



## A note on “A new method for ranking discovered rules from data mining by DEA”, and a full ranking approach

A.A. Foroughi

Department of Mathematics, University of Qom, Isfahan old road, Qom, Iran

### ARTICLE INFO

#### Keywords:

Data envelopment analysis  
Data mining  
Multiple criteria

### ABSTRACT

In a recent paper by Toloo et al. [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], they proposed a new integrated data envelopment analysis model to find most efficient association rule in data mining. Then, utilizing this model, an algorithm is developed for ranking association rules by considering multiple criteria. In this paper, we show that their model only selects one efficient association rule by chance and is totally depended on the solution method or software is used for solving the problem. In addition, it is shown that their proposed algorithm can only rank efficient rules randomly and will fail to rank inefficient DMUs. We also refer to some other drawbacks in that paper and propose another approach to set up a full ranking of the association rules. A numerical example illustrates some contents of the paper.

© 2011 Elsevier Ltd. All rights reserved.

### 1. Introduction

In multiple criteria decision making, each alternative is evaluated based on a number of different criteria. One popular solution method for a multiple criteria problem is to obtain weights for criteria and use the weighted sum of the criteria as the score for each alternative. These scores can be utilized for ranking the alternatives or selecting one with the biggest score as the final decision. The important question here is how to obtain these weights. There are many methods for this purpose; one of them is data envelopment analysis (DEA), a linear programming based method introduced by Charnes, Cooper, and Rhodes (1978). DEA uses the best favorable weights corresponding to each decision making unit (DMU) to obtain the scores.

In a recent paper, Chen (2007) used DEA in a data mining problem, to evaluate association rules with multiple criteria. For this purpose, Chen used the approach of Obata and Ishii (2003), which has been proposed for voting system. In Obata and Ishii approach, the proposed DEA/AR model of Cook and Kress (1990) is utilized first, to obtain efficiency scores, and then another model is used to discriminate efficient candidates. The candidates in data mining are the association rules (DMUs in DEA and alternatives in multiple criteria decision making), which are evaluated based on some criteria.

In a more recent paper, Toloo, Sohrabi, and Nalchigar (2009) proposed another DEA approach for data mining. They counted

some advantages of their method; one of the advantages was providing a full ranking for the association rules, which we will show to be not correct. We will refer to some disadvantages of their method and it will be shown that their proposed algorithm can only rank efficient DMUs randomly. In addition it will be shown that the model is used in their algorithm will be infeasible corresponding to inefficient DMUs. In general, their model is not convenient to rank efficient DMUs, and will fail to rank inefficient DMUs.

As a convenient full ranking approach for this problem, we suggest a slightly modified method of Foroughi and Tamiz (2005). In addition, to improve the approach and decrease the number of models and constraints, we have developed an algorithm, by modifying and simplifying the proposed algorithm of Toloo et al., to obtain efficient DMUs. The algorithm will improve the model particularly when we need only to rank efficient DMUs or select one association rule. It will be seen that the proposed approach will provide a full ranking of the rules for the example of market basket analysis, which was used by Chen (2007) and Toloo et al. (2009).

### 2. The drawbacks

By considering the criteria as outputs and the association rules as the DMUs, let  $y_{rj}$  ( $r = 1, 2, \dots, s$ ) be the data for the  $j$ th DMU ( $j = 1, 2, \dots, n$ ) and  $u_r$  ( $r = 1, 2, \dots, s$ ) be the sequence of weights given to the criteria. The proposed model of Toloo et al. (2009) to identify most efficient DMU was as follows:

E-mail addresses: [a-foroughi@qom.ac.ir](mailto:a-foroughi@qom.ac.ir), [aa\\_foroughi@yahoo.com](mailto:aa_foroughi@yahoo.com)

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

where,  $\varepsilon$  is a positive infinitesimal value to prevent selecting zero weights. Note that we have deleted the constraints  $w_i \geq \varepsilon$ ,  $i = 1, 2, \dots, m$ , from their model (model (4) in Toloo et al. (2009)) since they are obviously redundant (it seems a typos error). Here,  $d_j$  is a binary variable and they claim: DMU $j$  is most efficient if and only if  $d_j = 0$ . We will show that correspond to every efficient DMU $j$  we can obtain an optimal solution for model (1) with  $d_j = 0$ , that means each efficient DMU can be considered as most efficient. Hence the concept of most efficient DMU in their paper is not well defined.

Before proving our claim, and for later purposes, we first simplify this model. From the constraint of the model (1), in every feasible solution, one  $d_j$  is equal to zero and the others are equal to one. Hence, from the constraints  $M - d_j \geq 0$ ,  $j = 1, 2, \dots, n$ , the optimal value of this problem is equal to one, so all feasible solutions of (1) are optimal and these constraints can be removed from the model. Now, again from the constrains, we have  $\beta_j = \sum_{r=1}^s u_r y_{rj} + d_j - 1$ ,  $j = 1, 2, \dots, n$ . By replacing  $\beta_j$  in the model, solving model (1) is equivalent to determining a feasible solution for the following system:

$$\begin{aligned}
 0 \leq \sum_{r=1}^s u_r y_{rj} + d_j - 1 &\leq 1, \quad j = 1, 2, \dots, n \\
 \sum_{j=1}^n d_j &= n - 1 \\
 d_j &\in \{0, 1\}, \quad j = 1, 2, \dots, n \\
 u_r &\geq \varepsilon, \quad r = 1, 2, \dots, s
 \end{aligned} \tag{2}$$

Or equivalently:

$$\begin{aligned}
 1 - d_j \leq \sum_{r=1}^s u_r y_{rj} \leq 2 - d_j, \quad j = 1, 2, \dots, n \\
 \sum_{j=1}^n d_j &= n - 1 \\
 d_j &\in \{0, 1\}, \quad j = 1, 2, \dots, n \\
 u_r &\geq \varepsilon, \quad r = 1, 2, \dots, s
 \end{aligned} \tag{3}$$

Using the previous discussions we have proved the following result.

