The Sixteenth International Symposium on Artificial Life and Robotics 2011 (AROB 16th '11),
B-Con Plaza, Beppu,Oita, Japan, January 27-29, 2011

Discriminate approach for data selection in data envelopment analysis

Akio Naito and Shingo Aoki

Osaka Prefecture University, 1-1 Gakuencho Naka-ku Sakai Osaka
(Tel : 81-72-254-9353; Fax : 81-72-254-9915)
({naitou@mis., aoki@}cs.osakafu-u.ac.jp)

Abstract: DEA (Data Envelopment Analysis) is a well-known method for evaluating management efficiency of DMUs (Decision Making Units). To calculate efficiency of DMUs, analytical data are necessary. However, there are not clear criteria for data selection so that analysts have to choose the data on their own. Therefore, it is important to support data selection by reasonable ways to let analysis be informative and beneficial. In order to deal with this matter, new methods are proposed based on traditional ones. Support for data selection is realized by considering analyst's intention. Analytical data for making some specific DMUs efficient are obtained by reflecting knowledge or experience analysts have. TDS-DEA (Tight Data Selection based DEA) reflects the analyst's intention strongly and tries to make only intended DMUs efficient. On the other hand, LDS-DEA (Loose Data Selection based DEA) reflects it loosely and at least intended DMUs can be efficient. Then both methods should be examined more detail and how data selection is carried out effectively. On this point, this study prepares the experimental data to clarify the effectiveness and drawback of the methods. According to the experimental result, additional ideas such as discriminate approach or assurance region method are considered to improve the quality of data selection.

Keywords: Data Envelopment Analysis, Linear Programming, Decision Making Support, Data Selection

I. INTRODUCTION

DEA (Data Envelopment Analysis) is a method for measuring efficiency of DMUs (Decision Making Units) like company, hospital, municipal government or so. DMUs are evaluated by index called "efficiency score". Each DMU is classified as the state of efficient or that of inefficient based on the score [1]. DEA calculates efficiency score of each DMU based on Pareto optimal line which is called efficiency frontier consists of efficient DMUs. Then DEA shows a plan for improvement to inefficient DMUs.

DEA assumes activity of DMUs that produce output from input. This mechanism is interpreted as production function. DEA is a data-oriented method so that efficiency score depends on analytical data. Then all data related to DMUs can be possible to be selected for analysis. Therefore, data selection is really important. It is not easy to prove whether selected data correspond to purpose of analysis or not. Criteria of data selection are unclear practically. Hence, some methods were proposed to support data selection. The basic idea of the traditional methods is to utilize analyst's intention such as experience or knowledge regarding evaluated DMUs.

Though numerical experiment is carried out, but still need more trial to clarify the power and effectiveness. Therefore, the purpose of this study is to examine the traditional method in detail and find benefit and drawback to improve the approach.

II. DATA ENVELOPMENT ANALYSIS

1. Outline

DEA was proposed by A. Charnes et al. in 1978 as a method for management analysis [2]. DEA has room to treat a lot of data related to DMUs. Then necessary elements (capital, employee, etc) for operation are generally recognized as input and yielded elements (sales, profit rate, etc) are recognized as output. DEA calculates efficiency by input and output so that less input and larger output is more preferable. And the efficiency of each DMU is evaluated relatively among analyzed DMUs.

Efficient DMUs are regarded as best practice among DMUs and they get efficiency score as "one". Inefficient DMUs have that score less than one. Efficiency score is denoted as θ and calculated by dividing virtual output by virtual input. Virtual input and output are the useful for dealing with multi elements. DEA puts weight to each element and it is not fixed but variable. Therefore, it is possible to evaluate advantages of DMUs as much as possible [1].

©ISAROB 2011

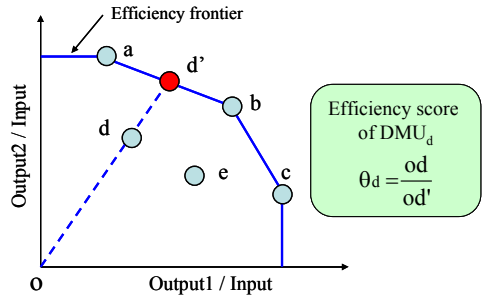

Fig.1. Efficiency frontier and efficiency score

Fig. 1 demonstrates efficiency frontier and efficiency score. There are five DMUs (a~e) with one input and two outputs. Each DMU is dotted based on score that output is divided by input. Hence, DMU located far from origin is more efficient. Here three DMUs are the state of efficient and they form efficiency frontier. DMU_c has largest output1 and DMU_a has largest output2. In addition, DMU_b has larger output1 and output2 with well-balanced. Characteristics of these three DMUs are evaluated well respectively. Thus DEA is able to reflect strength of DMUs for evaluation. Therefore, analytic data play an important role to extract characteristic of each DMU.

Efficiency score of inefficient DMU is calculated by distance to efficiency frontier that expresses ideal state. In case of DMU_d, ideal state of management is d'. Then efficiency score θ_d is calculated by the ratio of distance from origin to d and d'. In other words, efficiency score is calculated based on efficient DMUs among analytical objects.

2. Formulization

When analysis is carried out, linear programming is utilized. Here formulization of CCR model is shown. Assuming that there are *n* DMUs ($DMU_1, DMU_2, \cdots, DMU_n$) with *m* inputs and *s* outputs. DMU_k is characterized by inputs ($x_{1k}, x_{2k}, \cdots, x_{mk}$) and outputs ($y_{1k}, y_{2k}, \cdots, y_{sk}$). Then efficiency score of DMU_k is calculated by following formula.

$$\begin{aligned} &\text{Max} \sum_{r=1}^{s} u_r y_{rk} \\ &\text{s.t.} \ -\sum_{i=1}^{m} v_i x_{ij} + \sum_{r=1}^{s} u_r y_{rj} \le 0 \ \ (j=1,2,\cdots,n) \\ &\qquad \sum_{i=1}^{m} v_i x_{ik} = 1 \\ &\qquad v_i \ge 0 \ \ (i=1,2,\cdots,m), \ u_r \ge 0 \ \ (r=1,2,\cdots,s) \end{aligned} \tag{1}$$

Inputs and outputs are denoted as x_{ij}, y_{rk} while v_i, u_r are weights for input and output elements. Hence, $v_i x_{ij}$ and $u_r y_{rj}$ represent virtual input and output. One of the constraints works for virtual input of DMU_k to be "one". Then virtual output is maximized on condition that virtual input exceeds that of output for each DMUs. If virtual output is equal to one, DMU_k is the state of efficient. On the other hand, DMU_k is the state of inefficient if virtual output is less than one. Thus objective function, namely, virtual output signifies efficiency score directly.

As a result, efficiency score and weight value are shown by solving linear programming. Then weight that has value means corresponding input and output elements are employed in analysis. That is why DEA enables analyst to know strength of each DMU.

III. TRADITIONAL METHOD

1. Outline

The previous study focused on data selection for making intended DMUs efficient [3]. That is because some of the analysts are quite familiar with analyzed DMUs and they can predict result roughly. For beginners, deciding efficient DMUs they want is helpful to have reasonability of analytic data elements. Of course, it is possible to incorporate external information to the data selection.

Traditional methods need to take inverse procedure compared with fundamental DEA. It means result that some DMUs are efficient is assumed first and then used data are calculated based on the assumption. Fig. 2 shows stream of traditional method. Dotted line is route concerning fundamental DEA and procedure consists of data collection, analysis and result. However, traditional method assumes the result in advance and then data elements are calculated in order to guarantee assumed result (analyst's intention).

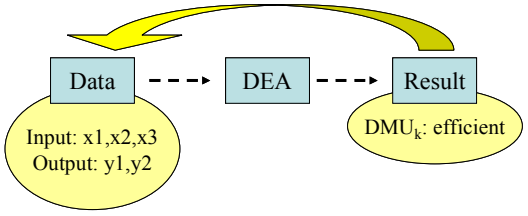

Fig.2. Procedure of the method

While idea of traditional method is formed, we consider possible users, namely, analysts who utilize the methods. There are mainly three types of users. Expert who has knowledge to DMUs, decision maker who

belongs to analyzed DMU, and beginners. The merits of the methods are mentioned for every type of user.

For experts, they have forecast for result based on their deep knowledge. That is why they can know effective data elements to realize their intentional analysis thanks to traditional methods. So it is effective for analyzing DMUs. For decision makers, they can find strengths of their company and competitive one by assuming those two DMUs as efficient. That is to say, management strategy can be planned by DEA since they know characteristic in detail for themselves and rival. For beginners, they are often confused when they decide data. Then traditional methods can show necessary data elements for them and support smooth analysis. Thus various analysts are able to get benefit by utilizing the method.

2. TDS-DEA

TDS-DEA (Tight Data Selection based on DEA) calculates data elements that make only intended DMUs efficient. Hence, it is possible to decrease the number of elements in analysis. TDS-DEA introduces condition for slack variable that treat surplus of input and lack of output. If slack variable has value, the DMU is the state of inefficient. On the other hand, the DMU is the ideal state if slack variable is zero. Thus only specific DMUs can be efficient and others can be inefficient. Here analyst would like to make DMU_k efficient. Then slack variable should be satisfied with following conditions;

$$\begin{aligned} &s_j = 0 \ \ (j=k) \\ &s_j > 0 \ \ (j=1,2,\cdots,n: j\neq k) \end{aligned} \tag{2}$$

Only slack variable of DMU_k is zero and others have value. Traditional method enables analyst's intention to be reflected strongly with these conditions. And the used elements are found by the weights v_i, u_r. If weight has value, corresponding element is used for analysis. Therefore, weight is a key to know which elements should be selected. Formula (3) is a linear programming regarding TDS-DEA. Assume that there are *n* DMUs with *m* inputs and *s* outputs. {J-k} signifies set including DMUs except for DMU_k. In order to realize concept of TDS-DEA, objective function are maximized. It means TDS-DEA tries to decrease efficiency of DMUs except for intended one as much as possible with keeping intended DMU efficient. When multi DMUs are assumed as efficient, elements are calculated by repeating analysis for each DMU.

$$\begin{aligned} \text{Max} \quad & \sum_{j\in\{J-k\}} s_j \\ \text{s.t.} \quad & -\sum_{i=1}^{m} v_i x_{ik} + \sum_{r=1}^{s} u_r y_{rk} = 0 \\ & -\sum_{i=1}^{m} v_i x_{ij} + \sum_{r=1}^{s} u_r y_{rj} + s_j = 0 \ \ (j=1,2,\cdots,n: j\neq k) \\ & \sum_{i=1}^{m} v_i x_{ik} = 1 \\ & v_i \geq 0 \ \ (i=1,2,\cdots,m), \ u_r \geq 0 \ \ (r=1,2,\cdots,s) \end{aligned} \tag{3}$$

3. LDS-DEA

As long as data selection relies on TDS-DEA, we need to care the accuracy of analyst's intention. That is because intended DMU surely get the state of efficient after calculation. At the same time extension of the method was necessary. It is to analyze DMUs by employing common elements for intended DMUs.

- The number of DMUs : n
- The number of intended DMUs : α
- Combination for choosing two DMUs from intended DMUs : h ($1, 2, \cdots, {}_{\alpha}C_2$)
- DMU No. of t-th DMU among intended ones : q_t ($t=1,2,\cdots,\alpha$)

$$\begin{aligned} \text{Min} \quad & \sum_{h=1}^{{}_{\alpha}C_2} \{\sum_{i=1}^{m} (d_{ih}^{x+} + d_{ih}^{x-}) + \sum_{r=1}^{s} (d_{rh}^{y+} + d_{rh}^{y-})\} \\ \text{s.t.} \quad & -\sum_{i=1}^{m} v_i^t x_{iq_t} + \sum_{r=1}^{s} u_r^t y_{rq_t} = 0 \ \ (t=1,2,\cdots,\alpha) \\ & -\sum_{i=1}^{m} v_i^t x_{ij} + \sum_{r=1}^{s} u_r^t y_{rj} + s_j^t = 0 \ \ (j=1,2,\cdots,n: j\neq t) \ \ (t=1,2,\cdots,\alpha) \\ & \sum_{i=1}^{m} v_i^t x_{iq_t} = 1 \ \ (t=1,2,\cdots,\alpha) \\ & v_i^k x_{iq_k} - v_i^l x_{iq_l} + d_{ih}^{x+} - d_{ih}^{x-} = 0 \\ & u_r^k y_{rq_k} - u_r^l y_{rq_l} + d_{ih}^{y+} - d_{ih}^{y-} = 0 \\ & (i=1,2,\cdots,m), (r=1,2,\cdots,s), (k,l\in t: k<l), (h=1,2,\cdots,{}_{\alpha}C_2) \\ & v_i^t \geq 0, \ v_i^t \geq 0 \ \ (i=1,2,\cdots,m) \\ & u_r^t \geq 0, \ u_r^t \geq 0 \ \ (r=1,2,\cdots,s) \end{aligned} \tag{4}$$

IV. NUMERICAL EXPERIMENT

To confirm the effectiveness, experimental data that has some features is prepared. There are twenty DMUs with six inputs and six outputs. The feature is that DMU_1 to DMU_6 have strong points in input1, input2, output1, and output2. DMU_7 to DMU_{12} have strength in input5, input6, output5, and output6. DMU_{15} to DMU_{20} has strength in input3, input4, output3, and output4. Then TDS-DEA is applied and calculates the data element to realize analyst's intention. Table. 1 shows the experimental data in this study.

Table 1. Experimental data

Element		DMU									
		DMU_1	DMU_2	DMU_3	・・・	DMU_{10}	DMU_{11}	・・・	DMU_{18}	DMU_{19}	DMU_{20}
Input	x1	0.111	0.222	0.444	・・・	0.444	0.556	・・・	0.556	0.667	0.444
	x2	0.778	0.667	0.444	・・・	0.778	0.333	・・・	0.778	0.556	1
	x3	0.897	0.691	0.918	・・・	0.804	0.763	・・・	0.433	1	0.351
	x4	0.545	0.364	0.852	・・・	0.523	0.511	・・・	1	0.477	0.273
	x5	0.311	0.378	0.333	・・・	0.200	0.244	・・・	0.622	0.711	0.422
	x6	0.317	0.293	0.610	・・・	0.244	0.512	・・・	0.976	0.512	0.659
Output	y1	0.778	1	0.889	・・・	0.556	0.222	・・・	0.243	0.220	0.060
	y2	0.625	0.500	0.875	・・・	0.500	0.375	・・・	0.875	0.500	0.250
	y3	1	0.188	0.800	・・・	0.350	0.800	・・・	0.753	0.753	0.494
	y4	0.351	0.485	0.897	・・・	0.472	0.485	・・・	1	0.763	0.371
	y5	0.899	0.528	0.876	・・・	0.270	0.506	・・・	0.461	0.360	0.247
	y6	0.214	0.827	0.459	・・・	0.663	0.867	・・・	0.337	0.306	0.112

Table 2. Result

Weight	DMU											
	DMU_1	DMU_2	DMU_3	DMU_4	DMU_9	DMU_{10}	DMU_{11}	DMU_{12}	DMU_{15}	DMU_{16}	DMU_{17}	DMU_{18}
v1	9.000	4.500	0	1.446	0	0	0	0	0	0	0	0
v2	0	0	0	0.804	0	0	0	0	0	3.000	0	0
v3	0	0	0	0	2.243	0	0	0	4.042	0	0	2.310
v4	0	0	0	0	0	0	0	0	0	0	6.286	0
v5	0	0	3.000	0	0	5.000	4.091	0.285	0	0	0	0
v6	0	0	0	0	0.626	0	0	2.854	0	0	0	0
u1	0	0.464	0	0	0	1.800	0	0	0	0	0	0
u2	0	0	0	1.143	0	0	0	0	0	0	0.766	0
u3	1.000	0	0	0	0	0	1.250	0	0	1.149	1.542	0
u4	0	0	1.115	0	0	0	0	0	1.738	0	0	1.000
u5	0	0	0	0	1.000	0	0	0	0.027	0	0	0
u6	0	0.648	0	0	0	0	0	1.000	0	0	0	0

Table. 2 shows the result for the input and output should be chosen. Shaded area is that each DMU's strong elements. According to the result, we find some knowledge through experimentation.

Let us focus on DMU_1 to DMU_4. They originally have strong input1, input2 and then TDS-DEA actually calculates those elements. The method signifies input5 is strong for DMU_3. That is because DMU_3 has larger input compared with other DMUs and choosing input5 is inevitable for individual efficiency. However, data selection for output is not enough since obtained result does not reflect the actual strength DMUs have. This is the drawback of the traditional method. That is to say, it is possible to get necessary data roughly but sometimes it is not reliable and accuracy. In this point, introducing assurance region method is helpful to improve the method. Also discriminate approach will have influence to data selection better.

The traditional method calculates some unsuitable element for other DMUs though they get desirable result. Analyst is able to set the analytical data based on certain reasonability and confidence thanks to the method. And it is beneficial for not only analysts who are expert but also analysts who don't have enough knowledge regarding evaluated DMUs. if analyst have problem for data selection, they just utilize the method and get the direction for their analytical procedure.

V. CONCLUSION

This paper examines the power of traditional method for data selection. The result shows how the method calculates necessary data. However, it is important to improve quality of the approach since some of the data are not calculated well based on analyst's intention. Then we consider introducing assurance region method or discriminate way in order to complement current method.

REFERENCES

[1] W. W. Cooper, L. M. Seiford and K. Tone (2007), Data Envelopment Analysis A comprehensive Text with Models, Applications, References and DEA-Solver Software Second Edition. Springer

[2] A. Charnes, W. W. Cooper and E. L. Rhodes (1978), Measuring the Efficiency of Decision Making Units. European Journal of Operational Research, vol. 2, pp. 429–444

[3] A. Naito and S. Aoki (2010), Support for Awareness of Data Selection in Data Envelopment Analysis. The 2nd International Symposium on Aware Computing (ISAC 2010)