or articulated (Whitten & Frank, 2005). Data mining, the extraction of implicit, previously unknown, and potentially useful information from data, can be viewed as a result of the natural evolution of Information Technology (IT). An evolutionary path has been passed in database field from data collection and database creation to data management, data analysis and understanding. According to Han & Camber (2001) the major reason that data mining has attracted a great deal of attention in information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis, to engineering design and science exploration. In other words, in today's business environment, it is essential to mine vast volumes of data for extracting patterns in order to support superior decision-making. Therefore, the importance of data mining is becoming increasingly obvious. Many data mining techniques have also been presented in various applications, such as association rule mining, sequential pattern mining, classification, clustering, and other statistical methods (Chen & Weng, 2008).

Association rule mining is a widely recognized data mining method that determines consumer purchasing patterns in transaction databases. Many applications have used association rule mining techniques to discover useful information, including market basket analysis, product recommendation, web page pre-fetch, gene regulation pathways identification, medical record analysis, and so on (Chen & Weng, 2009).

Extracting association rules has received considerable research attention and there are several efficient algorithms that cope with popular and computationally expensive task of association rule mining (Hipp et al., 2000). Using these algorithms, various rules may be obtained and only a small number of these rules may be selected for implementation due, at

least in part, to limitations of budget and resources (Chen, 2007). According to Liu et al. (2000) the interestingness issue has long been identified as an important problem in data mining. It refers to finding rules that are interesting/useful to the user, not just any possible rule. Indeed, there exist some situations that make necessary the prioritization of rules for selecting and concentrating on more valuable rules due to the number of qualified rules (Tan & Kumar, 2000) and limited business resources (Choi et al., 2005). The purpose of this chapter, briefly, is to propose a new methodology for prioritizing association rules resulted from the data mining, while considering their business values incorporating the conflicting criteria of business values. Toward this end, a decision analysis method, Data Envelopment Analysis (DEA) is applied.

Recent years have seen a great variety of applications of DEA for use in evaluating the performances of many different kinds of entities engaged in many different activities in many different contexts in many different countries (Cooper et al., 2007). Selection of best vendors (Weber et al., 1998 & Liu et al., 2000), ranking data mining rules (Chen, 2007 & Toloo et al., 2009), evaluation of data warehouse operations (Mannino et al., 2008), selection of flexible manufacturing system (Liu, 2008), assessment of bank branch performance (Camanho & Dyson, 2005), examining bank efficiency (Chen et al., 2005), analyzing firm's financial statements (Edirisinghe & Zhan, 2007), measuring the efficiency of higher education institutions (Johnes, 2006), solving Facility Layout Design (FLD) problem (Ertay et al., 2006) and measuring the efficiency of organizational investments in information technology (Shafer & Byrd, 2000) are samples of using DEA in various areas.

The rest of this chapter is organized as follows: Section 2 provides readers with basic concepts of DEA. Moreover, this section reviews DEA models for finding most efficient DMUs. Section 3 describes data mining association rules, their applications and algorithm. In Section 4, the problem which is addressed by this chapter is expressed. Section 5 provides a review of related studies for solving the problem. Section 6 presents a new methodology for the problem of chapter. Section 7 shows applicability of proposed method. Finally, this chapter closes with some concluding remark in Section 8.

2. Data envelopment analysis

2.1 Basic models

Data envelopment analysis (DEA) is a mathematical optimization technique that measures the relative efficiency of decision making units (DMUs) with multiple input–output. Based on Farrell's pioneering work, Charnes et al. (1978) first proposed DEA as an evaluation tool to measure and compare a DMU's relative efficiency. During last three decades, DEA has been widely recognized and discussed from the methodological as well as practical side in measuring the relative efficiency of units that utilize the same inputs to produce the same outputs. One advantage of DEA is that these inputs and outputs can remain in their natural physical units without reducing or transforming them into some common metric such as dollars. Indeed, DEA defines relative efficiency as the ratio of the sum of weighted outputs to the sum of weighted inputs:

$$\text{DEA efficiency} = \frac{\text{Sum of weighted outputs}}{\text{Sum of weighted inputs}}$$

The more output produced for a given amount of resources, the more efficient is the unit. The problem is how to weight each of the individual input and output variables, expressed

in their natural units; solving for these weights is the fundamental essence of DEA. For each DMU, the DEA procedure finds the set of weights that makes the efficiency of that DMU as large as possible. The values the weights any DMU can obtain is restricted through the evaluation of those weights in the input/output vectors for all the other comparable DMUs, where the resultant ratio of the sum of weighted outputs to the sum of weighted inputs is constrained to be no larger than 1. The procedure is repeated for all other DMUs to obtain their weights and associated relative efficiency score; ultimately providing decision makers with a listing of comparable DMUs ranked by their relative efficiencies.

Assume that there are n DMUs, ($\mathrm{DMU}_j : j = 1,2,...,n$). Some common input and output items for each of these n DMUs are selected as follows (Cooper et al., 2007):

1. Numerical data are available for each input and output, with the data assumed to be positive for all DMUs.
2. The items (inputs, outputs and choice of DMUs) should reflect an analyst's or a manager's interest in the components that will enter into the relative efficiency evaluations of the DMUs.
3. In principle, smaller input amounts are preferable and larger output amounts are preferable so the efficiency scores should reflect these principles.
4. The measurement units of the different inputs and outputs need not be congruent. Some may involve number of persons, or areas of floor space, money expended, etc.

Suppose each DMU consume m inputs ($\mathbf{x}_i : i = 1,2,\ldots,m$) to produce s outputs ($\mathbf{y}_r : r = 1,2,\ldots,s$). The CCR input oriented (CCR-I) model (Charnes et al., 1978) evaluates the efficiency of DMU_o, DMU under consideration, by solving the following linear program:

$$\begin{aligned} &\max \sum_{r=1}^{s} u_r y_{rj} \\ &\text{s.t.} \\ &\quad \sum_{i=1}^{m} w_i x_{io} = 1 \\ &\quad \sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} w_i x_{ij} \le 0 \qquad j = 1,2,...,n \\ &\quad w_i \ge \varepsilon \qquad i = 1,2,...,m \\ &\quad u_r \ge \varepsilon \qquad r = 1,2,...,s \end{aligned} \tag{1}$$

where x_{ij} and y_{rj} (all nonnegative) are the inputs and outputs of the DMUj, w_i and u_r are the input and output weights (also referred to as multipliers). x_{io} and y_{ro} are the inputs and outputs of DMU_o. Also, ε is non-Archimedean infinitesimal value for forestalling weights to be equal to zero. To find a suitable value for ε, there exists a polynomial time algorithm, Epsilon algorithm, which introduced by Amin & Toloo (2004). The CCR-I model must be run n times, once for each unit, to get the relative efficiency of all DMUs.