**Result 1.** Each feasible solution of system (3) is corresponded to an optimal solution for model (1), and vice versa.

Now, let DMU $o$  be an arbitrary DEA-efficient DMU and consider the following DEA model:

$$\begin{aligned}
 \max Z_o &= \sum_{r=1}^s u_r y_{ro} \\
 \text{s.t. } \sum_{r=1}^s u_r y_{rj} &\leq 1, \quad j = 1, 2, \dots, n, \\
 u_r &\geq \varepsilon, \quad r = 1, 2, \dots, s.
 \end{aligned} \tag{4}$$

Let  $u_r^*$ ,  $r = 1, 2, \dots, s$  be an optimal solution for this model. Since DMU $o$  is efficient, for this optimal solution we have  $Z_o^* = \sum_{r=1}^s u_r^* y_{ro} = 1$ , and  $\sum_{r=1}^s u_r^* y_{rj} \leq 1$ , for the other  $j$ . Now we put  $d_o = 0$ , and  $d_j = 1$ , for all  $j \neq o, j = 1, 2, \dots, n$ . This is clearly a feasible solution of system (3) and so, by Result 1, it is an optimal solution for model (1), by putting  $\beta_j = \sum_{r=1}^s u_r y_{rj} + d_j - 1$ ,

$j = 1, 2, \dots, n$ , and  $M = 1$ . This shows that all efficient DMUs are most efficient DMU by the proposed model of Toloo et al. (2009). Indeed we have proved the following result.

**Result 2.** Corresponding to every efficient DMU $o$ , we can obtain an optimal solution for model (1) with  $d_o = 0$ . Hence, by the definition of Toloo et al. (2009), all the efficient DMUs can be considered as the most efficient DMU, which is not a proper definition.

The previous discussions prove that the proposed method of Toloo et al. (2009) cannot discriminate between efficient DMUs and the discrimination in their paper is only depended on the selected optimal solution from alternative optimal solutions. Indeed, if different software is used or even another time the model is applied by the same software, different results can be obtained.

Now, we refer to another drawback in their paper. They proposed the following algorithm for ranking the first  $e$  best DMUs.

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

Step 1: Solve following model:

$$\begin{aligned}
 M^* &= \min M \\
 \text{s.t. } M - d_j &\geq 0, \quad j = 1, 2, \dots, n \\
 \sum_{r=1}^s u_r y_{rj} + d_j - \beta_j &= 1, \quad j = 1, 2, \dots, n \\
 \sum_{j=1}^n d_j &= n - 1 \\
 d_j &= 1 \quad \forall j \in T \\
 0 \leq \beta_j &\leq 1, \quad d_j \in \{0, 1\}, \quad j = 1, 2, \dots, n \\
 u_r &\geq \varepsilon, \quad r = 1, 2, \dots, s
 \end{aligned} \tag{5}$$

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

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

Step 3: If  $|T| = e$ , then stop; otherwise go to Step 1.

Now assume that the number of efficient DMUs, which is also depended on the value of  $\varepsilon$ , is  $k$ . As it was explained before, one of the efficient DMUs (dependent on the software, and indeed the good lock one) is selected as the most efficient in the first iteration. Then, similarly it can be seen that the other efficient DMUs are selected one after another until all efficient DMUs are ranked randomly after iteration  $l = \min\{e, k\}$ . Now, if we assume that  $l = k < e$ , then in iteration  $k + 1$  we should solve model (5) when  $d_j = 1 \quad \forall j \in E$ , where  $E$  is the index set of all efficient DMUs. In this case, model (5) is infeasible. Indeed, by contradiction, if we assume that it is feasible then we should have a feasible solution for (5) with  $d_p^* = 0$ , which  $p \notin E$ . This means we have positive weights for that  $\sum_{r=1}^s u_r y_{rp} \geq 1$ , and  $\sum_{r=1}^s u_r y_{rj} \leq 1$ , for all the other  $j$ , which is impossible since DMU $p$  is inefficient. Hence, we have the following result.

**Result 3.** The proposed algorithm of Toloo et al. (2009) will fail after  $k$  iteration, which  $k$  is the number of efficient DMUs.

Note: There is also another error in determining the assurance region for epsilon in model (5) in Toloo et al. (2009). Indeed, that model should be changed to the following one:

$$\begin{aligned}
 \varepsilon^* &= \max \varepsilon \\
 \text{s.t. } \sum_{r=1}^s u_r y_{rj} &\leq 1, \quad j = 1, 2, \dots, n, \\
 u_r &\geq \varepsilon, \quad r = 1, 2, \dots, s,
 \end{aligned}$$

which its optimal value can be obtained easily as:  $\varepsilon^* = \min_j (1/\sum_{r=1}^s y_{rj}), j = 1, 2, \dots, n$ .

As it was explained, the approach of Toloo et al. (2009) cannot rank efficient DMUs properly and will fail to rank inefficient DMUs.

متن کامل مقاله

دریافت فوری ←

**ISI**Articles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات