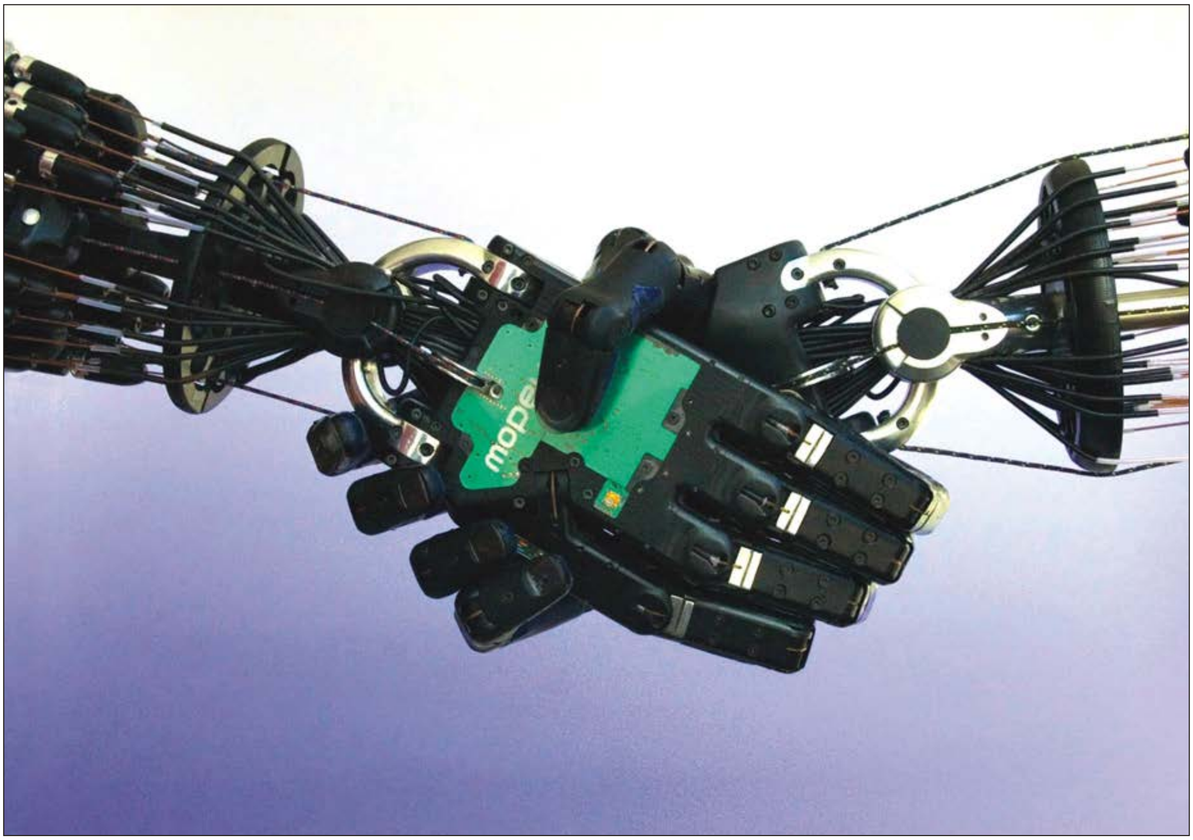

Journal of Artificial Intelligence

ISSN 1994-5450

Journal of Artificial Intelligence 4 (4): 279-287, 2011
ISSN 1994-5450 / DOI: 10.3923/jai.2011.279.287

A Neural Network Approach for Updating Ranked Association Rules, Based on Data Envelopment Analysis

Aliakbar Niknafs and Soodabeh Parsa
Department of Computer Engineering, Shahid Bahonar University of Kerman, Iran

Corresponding Author: Aliakbar Niknafs, Department of Computer Engineering, Shahid Bahonar University of Kerman, Iran

ABSTRACT

Mining association rules is one of the important tasks in data mining that finds sets of items which come together in many transactions. Ranking the association rules is very important in market basket analysis and decision making. Applying data envelopment analysis for finding the candidate rules and selecting efficient association rules has been an interesting research field in recent years. In this study, we propose a new method for updating the ranked table of association rules, when new transactions are added to the market basket. We apply artificial neural networks for refreshing the ranking table, this prevents repeating all the process of solving the linear programming problem by data envelopment analysis. An illustrative example is presented and the results are compared with the results of an earlier research.

Key words: Artificial neural networks, data envelopment analysis, association rules, market basket, data mining

INTRODUCTION

In recent years , the field of data mining has seen an explosion of interest from both academia and industry (Olafson *et al*., 2008) and a majority of researchers and managers have made use of some forms of data mining to make critical decisions. The reason is the rapid growth in the amount of information stored in databases.

As an information method, data mining is a technology to analyze a large amount of accumulated data to obtain information and knowledge valuable for decision making (Yun *et al*., 2003).

One of the main objectives of data mining is finding hidden patterns of relationship in a dataset. AR illustrate the relationship between transactions in different fields like market basket analysis. The resulting rules are helpful and interesting for decision makers in the market and also have noticeable effects on recommending the clients to purchase more items according to their preferences and needs. Therefore, finding Association Rules (AR) is an important objective in data mining.

The problem of discovering association rules has received considerable research attention and several fast algorithms for mining association rules have been developed (Srikant *et al*., 1997). Using these techniques, 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).

Given a set of items I, a rule in the form of x⇒y is an association rule when x⇒I and y⇒I. An association rule shows the interesting relationship among items in the dataset of transactions. The main result of AR is that if x is present in a transaction, there is a high possibility of presence of y in the same transaction. The support factor of a rule x$_\rightarrow$y is defined as the percentage of transactions which contain both x and y. The confidence of a rule x$_\rightarrow$y is the ratio of the number of transactions containing both x and y over the number of transactions containing x.

The minimum support and minimum confidence of an association rule are two factors that determine the strength of a rule. There are two main steps in the process of mining the AR. In the first step, the complete set of all itemsets whose support are greater than or equal to the minimum support, is identified. These combinations of items are called frequent items. In the second step, every non-empty proper subset of each frequent itemset is examined whether it can be the antecedent of an association rule.

Apriori is a well known algorithm for finding frequent itemsets (Shin and Lee, 2008). When a set of new transactions is added to a dataset, it is necessary to update the association rule. Generally, the algorithms utilize the previous results of AR in finding the up-to-date set of frequent itemsets.

Data Envelopment Analysis (DEA) is a linear programming based method for assessing the relative efficiency of decision making units (DMUs) (Allen and Thanassoulis, 2004). Since the original DEA study by Charnes *et al.* (1978), a rapid growth in the field has occurred. In the original DEA model, the weights are to be estimated in order to maximize the efficiency rating of DMU. The main constraint considered is that all the weights should be greater than a small positive value of ε.

Indeed, DEA is a non-parametric method in which multiple inputs and outputs could be used to measure an entity's performance (Guo, 2009). Therefore, DEA is a mathematical programming technique aiming at the measurement of DMUs relative efficiencies (Charnes *et al.*, 1978; Cooper *et al.*, 2007; Zhou *et al.*, 2007; Wagner and Shimshak, 2007).

Artificial Neural Networks (ANNs) have been extensively used in many fields to model complex real-world problems. Neural networks discover hidden patterns and relationships in large amounts of data by simulating the human brain in a computer system. An important property of ANNs is that, when correctly trained, they can appropriately process data that have not been used for training.

DATA ENVELOPMENT ANALYSIS

Data Envelopment Analysis (DEA) is an approach for evaluating the performance of a group of entities referred to as decision making units (DMUs). The DEA was introduced by Charnes *et al.* (1978). In their original DEA model, named as CCR, they proposed that the efficiency of a DMU can be obtained as the maximum of a ratio of weighted outputs to weighted inputs, subject to the condition that the same ratio for all DMUs must be less than or equal to one (Toloo *et al.*, 2009).

Several DEA models have been developed by researchers. We concern on two models developed by Cook and Kress (1990) and Obata and Ishii (2003). These two models were the basis of a research by Chen (2007) for ranking association rules. The DEA model proposed by Cook and Kress (1990) is as follows:

$$\text{Max } \sum_{j=1}^{K} w_j v_{oj} \tag{1}$$

s.t.

$$\sum_{j-1}^{\kappa} w_j v_{ij} \le 1, i=1,2,......,m \tag{2}$$

$$w_j - w_j +_1 \ge d(j,\varepsilon)\, j=1,2,..........\kappa-1 \tag{3}$$

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

where, W_j denotes the weight of the jth place; V_{ij} presents the number of the jth place votes of candidate i and d (.,ε), known as the discrimination intensity function, is nonnegative and non-descending in ε and satisfies d (.,0) = 0.

Model (1) should be resolved for each candidate o, o = 1,2,...,m. The resulting objective value is the preference score of the candidate o. Because of generating several efficient candidates by DEA (Cook and Kress, 1990), Chen proposed a method that uses another DEA model, proposed by Obata and Ishii (2003). This model employs only efficient association rules and is used for discriminating efficient rules.

The DEA model proposed by Obata and Ishii is as follows:

$$1/z = \text{Minimize}\,\|w\| \tag{5}$$

s.t.

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

$$\text{for all efficient io} \tag{7}$$

$$w_j - w_{j+1} \ge dj(j,\varepsilon),\ j=1,2,........\kappa-1 \tag{8}$$

$$W_j \ge d\ (k,\ \varepsilon) \tag{9}$$

The normalized preference score Z is obtained as a reciprocal of the optimal value. When L1-norm is used, the objective function becomes and then the model is a linear programming model.

Association rules: Association Rules (AR) (Agrawal *et al.*, 1993) are one of the important methods of data mining. AR tell you something about database that you did not already know. A market basket is a collection of items purchased by a customer in a single transaction. One common analysis is to find the set of items that come together in many transactions.

Let x and y be two itemsets. An association rule is an expression in the form x⇒Y where x is called the antecedent and y is the consequent of the rule. An example of an association rule is: 30% of transactions containing jam also contain butter, 4% of all transactions contain both jam and butter. In this example, 30% is called the confidence of the rule jam and 4% is the support of the rule.

If support is higher than a user-defined threshold, that itemset will be interesting. The confidence denotes the strength of implication and support indicates the frequency of the patterns occurring in the rule. Rules with high confidence and support are called strong rules. The aim of data mining in this field is to discover strong rules and then rank them in descending order.

Related works: By using data envelopment analysis (Charnes *et al.*, 1978) and (Cook and Kress, 1990) have proposed a method for estimating preference scores to analyze ranked voting data. The problem of ranked voting data arises when vectors select and rank more than one candidate with order of preference (Obata and Ishii, 2003). DEA often suggests that more than one unit are equally efficient. Therefore, a method for discriminating these efficient candidates is needed.

Obata and Ishii (2003) proposed a method that does not use information about inefficient candidates to discriminate efficient candidates; therefore the order of efficient candidates may not be changed by an inefficient candidate. In addition to support and confidence, domain knowledge can be further designed as measures to evaluate the discovered rules.

Chen (2007) considered the product value and cross-selling profit as essential measures to rule interestingness. Chen applied an example of market basket analysis to illustrate the DEA methodology for measuring the efficiency of association rules with multiple criteria.

In Chen's proposed approach, after mining association rules by using the Apriori algorithm with minimum support and confidence, the preference of rules were calculated by using Cook and Kress's DEA model. Obata and Ishii's model was used to discriminate the efficient rules. Finally, the rules were selected for implementation.

Proposed method: When there are a large number of rules in data set, the interestingness of a rule can be used to filter them and report only those which may be useful to decision makers (Mitra *et al.*, 2002). The thresholds of support and confidence are selected by only considering the database perspective. However, the interestingness of an rule is commonly application-dependent (Srikant *et al.*, 1997).

The idea of efficiency in DEA is a comparative concept (Serranto-Cinca *et al.*, 2005). Chen (2007) adopts Obata and Ishii's discriminate model (5)-(9), to discriminate the efficient candidates, which are generated in Cook and Kress's DEA model (1)-(4). The discriminate model can be used to identify one winner(the most favorable association rule) if such a winner exists.

The Chen's method has the following properties (Toloo *et al.*, 2009): it is time consuming, and the results of chen's method are immensely dependent on discrimination intensity function.

In market transactions , this is a fact that some new rules are generated every day and this makes necessary to update and refresh the ranking list of efficient and selected rules. But a question arises here: when a new rule is received to market basket database, is it necessary to repeat all the steps mentioned in Chen's method for inserting that rule in the ranked table?

In our proposed method the answer to this question is negative. The last step of the proposed model includes an artificial neural network that is trained through the previous entries of efficient ranked table. When the new rule comes, it is entered to the trained ANN model and the preference score (Z_i) will be found at the output. The value of Z'_i determines the position of the new rule among the previous rules of ranking table. According to the above mentioned discussion, the proposed approach is schematically illustrated in the flowchart in Fig. 1.

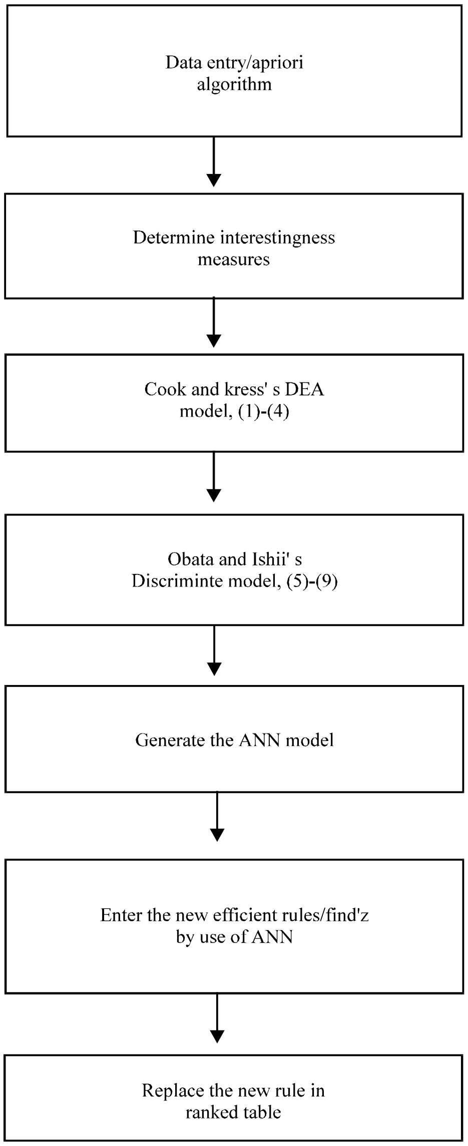
Data entry/apriori algorithm
Determine interestingness measures
Cook and kress' s DEA model, (1)-(4)
Obata and Ishii' s Discriminte model, (5)-(9)
Generate the ANN model
Enter the new efficient rules/find'z by use of ANN
Replace the new rule in ranked table

Fig. 1: Flowchart of the proposed updating algorithm

The proposed approach is described as follows:

Step 1: Input data and mine association rules by using the Apriori algorithm with minimum support and confidence

Step 2: Determine subjective interestingness measures by further considering the domain related knowledge

Step 3: Calculate the preference scores of association rules discovered in step 1 by using Cook and Kress's DEA model, (1)-(4)

Step 4: Use Obata and Ishii's discriminate model, (5)-(9) to discriminate the efficient association rules. The ranked list of rules is created by comparing the preference scores (Z'_i)

Step 5: Generate the ANN model and train it by using the ranked table

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

Rule No.	Supp. (%)	Conf. (%)	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	354.00	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

Using one hidden layer with alpha: 0.9, initial eta: 0.3, high eta: 0.1, low eta: 0.01 and eta decay:30

Step 6: Enter the new rules from market basket data base, find preference score (Z'_i) and determine the position of new rule in the list. Insert the new rule .The updating procedure of ranked list is complete now

Illustrative example: An example of market basket data is adopted from Chen (2007) to illustrate the applicability of the proposed method. First, association rules are discovered by the Apriori algorithm, where the minimum support and minimum confidence are set to 1 and 10%, respectively. Then, forty-six rules are selected. The itemset values and cross-selling profits for these 46 rules are then calculated and presented in Table 1.

Table 1 reveals 11 efficient association rules with preference score 1 (bold rows). The preference scores (Z_i) are calculated by Cook and Kress's DEA. These 11 efficient rules are further analyzed using Obata and Ishii's model and the preference scores (Z'_i) are obtained. According to Z'_i these 11 rules are ranked in descending order in Table 2. The results of ranking shown in Table 2 are calculated with respect to four criteria including two subjective measures of itemset value and cross-selling profit in addition to support and confidence.

In the proposed method , it is not needed to repeat all of the above mentioned calculations for the new rule. We produce the ANN model using the data set of Table 2 and when a new rule is received, it is entered to the model. The resultant Z'_i predicted by ANN determines the rank of the rule. In this example, the ANN model is produced.

Table 2: Summary 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

Table 3: Updated ranking using the proposed method

Rule no.	Supp. (%)	Conf. (%)	Itemset value	Cross-selling profit	Preference Score (Z'_i)	ANN output of 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		306.118
17	1.19	50.92	436.00	139.02	164.95	
101	1.43	39.00	610.00	101.00		2.723
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	
102	1.50	38.00	620.00	100.00		0.403
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	
14	2.44	15.22	554.00	41.56		0.100

Assume the rule 18 is not entered into Table 2. After entering this rule to ANN model we obtain Z'_i = 306.118, as shown in Table 3. This causes to place rule 18 after rule 22, that is just similar to the result of Chen's method.

Again consider rule 14. Entering this rule to ANN model results Z'i = 0.100, that puts the rule 14 in the last rank after the rule 12.

Suppose that a new rule, named rule 101 is entered to the dataset. In order to insert it to the ranking table, the Z'_i is obtained using the ANN model. The resultant value Z'_i=2.723 causes to put this rule after rule 17 and before rule 7. Similarly another imaginary rule 102 is entered and Z'_i=0.403 is calculated that places this rule between rules 6 and 43.

CONCLUSIONS

Using the values of minimum support and confidence, the efficient candidates can be determined by means of DEA. According to other criteria like item value and cross-selling profit, the interesting rules can be determined by using another DEA model that considers only the efficient candidate rules. Then ranking the rules is possible. When new transactions are added to the market basket, it is needed to update the ranked list of rules. Applying ANN prevents repeating all the procedure of DEA linear programming model. This leads to more time efficiency and speed of refreshing the ranked lists. In future works, the instant based reasoning (case based reasoning) could be examined for achieving this goal. Also the constraints of DEA could be changed so that faster approaches be achieved.

REFERENCES

Agrawal, R., T. Imielinski and A. Swami, 1993. Mining association between sets of items in massive database. Proceedings of the ACM SIGMOD International Conference on Management of Data, June 1, 1993, ACM New York, USA., pp: 207-216.

Allen, R. and E. Thanassoulis, 2004. Improving envelopment in data envelopment analysis. Eur. J. Oper. Res., 154: 363-379.

Charnes, A., W.W. Cooper and E. Rhodes, 1978. Measuring the efficiency of decision-making units. Eur. J. Oper. Res., 2: 429-444.

Chen, M.C., 2007. Ranking discovered rules from data mining with multiple criteria by data envelopment analysis. Expert Syst. Appl., 33: 1110-1116.

Cook, W.D. and M. Kress, 1990. A data envelopment model for aggregating preference rankings. Manage. Sci., 36: 1302-1310.

Cooper, W.W., J.L. Ruiz and I. Sirvent, 2007. Choosing weights from alternative optimal solutions of multipliers in DEA. Eur. J. Oper. Res., 180: 443-458.

Guo, P., 2009. Fuzzy data envelopment analysis and its application to location problem. Inform. Sci., 179: 820-829.

Mitra, S., S.K. Pal and P. Mitra, 2002. Data mining in soft computing framework: A survey. IEEE. Trans. Neural Networks, 13: 3-14.

Obata, T. and H. Ishii, 2003. A method for discriminating efficient candidates with ranked voting data. Eur. J. Oper. Res., 151: 233-237.

Olafson, S., X. Li and S. Wu, 2008. Operation research and data mining. Eur. J. Oper. Res., 187: 1429-1448.

Serranto-Cinca, C., Y. Fuertes-Callen and C. Mar-Molinero, 2005. Measuring DEA efficiency in internet companies. Decis. Support Syst., 38: 557-573.

Shin, S.J. and W.S. Lee, 2008. On-line generation association rules over data streams. Inform. Software Technol., 50: 569-578.

Srikant, R., Q. Vu and R. Agrawal, 1997. Mining association rules with item constraints. Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Augast 14-17, 1997, California, USA., pp: 66-73.

Toloo, M., B. Sohrabi and S. Nalchigar, 2009. A new method for ranking discovered rules from data mining by DEA. Expert Syst. Appl.: Int. J., 36: 8503-8508.

Wagner, J.M. and D.G. Shimshak, 2007. Stepwise selection of variables in data envelopment analysis, procedures and managerial perspectives. Eur. J. Oper. Res., 180: 57-67.

Yun, H., D. Ha, B. Hwang and K.H. Ryu, 2003. Mining association rules on significant rare data using relative support. J. Syst. Softwares, 167: 181-191.

Zhou, P., K.L. Poh and B.W. Ang, 2007. A non-radial DEA approach to measuring environmental performance. Eur. J. Oper. Res., 178: 1-9.