It should be noted that Model (1) assumes that the production function exhibits constant returns-to-scale. As a theoretical extension, Banker et al. (1984) developed a variable returns to scale variation of Model (1). The BCC model (Banker et al., 1984) adds an additional constant variable in order to permit variable returns-to-scale. The BCC input oriented (BCC-I) model evaluates the efficiency of DMU_o, DMU under consideration, by solving the following linear program:

$$
\begin{aligned}
&\max \sum_{r=1}^{s} u_r y_{rj} - u_0 \\
&\text{s.t.} \\
&\quad \sum_{i=1}^{m} w_i x_{io} = 1 \\
&\quad \sum_{r=1}^{s} u_r y_{rj} - u_0 - \sum_{i=1}^{m} w_i x_{ij} \le 0 \qquad j = 1,2,...,n \\
&\quad w_i \ge \varepsilon \qquad i = 1,2,...,m \\
&\quad u_r \ge \varepsilon \qquad r = 1,2,...,s \\
&\quad u_0 \ \text{free}
\end{aligned}
\tag{2}
$$

The structure and variables of this model are similar to Model (1). It is clear that a difference between the CCR and BCC models is present in the free variable u_0, which is used to measure the return to scale of DMU_o.

New applications and extensions with more variables and more complicated models are being introduced (Emrouznejad et al., 2007). In many applications of DEA, finding the most efficient DMU is desirable. The next section of this chapter introduces readers with some new DEA model for finding the most efficient DUMS. It is noteworthy to mention that Cook & Seiford (2009) provide a sketch of some of the major research thrusts in DEA over the three decades. Interested readers can refer to this paper of for further discussion on DEA, and a comprehensive review on it.

2.2 DEA model for finding the most efficient DMU

By applying basic DEA models (CCR and BCC), DMUs are grouped into two sets: efficient and inefficient DMUs. On the other hand, often decision-makers are interested in a complete ranking, beyond the dichotomized classification, in order to refine the evaluation of the units and find most efficient DMUs. Recently, the problem of finding most efficient DMUs in DEA has gained attention between researchers. For instance Ertay et al. (2006) integrated DEA and Analytic Hierarchy Process (AHP) and presented a decision-making methodology for evaluating Facility Layout Designes (FLDs). In the last step of their methodology, they extended minimax DEA model to identify single most efficient DMU. Amin & Toloo (2007) extended their work and proposed an integrated DEA model in order to detect the most CCR-efficient DMU. It was able to find the most CCR-efficient DMU without solving the model n times (one Linear Programming (LP) for each DMU) and therefore allowed the user to get faster results. Amin & Toloo (2007)'s model is as follows:

$$
\begin{aligned}
&\min M \\
&\text{s.t.} \\
&\quad M - d_j \ge 0 \qquad j = 1,2,...,n \\
&\quad \sum_{i=1}^{m} w_i x_{ij} \le 1 \qquad j = 1,2,...,n \\
&\quad \sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} w_i x_{ij} + d_j - \beta_j = 0 \qquad j = 1,2,...,n
\end{aligned}
\tag{3}
$$

$$\begin{aligned}
&\sum_{j=1}^{n} d_j = n-1 \\
&0 \le \beta_j \le 1 \qquad j=1,2,...,n \\
&d_j \in \{0,1\} \qquad j=1,2,...,n \\
&w_i \ge \varepsilon \qquad i=1,2,...,m \\
&u_r \ge \varepsilon \qquad r=1,2,...,s \\
&M \ \text{free}
\end{aligned} \tag{3}$$

where d_j as a binary variable represents the deviation variable of DMU_j. β_j is considered in the Model (3) because of discrete nature of d_j and M represents maximum inefficiency which should be minimized. DMU_j is most efficient if and only if $d_j^*=0$.

First constraint of Model (3) implies that M is equal to maximum inefficiency. Second constraint shows input-oriented nature of the Model (2). Third constraint causes efficiency of all units to be less that 1. The last one implies among all the DMUs for only most efficient unit, say DMU_p, which has $d_p^*=0$ in any optimal solution. In addition, to determine the non-Archimedean epsilon, Amin & Toloo (2007) developed an epsilon model.

It should be noted that Model (3) is based on CCR model and identify most CCR-efficient DMU. Indeed, Model (3) is not applicable for situations in which DMUs operating in variable return to scale. To overcome this drawback, Toloo & Nalchigar (2009) proposed an integrated model which is able to find most BCC-efficient DMU. They developed Model (3) as a new integrated model for finding the most BCC-efficient DMU.

$$\begin{aligned}
&\min M \\
&\text{s.t.} \\
&M - d_j \ge 0 \qquad j=1,2,...,n \\
&\sum_{i=1}^{m} w_i x_{ij} \le 1 \qquad j=1,2,...,n \\
&\sum_{r=1}^{s} u_r y_{rj} - u_0 - \sum_{i=1}^{m} w_i x_{ij} + d_j - \beta_j = 0 \qquad j=1,2,...,n \\
&\sum_{j=1}^{n} d_j = n-1 \\
&0 \le \beta_j \le 1 \qquad j=1,2,...,n \\
&d_j \in \{0,1\} \qquad j=1,2,...,n \\
&w_i \ge \varepsilon \qquad i=1,2,...,m \\
&u_r \ge \varepsilon \qquad r=1,2,...,s \\
&M, u_0 \qquad \text{free}
\end{aligned} \tag{4}$$

Model (4) is computationally efficient and also has wider range of application than models which find most CCR-efficient DMU (Model (3)), because is capable for situation in which return to scale is variable. They illustrated the applicability of their model on a real case data.

Recently, Amin (2009) extended the work of Amin & Toloo (2007) and indicated the problem of using the Model (3). He indicated that Model (3) may identify more than one efficient DMU in a given data set. Then, he presented an improved Mixed Integer Non-Linear Programming (MINLP) integrated DEA model for determining the best CCR-efficient unit, as follows:

$$
\begin{aligned}
&\min M \\
&\text{s.t.} \\
&M - d_j \geq 0 \qquad j = 1,2,...,n \\
&\sum_{i=1}^{m} w_i x_{ij} \leq 1 \qquad j = 1,2,...,n \\
&\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} w_i x_{ij} + d_j = 0 \qquad j = 1,2,...,n \\
&\sum_{j=1}^{n} \theta_j = n - 1 \\
&\theta_j - d_j \beta_j = 0 \qquad j = 1,2,...,n \\
&\beta_j \geq 1 \qquad j = 1,2,...,n \\
&d_j \geq 0 \qquad j = 1,2,...,n \\
&\theta_j \in \{0,1\} \qquad j = 1,2,...,n \\
&w_i \geq \varepsilon \qquad i = 1,2,...,m \\
&u_r \geq \varepsilon \qquad r = 1,2,...,s \\
&M \qquad \text{free}
\end{aligned}
\tag{5}
$$

Obviously, variables β_j (j = 1, ... ,n) are eliminated from the third type constraints of Model (3) and new binary variables θ_j (j = 1, ... ,n) are added in Model (5). Also the constraints $0 \leq \beta_j \leq 1$ are changed to $\beta_j \geq 1$ (j = 1, ... ,n). Moreover, the nonlinear constraints $\theta_j - d_j \beta_j = 0$ (j = 1, ... ,n) beside the constraint $\sum_{j=1}^{n} \theta_j = n - 1$ implies for one and only one of the deviation variables d_j can be vanished, meaning that only one CCR-efficient DMU can be achieved, as the most CCR-efficient DMU, by Model (5). It should be noted that Model (5) is computationally difficult to be used since it is MINLP in nature.

In order to overcome this drawback, Toloo (2010) proposed a new Mixed Integer Linear Programming (MILP) as follows:

$$
\begin{aligned}
&\min M \\
&\text{s.t.} \\
&M - d_j \geq 0 \qquad j = 1,2,...,n \\
&\sum_{i=1}^{m} w_i x_{ij} \leq 1 \qquad j = 1,2,...,n \\
&\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} w_i x_{ij} + d_j = 0 \qquad j = 1,2,...,n
\end{aligned}
\tag{6}
$$

$$\begin{aligned}
&\sum_{j=1}^{n}\theta_j = n-1 \\
&m\theta_j \le d_j \le \theta_j && j=1,2,\ldots,n \\
&d_j \ge 0,\ \theta_j \in \{0,1\} && j=1,2,...,n \\
&w_i \ge \varepsilon && i=1,2,...,m \\
&u_r \ge \varepsilon && r=1,2,...,s \\
&M && \text{free}
\end{aligned} \tag{6}$$

In this model, if $\theta_j = 0$, then constraint $d_j \le \theta_j$ forces that $d_j = 0$ and if $\theta_j = 1$, then constraint $m\theta_j \le d_j$ forces that $d_j > 0$. Hence:

$$d_j \begin{cases} = 0 & \text{if } \theta_j = 0 \\ > 0 & \text{if } \theta_j = 1 \end{cases}$$

These constraints added to the constraint $\sum_{j=1}^{n}\theta_j = n-1$ imply for one and only one of the deviation variables d_j can be vanished. According to Toloo (2010) there exists a basic model that conceptually underlies Models (3) to (6) as follows:

$$\begin{aligned}
&\min M \\
&\text{s.t.} \\
&M - d_j \ge 0 && j=1,2,...,n \\
&\sum_{i=1}^{m} w_i x_{ij} \le 1 && j=1,2,...,n \\
&\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} w_i x_{ij} + d_j = 0 && j=1,2,...,n \\
&d_j \ge 0 && j=1,2,...,n \\
&w_i \ge \varepsilon && i=1,2,...,m \\
&u_r \ge \varepsilon && r=1,2,...,s \\
&M && \text{free}
\end{aligned} \tag{7}$$

Model (7) determines efficient unit(s) with a common set of optimal weighs $(\mathbf{u}^*, \mathbf{w}^*)$, i.e. all DMUs are evaluated by a common set of weights. Indeed, in Model (7) DMU_k is an efficient unit iff $\mathbf{u}^*\mathbf{y}_k - \mathbf{w}^*\mathbf{x}_k = 0$ (or equivalently $d_k^* = 0$). Let J be a set of indexes of efficient DMU(s) mathematically, $J = \{j \mid d_j^* = 0\ j = 1,\ldots,n\}$. Clearly, if $|J|=1$, then definitely DMU_k is unique efficient DMU and hence is the best efficient DMU with the common set of optimal weights, $(\mathbf{u}^*, \mathbf{w}^*)$, iff $k \in J$. In this case, the best efficient unit can be easily determined by model (7) and no more models are needed. Otherwise, if $|J|>1$, then Model (7) cannot be used to find a single CCR-efficient unit. As he mentioned, to encounter this situation, some suitable constraints can be added to force this model to find only a single efficient unit. In other

words, by restricting the feasible region of Model (7) only one efficient DMU can be achieved. Toward this end, Amin (2009) added the following constraints to the basic model (Model (7)):

$$\begin{aligned} &\sum_{j=1}^{n}\theta_j = n-1 \\ &\theta_j - d_j\beta_j = 0 \qquad j=1,2,\ldots,n \\ &\beta_j \geq 1 \qquad\qquad j=1,2,...,n \\ &\theta_j \in \{0,1\} \qquad j=1,2,...,n \end{aligned} \tag{8}$$

and clearly, as mentioned before, the resulting model is a MINLP. Toloo (2010), instead of mixed integer non-linear constraints (8) , adjoined the following mixed integer linear constraints:

$$\begin{aligned} &\sum_{j=1}^{n}\theta_j = n-1 \\ &m\theta_j \leq d_j \leq \theta_j \qquad j=1,2,\ldots,n \\ &d_j \geq 0,\ \theta_j \in \{0,1\} \quad j=1,2,...,n \end{aligned}$$

where, $0 < m \ll 1$ is a positive parameter. Briefly, Model (6) is resulted from adding these equations to the basic model. Toloo (2010) mathematically proved that in Model (6) there exist only a single efficient DMU. Due to advantages of Model (6), this model is used to propose a new methodology for ranking data mining association rules. The next section introduces readers with these rules, their applications, and algorithms.

3. Data mining association rules

Association rules are valuable patterns that can be derived from large databases. Conceptually, an association rule indicates that the presence of a set of items (itemset) in a transaction would imply the occurrence of other items in the same transaction. The problem was first introduced by Agrawal et al. (1993), who defined it as finding all rules from transaction data satisfying the minimum support and the minimum confidence constraints. In brief, the association rule discovery problem could be divided into two separate tasks: (1) to discover all itemsets having support above a user-defined threshold, and (2) to generate rules from the frequent itemsets (Tan & Kumar, 2000).

Since introduction of association rules, this branch of data mining has gained great deal of attention by both researchers and practitioners. Today, the mining of such rules is still one of the most popular pattern discovery methods (Hipp et al., 2000). Nowadays, many applications have used association rule mining to discover useful information, including market basket analysis (Agrawal et al., 1993), web personalization (Mobasher et al., 2000 and Mulvenna et al., 2000) product recommendation (Adomavicius & Tuzhilin, 2005), soil quality assessment (Ju et al., 2010), extraction of failure patterns and forecast failure sequences of aircrafts (Han et al., 2009), credit card fraud detection (Sanchez et al., 2009), evaluation of agility in supply chains (Jain et al., 2008), exploration of cause–effect relationships in occupational accidents (Cheng et al., 2010), etc. In addition, due to its great

success and widespread application, many algorithms have been proposed for association rule mining. Based on data types which are handled by algorithm, they can be classified into three categories: nominal/Boolean data, ordinal data, and quantitative data. First, according to Agrawal et al.'s definition, transaction data is merely a set of items bought in that transaction. In other words, we can view transaction data as a set of Boolean variables, each of which corresponds to whether an item is purchased or not. The algorithms in this category find association rules from Boolean data. Second, since many data in the real world are nominal, such as hair color, grade, and birthplace, a natural extension is to modify the algorithms in the first category so that they can find association rules from nominal data. Usually, the algorithms in the first category can be easily adapted to handle nominal data. Therefore, from the algorithm's point of view, these two categories can be merged into a single category. Finally, the third category extends the algorithms so that they can find association rules from quantitative data, such as salary, height, humidity, and so on (Chen & Weng, 2008).

According to Agrawal & Sirkant (1994), given an item set $I = \{i_1, i_2, ..., i_m\}$ and given D represent a set of transaction, where each transaction T is a subset of I, $T \subset I$. A unique identifier, namely TID, is associated with each transaction. A transaction T is said to contain X, a set of items in I, if $X \subseteq T$. An association rule is said to be an implication of the form $X \Rightarrow Y$ denoting the presence of Itemset X and Y in some of the T transactions, assuming that $X, Y \subset I, X \bigcap Y = \varphi$ and $X, Y \neq \varphi$. The rule $X \Rightarrow Y$ 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 $X \bigcup Y$. It is noteworthy to mention that the idea of mining association rules originates from the analysis of market-basket data where rules like "A customer who buys product x_1 and x_2 will also buy product y with probability c%." are found. Their direct applicability to business problems together with their inherent understandability – even for non data mining experts – made association rules a popular mining method. In addition, it became clear that association rules are not restricted to dependency analysis in the context of retail applications, but are successfully applicable to a wide range of business problems (Hipp et al., 2000). In order to extract these rules, an efficient algorithm is needed that restrict 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) is such an algorithm. However, the interestingness of rule is only based on support and confidence. The Apriori algorithm is as follows:

(1) $L_1 = \text{find\_large\_1} - \text{itemsets};$
(2) $\text{for}(k = 2; L_{k-1} \neq \varphi; k++)\,\text{do begin}$
(3) $C_k = \text{apriori\_gen}(L_{k-1}); //\,\text{new candidates}$
(4) $\text{forall TID } T \in D \text{ do begin}$
(5) $C_T = \text{subset}(C_k, T); //\,\text{candidates contained in } T$
(6) $\text{forall candidates } C \in C_T \text{ do}$ (9)
(7) $C.count++;$
(8) end
(9) $L_k = \{C \in C_k \mid C.count / no\_of\_data \geq \text{minimum support threshold}\}$
(10) end
(11) $\text{Return } L = \bigcup_k Lk.$

In the above Apriori algorithm, the aprioir_gen procedure generates candidates of itemset and then uses the minimum support criterion to eliminate infrequent itemsets. The aprioir_gen procedure performs two actions, namely, join and prune, which are discussed in Han & Kamber (2001). In join step, L_{k-1} is joined with $L_{k-1}$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 and only if its (*k*-1) -subsets are also frequent; consequently aprioir_gen only generates candidates with this property, a situation easily achievable given the set L_{k-1} (Chen, 2007).
Generally, support and confidence are considered as two main criteria to evaluate the usefulness of association rules (Agrawal et al., 1993; Srikant & Agrawal, 1997). Association rules are regarded as interesting if their support and confidence are more than minimum support and minimum confidence, defined by user. In data mining, it is important but difficult to appropriately determine these two thresholds of interestingness.

4. The problem

According to Hipp et al. (2000), when mining association rules there are mainly two problems to deal with: First of all there is the algorithmic complexity. The number of rules grows exponentially with the number of items. It is to be noted that new algorithms are able to prune this immense search space based on minimal thresholds for quality measures on the rules. Second, interesting rules must be picked from the set of generated rules. This is important and costly because applying association rule algorithms on datasets results in quite large number of rules and in contrast the percentage of useful rules in typically only a very small fraction. This chapter generally addresses the second problem.
In existing data mining techniques, there exist some situations that make necessary the prioritization of rules for selecting and concentrating on more valuable rules due to the number of qualified rules (Tan & Kumar, 2000) and limited business resources. In other words, one main problem of association rule induction is that there are so many possible rules. Obviously such a vast amount of rules cannot be processed by inspecting each one in turn (Tajbakhsh et al., 2009). Even though the purpose of data mining is rule (pattern) extraction that is valuable for decision making, patterns are deemed 'interesting' just on the basis of passing certain statistical tests such as support/confidence in data mining. To the enterprise, however, it remains unclear how such patterns can be used to maximize business values. Choi et al. (2005) believe that the major obstacle lies in the gap between statistic-based summaries (the statistic-based pattern extraction) extracted by traditional rule mining and a profit-driven action (the value-based decision making) required by business decision making which is characterized by explicit consideration of conflicts of business objectives and by multiple decision makers' involvement for corporate decision making.
It is to be noted that confidence and support of the rules are not sufficient measures to select ''interesting'' rules (Tajbakhsh et al., 2009). An association rule which is advantageous and profitable to sellers may not be discovered by setting constraints of minimum support and minimum confidence in the mining process because high value products are relatively uncommonly bought by customers, (Chen, 2007). Consider the following case, entitled the Ketel vodka and Beluga caviar in the market basket 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 et al. (2003). The association rule of [wine $\Rightarrow$ salmon, 1%, 80%] may be more interesting to analysts than [bread $\Rightarrow$ 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. This chapter proposes a new method for estimating and ranking the efficiency (interestingness or usefulness) of association rules with multiple criteria by using a non-parametric approach, DEA. The interestingness of association rules is measured by considering 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.

5. Previous related studies

The problem of ranking discovered rules of data mining has gained attention by some researchers. Sirkant et al. (1997) presented three integrated algorithms for mining association rules with item constraint. Moreover, Lakshmanan et al. (1998) extended the approach presented by Srikant et al. to consider much more complicated constraints, including domain, class, and SQL-style aggregate constraints. Liu et al. (2000) presents an Interestingness Analysis System (IAS) to help the user identify interesting association rules. In their proposed method, they consider two main subjective interestingness measures, unexpectedness and actionability. The degree of unexpectedness of rules can be measured by the extent to which they surprise the analyst. Meanwhile, the degree of actionability can be measured by the extent to which analysts can use the discovered rules to their advantage. Choi et al. (2005), using Analytic Hierarchy Process (AHP) presented a method for association rules prioritization which considers the business values which are comprised of objective metric or managers' subjective judgments. They believed that proposed method makes synergy with decision analysis techniques for solving problems in the domain of data mining. Nevertheless this method requires large number of human interaction to obtain weights of criteria by aggregating the opinions of various managers. Chen (2007) developed their work and proposed a Data Envelopment Analysis (DEA) based methodology for ranking association rules while considering multiple criteria. During his ranking procedure, he uses a DEA model, proposed by Cook & Kress (1990), to identify efficient association rules.

In fact, his proposed method uses a DEA model, proposed by Cook & Kress (1990), for identifying efficient association rules. This model is as follows:

$$
\begin{aligned}
&\max \sum_{j=1}^{k} \mathrm{w_j} v_{oj} \\
&\text{s.t.} \\
&\quad \sum_{j=1}^{k} \mathrm{w_j} v_{ij} \le 1 \qquad i = 1,2,...,m \\
&\quad w_j - w_{j+1} \ge d(j,\varepsilon) \qquad j = 1,2,...,k-1 \\
&\quad w_k \ge d(k,\varepsilon)
\end{aligned}
\tag{10}
$$

where w_j denotes the weight of the jth place; v_{ij} represents the number of jth place votes of candidate $i\,(i=1,2,...,m,\ j=1,2,...,k)$ and $d(\bullet,\varepsilon)$, known as the discrimination intensify function, is nonnegative and nondecreasing in ε and satisfies $d(\bullet,\varepsilon)=0$.
Model (3) should be resolved for each candidate $o, o=1,2,...,m$. The resulting objective value is the preference score of candidate o. Because of the fact that DEA frequently generates several efficient candidates (Obata & Ishii, 2003), Chen's proposed method uses another DEA model, proposed by Obata & Ishii (2003), for discriminating efficient association rules. This model is as follows:

$$\begin{aligned}
&\max \sum_{j=1}^{k} w_j \\
&\text{s.t.} \\
&\sum_{j=1}^{k} w_i v_{oj} = 1 \\
&\sum_{j=1}^{k} w_j v_{ij} \le 1 \quad \text{for all efficient } i \ne o \\
&w_j - w_{j+1} \ge d(j,\varepsilon) \qquad j=1,2,...,k-1 \\
&w_j \ge d(j,\varepsilon)
\end{aligned} \tag{11}$$

It should be noted that this model does not employ any information about inefficient candidates and should be solved only for efficient association rules. It should be noted that his proposed method requires the first model to be solved for all DMUs and the second model to be solved for efficient DMUs. As a drawback, this approach requires considerable number of Linear Programming (LP) models to be solved. Toloo et al. (2009) mentioned following problems in using Chen's proposed method:

- Chen's method requires computing v_{ij} from y_{ij} (j^{th} outputs of i^{th} association rule). Although, the algorithm of computing v_{ij} from y_{ij} is polynomial, it is time consuming. Identifying efficient association rules can be done through a more simple and efficient way. Interested readers are referred to Toloo et al. (2009) for further explanations.
- Result of Chen's method is immensely dependent on discrimination intensify function.
- Suppose that there are e efficient association rules which are obtained from Model (10). To rank e efficient units, Chen's method includes solving $(n+e)$ LPs.
- To overcome above problems, Toloo et al. (2009) improved the work of Chen (2007) and proposed a methodology which ranked association rules by solving less numbers of LP models. Their methodology was based on Model (3) and include following steps:

Step 0. Let $T=\phi$ and $e=$ number of association rules to be ranked.

Step 1. Solve following model:

$$\begin{aligned}
&\min M \\
&\text{s.t.} \\
&M - d_j \ge 0 \qquad\qquad j=1,2,...,n \\
&\sum_{r=1}^{s} u_r y_{rj} + d_j - \beta_j = 1 \qquad j=1,2,...,n
\end{aligned} \tag{12}$$

$$
\begin{aligned}
&\sum_{j=1}^{n} d_j = n-1 \\
&d_j = 1 && \forall j \in \mathrm{T} \\
&0 \le \beta_j \le 1 && j = 1,2,...,n \\
&d_j \in \{0,1\} && j = 1,2,...,n \\
&w_i \ge \varepsilon && i = 1,2,...,m \\
&u_r \ge \varepsilon && r = 1,2,...,s \\
&M && \text{free}
\end{aligned}
\tag{12}
$$

Suppose in optimal solution $d_p^* = 0$.

Step 2. Let $\mathrm{T} = \mathrm{T} \cup \{p\}$.

Step 3. If $|\mathrm{T}| = e$, then stop; otherwise go to Step 1.

Indeed, in Step 1 of Toloo et al.'s algorithm, an association rule is identified as best efficient rule. It is noteworthy to mention that Model (12) is based on Model (3) and the only difference is that Model (12) considers a single input with equal value for all association rules. This is because of the fact that all evaluation criteria of association rules are output in nature. Clearly using DEA models (e.g. Model (3)) requires input data of DMUs and consequently Model (12) were developed by Toloo et al. (2009) to handle this situation and be applicable for ranking association rules. In Step 2, the best efficient association rule identified in Step 1 is added to T. Next, in step 3, if all rules are ranked, the algorithm finishes, else it goes to next iteration; finally, after e iterations all association rules are ranked. Although they improved Chen's method, their methodology was based on Model (3) which suffers from some drawbacks, as mentioned in Section 2.2. In the next section we propose a new methodology based on the latest developments by Toloo (2010).

6. Proposed method

This section proposes a new DEA-based methodology for ranking units. Previously, various methodologies have been proposed to rank DMUs, most of which are reviewed by Adler (2002). Interested readers can refer to this reference for further discussion on ranking methods. DEA is able to compare DMUs using different criteria as the basis for comparison, while utilizing all inputs and outputs simultaneously. Generally, in DEA applications, the criteria that should be maximized are considered as outputs and ones that should be minimized are treated as inputs. In case of ranking data mining association rules, previous studies such as Chen (2007) and Toloo et al. (2009) considered criteria that are outputs in nature. In other words, the more value of those criteria the more interesting association rule for business. In this section, a new DEA model is presented which identifies the best efficient unit by considering only output data of DMUs. The model proposed as:

$$
\begin{aligned}
&\min M \\
&\text{s.t.} \\
&\quad M - d_j \ge 0 && j = 1,2,...,n
\end{aligned}
\tag{13}
$$

$$
\begin{aligned}
&\min M \\
&\text{s.t.} \\
&\quad M - d_j \geq 0 && j = 1,2,...,n \\
&\quad \sum_{r=1}^{s} u_r y_{rj} + d_j = 1 && j = 1,2,...,n \\
&\quad \sum_{j=1}^{n} \theta_j = n - 1 \\
&\quad m\theta_j \leq d_j \leq \theta_j && j = 1,2,...,n \\
&\quad d_j \geq 0 && j = 1,2,...,n \\
&\quad \theta_j \in \{0,1\} && j = 1,2,...,n \\
&\quad w_i \geq \varepsilon && i = 1,2,...,m \\
&\quad u_r \geq \varepsilon && r = 1,2,...,s \\
&\quad M && \text{free}
\end{aligned}
\tag{13}
$$

The structure of Model (13) is similar to Model (6) and the main idea is trying to find only one most efficient DMU. However, Model (6) considers various criteria as inputs and outputs and Model (13) considers only output data of DMUs. In other words, Model (13) is applicable for situations in which all evaluation criteria are output in nature (e.g. association rules). In simple words, Model (13) is a customized version of Model (6). Using Model (13), in this section Toloo et al.'s methodology is improved as follows:

Step 0. Let $T = \varphi$ and $e =$ number of association rules to be ranked.

Step 1. Solve following model:

$$
\begin{aligned}
&\min M \\
&\text{s.t.} \\
&\quad M - d_j \geq 0 && j = 1,2,...,n \\
&\quad \sum_{r=1}^{s} u_r y_{rj} + d_j = 1 && j = 1,2,...,n \\
&\quad \sum_{j=1}^{n} \theta_j = n - 1 \\
&\quad m\theta_j \leq d_j \leq \theta_j && j = 1,2,...,n \\
&\quad \theta_j = 1 && j \in T \\
&\quad d_j \geq 0 && j = 1,2,...,n \\
&\quad \theta_j \in \{0,1\} && j = 1,2,...,n \\
&\quad w_i \geq \varepsilon && i = 1,2,...,m \\
&\quad u_r \geq \varepsilon && r = 1,2,...,s \\
&\quad M && \text{free}
\end{aligned}
\tag{14}
$$

Suppose in optimal solution $d_p^* = 0$.

Step 2. Let $\mathrm{T} = \mathrm{T} \cup \{p\}$.

Step 3. If $|\mathrm{T}| = e$, then stop; otherwise go to Step 1.

Step 1 ensures that one and only one DMU is selected as the best efficient unit. Step 2 adds this DMU to T and Step 3 ensures that all DMUs are ranked. Although the proposed methodology is similar to Toloo et al.'s methodology in structure, it overcomes the former drawbacks. In other words, as contribution, proposed methodology is based on latest development in DEA and overcome the problems of previous DEA models.

7. Illustrative example

In this section, to indicate the application of proposed method and compare its results with previous methods, an example of market basket data is adopted from Chen (2007). 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 and presented in Table (1).

By applying Model (14) to data presented in Table (1), DMU_{12} is identified as the most efficient association rule (considering m=0.001). In Step 2, 12 is added to T and in Step 3, methodology enters second iteration. Based on the methodology, in second iteration the constraint $\theta_{12} = 1$ is added to model. Solving Model (14) in second iteration resulted in ($\theta_{18}^* = 0$, $\theta_{j\neq 18}^* = 1$) implies that DMU_{18} is second efficient association rule. Table (2) presents results of ranking efficient rules in comparison to Chen's method and Toloo et al.'s method.

Table 2 shows that the results of proposed method are different from results of previous methods. In order to provide readers with further insight, basic model has been applied to data set of Table.1[1]. As mentioned in Section 2.2, basic model determines DMU(s) which are candidate to be the best efficient DMU with considering common set of weights. The results show that $d_{12}^* = 0$ meaning that DMU_{12} should be the highest ranked DMU, since there is no other candidate to be the best efficient DMU. It is notable that this DMU is ranked 10th by Chen's method and 4th by Toloo et al.'s method. Obviously, proposed method provides decision maker with more accurate results as its main advantage to previous methods.

Association Rule Number (DMU)	Support (%)	Confidence (%)	Itemset value	Cross-selling profit
1	3.87	40.09	337.00	25.66
2	1.42	18.17	501.00	11.63
3	2.83	17.64	345.00	11.29
4	2.34	30.83	163.00	19.73
5	2.63	23.90	325.00	15.30
6	1.19	55.65	436.00	35.61
7	1.19	47.42	598.00	30.35
8	1.19	15.70	436.00	52.91

[1] Appendix A indicates GAMS program of basic model.

Association Rule Number (DMU)	Support (%)	Confidence (%)	Itemset value	Cross-selling profit
9	1.19	10.82	598.00	36.45
10	1.19	12.32	436.00	20.08
11	1.19	12.32	598.00	40.04
12	3.87	38.08	337.00	103.97
13	1.18	15.09	710.00	41.19
14	2.44	15.22	554.00	41.56
15	2.14	28.21	372.00	77.02
16	2.51	22.81	534.00	62.26
17	1.19	50.92	436.00	139.02
18	1.19	45.25	598.00	123.52
19	1.19	11.70	436.00	43.54
20	1.19	11.70	598.00	62.50
21	1.42	13.99	501.00	61.16
22	1.18	12.23	710.00	53.45
23	1.50	13.64	698.00	59.59
24	2.83	27.82	345.00	78.17
25	2.44	25.27	554.00	71.00
26	1.25	15.97	718.00	44.87
27	1.22	34.89	339.00	98.04
28	1.30	35.12	435.00	98.68
29	1.42	33.81	534.00	95.01
30	1.91	25.26	380.00	70.97
31	1.43	37.14	618.00	104.35
32	2.38	21.63	542.00	60.78
33	1.18	30.24	366.00	84.98
34	1.23	29.36	626.00	82.51
35	1.58	22.65	354.00	63.64
36	2.34	22.99	163.00	22.76
37	2.14	22.14	372.00	21.92
38	1.91	11.94	380.00	11.82
39	2.03	18.42	360.00	18.23
40	1.19	30.73	436.00	30.43
41	2.63	25.87	325.00	67.52
42	2.51	25.98	534.00	67.81
43	1.50	19.16	698.00	50.02
44	2.38	14.85	542.00	38.75
45	2.03	26.73	360.00	69.78
46	1.19	30.73	598.00	80.22

Table 1. Data of Association Rules

Ranking	Association Rule Number (DMU)		
	Chen's Method	Toloo et al.'s Method	Proposed Method
1	26	18	12
2	22	23	18
3	18	26	26
4	17	12	43
5	7	31	23
6	23	43	31
7	6	22	1
8	43	6	7
9	31	17	6
10	12	1	17
11	1	7	22

Table 2. Ranking of Proposed Method in Comparison to Chen's method and Toloo et al.'s method

8. Conclusion

Association rule discover is one of widely recognized data mining techniques which has gained great deal of attention recently. Association rules are valuable patterns because they offer useful insight into the types of dependencies that exist between attributes of a data set. By applying association rules algorithms, there exist some situations that make necessary the prioritization of rules for selecting and concentrating on more valuable rules due to the number of qualified rules and limited business resources. In other words, one main problem of association rule induction is that there are so many possible rules. Hence, evaluating the interestingness or usefulness of association rules and ranking them is a critical task in data mining applications. Indeed, selecting the more valuable rules for implementation increases the possibility of success in data mining. In this chapter, a new methodology proposed for ranking association rules of data mining. This method uses a non-parametric linear programming technique, DEA, for ranking the units. As an advantage, the proposed method utilizes the latest developments in DEA models and finds the best efficient association rule by solving only one MILP. The applicability of proposed method is indicated and its results are compared with the results of previous methods. Using basic model presented in this chapter, it is shown that results of new proposed method is more advantageous than previous ones since it results in more accurate results.
As directions for further researches, extending the applicability of proposed method to imprecise/fuzzy situations is suggested. Obviously, in many real world business applications of data mining, data of association rules is imprecise/fuzzy. Besides, future researchers could extend the applicability of proposed method to solve other business decision problems such as supplier selection, ranking of projects, and etc.

9. References

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

Adomavicius, G., Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowledge Data Engrg*. 17, 734–749.

Agrawal, R., Imielinski, T., Swami, A. (1993). Mining association between sets of items in massive database. *International proceedings of the ACM-SIGMOD international conference on management of data*, 207–216.

Agrawal, R., Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the international conference on very large data bases*, 407–419.

Amin, Gholam R., Toloo, M. (2004). A polynomial-time algorithm for finding Epsilon in DEA models, *Computers and Operations Research,* 31, 803–805.

Amin, Gholam R., Toloo, M. (2007). Finding the most efficient DMUs in DEA: An improved integrated model. *Computers & Industrial Engineering*, 52, 71-77.

Amin, Gholam R., (2009). Comments on finding the most efficient DMUs in DEA: An improved integrated model. *Computers & Industrial Engineering*, 56, 1701-1702.

Banker, R. D., Charnes, A., Cooper, W. W. (1984). Some models for estimating technical and scale inefficiency in data envelopment analysis. *Management Science*, 30, 1078–1092.

Camanho, A. S., Dyson, R.G. (2005). Cost efficiency measurement with price uncertainty: a DEA application to bank branch assessments. *European Journal of Operational Research*, 161, 432-446.

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

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, X., Skully M., Brown, K. (2005). Banking efficiency in China: Application of DEA to pre- and post-deregulation eras: 1993-2000. *China Economic Review*, 16, 229-245.

Chen, Y.L., Weng, C.H. (2008). Mining association rules from imprecise ordinal data. *Fuzzy Sets and Systems*, 159, 460-474.

Chen, Y.L., Weng, C.H. (2009). Mining fuzzy association rules from questionnaire data. *Knowledge-Based Systems*. 22, 46–56.

Cheng, C.W., Lin, C.C., Leu, S.S. (2010). Use of association rules to explore cause–effect relationships in occupational accidents in the Taiwan construction industry. *Safety Science*, 48, 436–444.

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, 867-878.

Cook, W.D., Seiford, L.M. (2009). Data Envelopment Analysis (DEA): Thirty years on. *European Journal of Operational Research*, 192, 1-17.

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

Cooper, W.W., Seiford,L.M., Tone, K. (2007). Data Envelopment Analysis: A Comprehensive text with Models, Applications, References and DEA-Solver Software. Springer, 978-0387-45283-8.

Edirisinghe, N.C.P., Zhang, X. (2007). Generalized DEA model of fundamental analysis and its application to portfolio optimization. *Journal of Banking & Finance*, 31, 3311-3335.

Emrouznejad, A., Tavares, G., Parker, B. (2007). A bibliography of data envelopment analysis (1978–2003). *Socio-Economic Planning Sciences*, 38, 159-229.

Ertay, T., Ruan, D., Tuzkaya, U. R. (2006). Integrating data envelopment analysis and analytic hierarchy for the facility layout design in manufacturing systems. *Information Sciences*, 176, 237–262.

Han, J.W., Kamber, M. (2001). Data Mining: Concepts and Techniques, *MORGAN KAUFMANN PUBLIOSHERS*, San Francisco.

Han, H.K., Kim, H.S., Sohn, S.Y. (2009). Sequential association rules for forecasting failure patterns of aircrafts in Korean airforce. *Expert Systems with Applications*, 36, 1129–1133.

Hipp, J., Guntzer, U., Nakhaeizadeh, G. (2000). Algorithms for Association Rule Mining – A General Survey and Comparison. *SIGKDD Explorations*, 2, 58-64.

Jain, V., Benyoucef, L., Deshmukh, S.G. (2008). A new approach for evaluating agility in supply chains using

Fuzzy Association Rules Mining. *Engineering Applications of Artificial Intelligence*. 21, 367–385.

Johnes, J. (2006). Measuring teaching efficiency in higher education: An application of data envelopment analysis to economics graduates from UK Universities 1993. *European Journal of Operational Research*, 174, 443-456.

Liu, J., Ding, F.Y., Lall, V. (2000). Using data envelopment analysis to compare suppliers for supplier selection and performance improvement, *Supply Chain Management,* 5, 143–150.

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

Liu, S.T. (2008). A fuzzy DEA/AR approach to the selection of flexible manufacturing systems. *Computers & Industrial Engineering*, 54, 66-76.

Mannino, M., Hong, S.N., Choi, I.J. (2008). Efficiency evaluation of data warehouse operations. *Decision Support Systems*, 44, 883-898.

Mobasher, B., Cooley, R., Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM*, 43, 142–151.

Mulvenna, M. D., Anand, S. S., Bu¨chner, A. G. (2000). Personalization on the net using Web mining. *Communications of the ACM*, 43, 123–125.

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*, 13–24.

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

Sanchez, D., Vila, M.A., Cerda, L., Serrano, J.M. (2009). Association rules applied to credit card fraud detection. *Expert Systems with Applications*, 36, 3630–3640.

Shafer, S.M., Byrd, T.A. (2000). A framework for measuring the efficiency of organizational investments in information technology using data envelopment analysis. *Omega*, 28, 125-141.

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*, 67–73.

Tajbakhsh, A., Rahmati, M., Mirzaei, A. (2009). Intrusion detection using fuzzy association rules. *Applied Soft Computing*, 9, 462-469.

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, 661–666.

Toloo, M. (2010). A new mixed integer linear programming integrated model for finding the most efficient unit in data envelopment analysis. *Computers & Industrial Engineering* (Submitted Manuscript Number: CAIE-D-10-00438.

Toloo, M., Nalchigar, S. (2009). A new integrated DEA model for finding most BCC-efficient DMU. *Applied Mathematical Modelling*, 33, 597-604.

Toloo, M., Sohrabi. B., Nalchigar, S. (2009). A new method for ranking discovered rules from data mining by DEA. *Expert Systems with Applications*, 36, 8503-8508.

Weber, C.A., Current, J.R., Desai, A. (1998). Non-cooperative negotiation strategies for vendor selection. *European Journal of Operational Research*, 108, 208-223.

Whitten, I.H., Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. *MORGAN KAUFMAN PUBLISHERS*. 0-12-088407-0

Xue, Y. J., Liu, S. G., Hu, Y. M. and Yang, J. F. (2010). Soil quality assessment using weighted fuzzy association rules. *Pedosphere*. 20, 334–341.

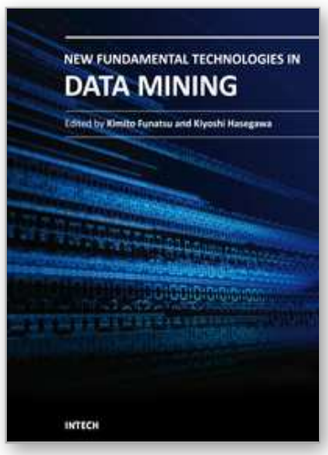
Appendix A

Basic Model in GAMS

```
$title Basic Model
$ontext
This program is written as a part of a book chapter with following specifications. In brief, this
program shows applicability of a mathematical model which is proposed by Toloo (2010) for finding
the best data mining association rule.

Authors: Mehdi Toloo & Soroosh Nalchigar
Book Name: Data Mining
Chapter Name:
On Ranking Discovered Rules of Data Mining by Data Envelopment Analysis: Some New Models
with Applications
Publisher: INTECH
2011

$offtext
SETS
  J   "Number of DMUs"     /01*46/
  O   "Number of outputs"   /1*4/

ALIAS  (J,L);

PARAMETERS
   Yo(O)     "Output vector of DMUo"
  ep;
```

```
ep=0.0001;

TABLE Y(J,O)   "Output vectors of all DMUs"
          1          2          3          4
01        3.87       40.09      337        25.66
02        1.42       18.17      501        11.63
03        2.83       17.64      345        11.29
04        2.34       30.83      163        19.73
05        2.63       23.9       325        15.3
06        1.19       55.65      436        35.61
07        1.19       47.42      598        30.35
08        1.19       15.7       436        52.91
09        1.19       10.82      598        36.45
10        1.19       12.32      436        20.08
11        1.19       12.32      598        40.04
12        3.87       38.08      337        103.97
13        1.18       15.09      710        41.19
14        2.44       15.22      554        41.56
15        2.14       28.21      372        77.02
16        2.51       22.81      534        62.26
17        1.19       50.92      436        139.02
18        1.19       45.25      598        123.52
19        1.19       11.7       436        43.54
20        1.19       11.7       598        62.5
21        1.42       13.99      501        61.16
22        1.18       12.23      710        53.45
23        1.5        13.64      698        59.59
24        2.83       27.82      345        78.17
25        2.44       25.27      554        71
26        1.25       15.97      718        44.87
27        1.22       34.89      339        98.04
28        1.3        35.12      435        98.68
29        1.42       33.81      534        95.01
30        1.91       25.26      380        70.97
31        1.43       37.14      618        104.35
32        2.38       21.63      542        60.78
33        1.18       30.24      366        84.98
34        1.23       29.36      626        82.51
35        1.58       22.65      354        63.64
36        2.34       22.99      163        22.76
37        2.14       22.14      372        21.92
38        1.91       11.94      380        11.82
39        2.03       18.42      360        18.23
40        1.19       30.73      436        30.43
41        2.63       25.87      325        67.52
```

```
42      2.51      25.98      534      67.81
43      1.5       19.16      698      50.02
44      2.38      14.85      542      38.75
45      2.03      26.73      360      69.78
46      1.19      30.73      598      80.22;

VARIABLES
Mstar
M
 POSITIVE VARIABLES
 u(O)
 d(j)

EQUATIONS
obj
const1
const2
const3;
    obj..          Mstar =e=M ;
    const1(j)..   M-d(j)=g=0;
    const2(j)..   sum(o,u(o)*y(j,o))+d(j)=e= 1 ;
    const3(o)..  u(o)=g=ep;

FILE  result/Basic_Model_Results.txt/;

MODEL Basic_Model /all/ ;

SOLVE Basic_Model using LP minimizing Mstar ;
PUT result ;
PUT "M*=", PUT Mstar.L:9:6,"  ", PUT /;
LOOP (j, PUT "d*_", PUT  j.tl, PUT@7 "= " PUT d.l(j):10:8, PUT /);
```



New Fundamental Technologies in Data Mining

Edited by Prof. Kimito Funatsu

ISBN 978-953-307-547-1
Hard cover, 584 pages
Publisher InTech
Published online 21, January, 2011
Published in print edition January, 2011

The progress of data mining technology and large public popularity establish a need for a comprehensive text on the subject. The series of books entitled by "Data Mining" address the need by presenting in-depth description of novel mining algorithms and many useful applications. In addition to understanding each section deeply, the two books present useful hints and strategies to solving problems in the following chapters. The contributing authors have highlighted many future research directions that will foster multi-disciplinary collaborations and hence will lead to significant development in the field of data mining.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Mehdi Toloo and Soroosh Nalchigar (2011). On Ranking Discovered Rules of Data Mining by Data Envelopment Analysis: Some Models with Wider Applications, New Fundamental Technologies in Data Mining, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-547-1, InTech, Available from: http://www.intechopen.com/books/new-fundamental-technologies-in-data-mining/on-ranking-discovered-rules-of-data-mining-by-data-envelopment-analysis-some-models-with-wider-appli

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen