Australian Journal of Basic and Applied Sciences, 6(2): 158-163, 2012
ISSN 1991-8178

Application of Rough Set Theory in Performance Analysis

[1]Mahnaz Mirbolouki, [2]Mohammad Hassan Behzadi, [1]Leila Karamali

[1]Department of Mathematics, Shahre-Rey Branch, Islamic Azad University, Tehran, Iran.
[2]Department of Statistics, Science and Research Branch, Islamic Azad University, Tehran.

Abstract
Data envelopment analysis (DEA) is a mathematical technique based on linear programming for evaluating the efficiency of a set of decision making units (DMUs). Every DMU use several inputs to produce several outputs. In order to derive the efficiency values from DEA models meaningfully, it is conventionally assumed that three times the total number of inputs and outputs factors is less than or equal to the number of units. But some practical issues contain large number of these factors. Thus providing a method that can reduce the number of these factors is felt necessary. While this paper is reviewing some preliminary relationships of Rough Set Theory (RST), it is mentioning the application of RST in data envelopment analysis. One of the applications is reducing the number of inputs and outputs. Also, a numerical example is provided to show implementation of this method.

Key words: Data envelopment analysis (DEA), Performance evaluation, Rough set theory (RST).

INTRODUCTION

DEA, a mathematical technique based on linear programming first introduced by Charnes et al., 1978, is a way of determining the efficiency for a group of decision making units (DMUs) when measured over a set of multiple input and output variables. For a given set of input and output variables, DEA produces a single comprehensive measure of performance (efficiency score) for each DMU.

Rough set theory (RST) was initiated in the early 1980s by Pawlak, 1982. This theory deals with the analysis of data tables which are called information systems. The data can be quantitative or qualitative.

Usually in the real issues, several pieces of information are inaccurate, incomplete or unreliable. Therefore in order to drive suitable conclusions related to such information, the information must be processed first. Rough sets theory and fuzzy sets theory are two well-known tools for processing different types of inaccurate, unreliable and ambiguous data. However, these theories are included different concepts. RST is mainly differed in comparison with exact and fuzzy sets theories by membership function definition. In the usual set theory, a set of accurate data is uniquely identified with its members. Membership function describing the elements of the reference set is only getting zero or one values. The Fuzzy set theory in which data is dealing with uncertainty, the set membership function get values of [0,1] interval. However, Rough sets theory membership is not common concept. Rough sets offer a different approach for vague and uncertain. Definition of a set in RST relates to the available information and relations between data in the information system. Members are specified with relevant to information and properties. Thus, two different members of a set may not be distinct clearly. Rough set can be introduced as a framework for discovering facts from imperfect data. Rough set results provide categories or decision rules which are derived from a set of samples.

The main purpose in the analysis of RST is obtaining conceptual approximate of acquired data. While this approximation of a crisp set contains a pair of sets which give the *lower* and the *upper* approximation of the original set. In the usual version of RST, the lower- and upper-approximation sets are crisp sets, but in other variations, the approximating sets may be fuzzy sets. RST is a powerful mathematical tool for uncertainty reasoning helps to gain insight into the problem at hand by analyzing the constructed model. RST provides procedures for removing and reducing additional information or irrelevant knowledge of the database. This process eliminates unnecessary data from the main task of the system without losing basic data. The reduced data set makes decision much easier. Therefore, given the explosive growth of data volume, RST can be very effective in Decision Support Systems. (see e.g. Pawlak, 1991; Polkowski and Skowron, 1998; Polkowski, 2002).

Counterintuitively, using too many inputs and outputs in DEA will be less helpful because when the number of inputs and outputs increases, more decision-making units tend to get an efficiency score of 1 as they become too specialized to be evaluated with respect to other units. A rule of thumb is that there should be a minimum of three funds per input and output in implementing a DEA model (Bowlin, 1998; Raab and Lichty, 2002). Thus, for practical reasons, there needs to be some limit on the number of inputs and outputs. In this paper, considering the properties of Rough set theory in the criteria classification and distinguishing the criteria

Corresponding Author: Mahnaz Mirbolouki, Department of Mathematics, Shahre-Rey Branch, Islamic Azad University, Tehran, Iran.
E-mail: m.mirbolouki@srbiau.ac.ir.

minimal sets, along with providing an example the implementation of RST in input and output reduction is proposed. By this approach the unnecessary inputs and outputs can be discovered.

The remainder of this paper is organized as follows: preliminaries of RST concepts are described in section 2 and in section 3 during a numerical example, the application of RST in input and output reduction is proposed. Section 4 provides the conclusion of the paper.

2-RST Preliminaries:

This section contains an explanation of the basic framework of rough set theory, along with some of the key definitions.

2-1 Information System:

Let $I = (\mathbb{U}, \mathbb{A})$ be an information system (attribute-value system), where $\mathbb{U}$ is a non-empty set of finite objects, $\mathbb{U} = \{x_1, x_2, ..., x_m\}$, and $\mathbb{A}$ is a non-empty finite set of attributes such that $\mathbb{U}: a \to V_a$ for every $a \in \mathbb{A}$. V_a is the set of values that attribute a may take. The information table assigns a value $a(x)$ from V_a to each attribute a and object x in the universe $\mathbb{U}$. For any $P \subseteq \mathbb{A}$ there is an associated equivalence relation $IND(P)$, indiscernibility relation,

$$IND(P) = \{(x, y) \in \mathbb{U}^2 \mid \forall a \in P, a(x) = a(y)\}.$$

The partition of $\mathbb{U}$ is a family of all equivalence classes of $IND(P)$ and is denoted by $\mathbb{U}/IND(P)$. If $(x, y) \in IND(P)$, then x and y are indiscernible (or indistinguishable) by attributes from P. Let $X \subseteq \mathbb{U}$ be a subset that we are going to represent using attribute subset P. In general, X cannot be expressed precisely, because the set may include and exclude objects which are indistinguishable on the basis of attributes P. However, the target set X can be approximated using the information which is involved in P by defining the lower and upper approximations of X:

$$\begin{aligned} \underline{P}X &= \{x \mid [x]_P \subseteq X\} \\ \overline{P}X &= \{x \mid [x]_P \cap X \neq \varnothing\} \end{aligned} \tag{1}$$

The tuple $\langle \underline{P}X, \overline{P}X \rangle$ is called a rough set. So, a rough set is composed of two crisp sets. The accuracy of the rough set representation of the set X is defined as follows:

$$\alpha_P(X) = \frac{\underline{P}X}{\overline{P}X} \tag{2}$$

Generally the upper and lower approximations are not equal. Thus target set X is indefinable or roughly definable on attribute set P. When the upper and lower approximations are equal, then the target set X is definable on attribute set P.

2-2 Reduct and Core:

Reducts and core indicate the attributes in the information system which are more important to the knowledge represented in the equivalence class structure than other attributes. reduct is a subset of attributes which can fully characterize the knowledge in the database. Formally, a reduct is a subset of attributes $RED \subseteq P$ such that

- $[x]_{RED} = [x]_P$, that is, the equivalence classes induced by the reduced attribute set RED are the same as the equivalence class structure induced by the full attribute set P.
- The attribute set RED is minimal, in the sense that $[x]_{(RED-\{a\})} \neq [x]_P$ for any attribute $a \in RED$, in other words, no attribute can be removed from set RED without changing the equivalence classes $[x]_P$.

A reduct can be considered as a sufficient set of features to represent the category structure. The set of attributes which is the intersection of all reducts is called core. Core is the set of attributes which is possessed by every legitimate reduct, and therefore consists of attributes which cannot be removed from the information system without causing collapse of the equivalence-class structure. The core may be thought of as the set of necessary attributes for the category structure to be represented.

2-3 Attribute Dependency:

One of the most important aspects of database analysis or data acquisition is identifying of attribute dependencies. I means the detection of the variables which are strongly related to other variables. For this purpose, let $[x]_Q = \{Q_1, Q_2, ..., Q_N\}$, where Q_i is a given equivalence class from the equivalence-class structure induced by attribute set Q. Thus, the dependency of attribute set Q on attribute set P, $\gamma_P(Q)$, is as following:

$$\gamma_P(Q) = \frac{\sum_{i=1}^{N} |\underline{PQ}_i|}{|\mathbb{U}|} \leq 1$$

That is, for each equivalence class Q_i in $[x]_Q$, the size of its lower approximation is added up by the attributes in P. Added across all equivalence classes in $[x]_Q$, the numerator above represents the total number of objects which is based on attribute set P can be positively categorized according to the classification induced by attributes Q. The dependency ratio indicates the proportion of such classifiable objects. The dependency $\gamma_P(Q)$ can be thought as a proportion of such objects in the information system for which it suffices to know the values of attributes in P to determine the values of attributes in Q.

3-Inputs and Outputs Reduction by RST:

In this section, a numerical example of an application is provided to show utilization of RST in reducing the number of input and output components.

Data envelopment analysis (DEA) initiated by Charnes *et al*. (1978), and the first model was called CCR model. This model can evaluate the relative efficiency of a set of DMUs, DMU_j ; $j = 1,...,n$, which use a vector of inputs, $x_j = (x_{1j},...,x_{mj}) \in R^{m+}$, to produce a vector of outputs, $y_j = (y_{1j},...,y_{sj}) \in R^{s+}$. Input oriented CCR model for efficiency evaluation of DMU_o ; $o \in \{1,...,n\}$ is as following:

$$\begin{aligned} &\min \quad \theta \\ &s.t. \quad \sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta x_{io}, \qquad i = 1,...,m, \\ &\qquad \sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{ro}, \qquad r = 1,...,s, \\ &\qquad \lambda_j \geq 0, \qquad j = 1,...,n. \end{aligned} \tag{3}$$

In model (3), θ in optimal solution means efficiency and it is a value between zero and one. If $\theta^* = 1$ then the under assessment DMU is an efficient DMU. Counterintuitively, using too many inputs and outputs in DEA will be less helpful because when the number of inputs and outputs increases, more DMUs tend to get an efficiency score of 1 as they become too specialized to be evaluated with respect to other units. A rule of thumb is that there should be a minimum of three funds per input and output in implementing a DEA model [1,7]. Thus, for practical reasons, there needs to be some limit on the number of inputs and outputs. Here, we consider an application of banking industry which is containing 20 bank branches with 4 inputs and 16 outputs. These inputs and outputs are gathered in Table 1 and Table 2 respectively. Since the number of inputs and outputs of each DMU with comparison of amount of DMUs is high, therefore all units are detected as efficient DMU after evaluation by model (3). To solve this problem we applied RST in inputs and outputs sets. In this approach DMUs are the objects and inputs and outputs (components) are attributes. Also, the classification of components are considered as the attributes values, V_a. These classifications are based on the uniform distributions. The classification of inputs and outputs in 4 groups are in Table 3.

Discernibility relation between DMUs (attributes) can be detected from Table 3 by comparing inputs and outputs of DMUs. For example DMUs 2, 3, 4, 5, 6, 8, 9, 11, 12, 13 are in a same class of inputs set (note that here we just consider inputs attributes in order to find reducts and cores of inputs). Therefore those DMUs which are not in a same class can discern each other. Discernibility matrix can be defined as $D = [d_{ij}]_{n \times n}$, where

$$d_{ij} = \{I_k \mid DMU_i \text{ can discern from } DMU_j \text{ by } I_k, k = 1,...,4\}.$$

For example, d_1, first column of D, is as

–
I_2
I_2
I_2
I_2
I_2
I_1, I_2, I_3, I_4
I_2
I_2
I_2
–
–
I_1, I_2, I_3, I_4
I_3, I_4
–
I_2
–

Table 1: Inputs

DMU	I1	I2	I3	I4
DMU1	12.23	0.25	1011.99	1011.99
DMU2	49.22	3.68	8546.15	8546.15
DMU3	72.64	3.94	11792.83	11792.83
DMU4	69.37	2.34	1041.10	1041.10
DMU5	35.03	2.68	5567.14	5567.14
DMU6	136.03	3.61	12326.69	12326.69
DMU7	392.51	7.84	19497.40	19497.40
DMU8	63.67	2.76	1684.76	1684.76
DMU9	20.71	2.77	1727.79	1727.79
DMU10	683.28	2.47	53676.87	53676.87
DMU11	117.67	2.19	6269.98	6269.98
DMU12	142.65	2.29	6980.21	6980.21
DMU13	76.11	2.67	4310.78	4310.78
DMU14	178.22	1.12	8625.49	8625.49
DMU15	50.90	1.53	2409.80	2409.80
DMU16	258.41	3.09	17029.92	17029.92
DMU17	125.42	0.24	14286.10	14286.10
DMU18	101.29	1.76	6515.61	6515.61
DMU19	32.31	4.52	2829.56	2829.56
DMU20	12.76	1.60	980.07	980.07

Table 2: Outputs.

DM U20	DM U19	DM U18	DM U17	DM U16	DM U15	DM U14	DM U13	DM U12	DM U11	DM U10	DM U9	DM U8	DM U7	DM U6	DM U5	DM U4	DM U3	DM U2	DM U1	DM U
69	42	42	231	6	10	327	27	62	27	39	92	146	229	24	58	45	212	89	38	O1
20	42	28	89	1	1	10	48	94	6	49	22	21	294	48	53	6	154	58	21	O2
13	32	101	125	258	51	178	76	143	118	683	21	64	393	136	35	69	73	49	12	O3
25	51	12	68	59	5	74	9	14	14	350	308	285	829	464	774	62	202	270	66	O4
980	2830	6516	14286	17030	2410	8625	4311	6980	6270	53677	1728	1685	19497	12327	5567	1041	11793	8546	1012	O5
300	385	514	231	9	4	147	310	230	57	328	382	574	1635	236	826	113	1704	199	632	O6
0	1772	281	2591	49	26	11	3824	172	5425	1212	1325	649	2346	525	0	173	130	419	39	O7
292	1258	535	9180	363	239	234	1552	821	783	1149	680	696	5882	6426	224	306	1661	7733	462	O8
80	117	66	62	24	70	124	59	206	132	128	65	37	1269	70	90	74	2070	74	120	O9
285	783	2802	2045	16472	1978	696	499	5409	3005	46775	634	516	3639	4805	4936	474	7881	435	264	O10
322	672	3113	2999	171	122	7572	2202	545	2351	5625	349	435	8707	1026	317	187	181	304	166	O11
945	1964	1484	3397	429	367	448	3140	2841	6195	6675	15258	11115	23662	1829	60147	892	8142	4915	3710	O12
262	675	330	133	0	113	30	2024	33	25	6932	27008	6807	12804	1313	14645	147	7438	5944	933	O13
27	4	60	16	0	0	0	190	0	3	0	0	0	0	6	0	2	7	2	66	O14
0	17039	13543	58657	0	0	455	53960	7885	2388	34532	25260	42084	1304733	2468	93640	0	10527	2483	2327	O15
3357	8463	6807	25128	22022	1511	1598	4861	9578	12003	9550	14010	30454	99998	46092	34766	12133	20495	29779	8685	O16

In order to find reducts and Cores, Discernibility function must be constructed. This function, which can be computed by discernibility matrix, has the following form:

$$f(A) = f_1(A) \times f_2(A) \times \ldots \times f_n(A)$$

where $f_i(A)$ is discernibility function related to column i. For example

$$f_1(A) = I_2 \times (I_1 + I_2 + I_3 + I_4) \times (I_3 + I_4)$$
$$= I_2 \times (I_3 + I_4)$$
$$= I_2 \times I_3 + I_2 \times I_4$$

$f_1(A)$ is obtained by definition of operations $+$ and $\times$, rough set operations [8]. Total discernibility function (of inputs) can be obtained similarly. Here, It is equals $f_1(A)$, i.e. $f(A) = I_2 \times I_3 + I_2 \times I_4$. Therefore $\{I_2, I_3\}$, $\{I_2, I_4\}$ are reducts of inputs and I_2 is the core. It means we can consider one of reducts instead of all inputs. Also, I_2, core, is the most important input.

Corresponding reducts of outputs can be attained as follows:

- reducts: $\{O_2, O_7, O_{11}, O_4, O_8\}$, $\{O_2, O_7, O_{11}, O_4, O_{16}\}$, $\{O_2, O_7, O_{11}, O_4, O_{13}\}$, $\{O_2, O_7, O_{11}, O_{13}, O_8\}$
- cores: O_2, O_7, O_{11}

Table 3: Classification of inputs and outputs component.

	I1	I2	I3	I4	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10	O11	O12	O13	O14	O15	O16
DMU1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1
DMU 2	1	2	1	1	2	1	1	2	1	1	1	4	1	1	1	1	1	1	1	2
DMU 3	1	2	1	1	3	3	1	1	1	4	1	1	4	1	1	1	2	1	1	1
DMU 4	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DMU 5	1	2	1	1	1	1	1	4	1	2	1	1	1	1	1	4	3	1	1	2
DMU 6	1	2	1	1	1	1	1	3	1	1	1	3	1	1	1	1	1	1	1	2
DMU 7	3	4	2	2	3	4	3	4	2	4	2	3	3	1	4	2	2	1	4	4
DMU 8	1	2	1	1	2	1	1	2	1	2	1	1	1	1	1	1	2	1	1	2
DMU 9	1	2	1	1	2	1	1	2	1	1	1	1	1	1	1	1	4	1	1	1
DMU 10	4	2	4	4	1	1	4	2	4	1	1	1	1	4	3	1	2	1	1	1
DMU 11	1	2	1	1	1	1	1	1	1	1	4	1	1	1	2	1	1	1	1	1
DMU 12	1	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DMU 13	1	2	1	1	1	1	1	1	1	1	3	1	1	1	1	1	1	4	1	1
DMU 14	1	1	1	1	4	1	1	1	1	1	1	1	1	1	4	1	1	1	1	1
DMU 15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DMU 16	2	2	2	2	1	1	2	1	2	1	1	1	1	2	1	1	1	1	1	1
DMU 17	1	1	2	2	3	2	1	1	2	1	2	4	1	1	2	1	1	1	1	1
DMU 18	1	1	1	1	1	1	1	1	1	2	1	1	1	1	2	1	1	2	1	1
DMU 19	1	3	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1
DMU 20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 4: Computational results of model (3) related to input and output reducts

	$\{O_2, O_7, O_{11}, O_4, O_8\}$ $\{I_2, I_3\}$	$\{O_2, O_7, O_{11}, O_4, O_{16}\}$ $\{I_2, I_3\}$	$\{O_2, O_7, O_{11}, O_4, O_{13}\}$ $\{I_2, I_3\}$	$\{O_2, O_7, O_{11}, O_{13}, O_8\}$ $\{I_2, I_3\}$
DMU1	1.0000	1.0000	1.0000	1.0000
DMU2	0.8450	0.7922	0.7748	1.0000
DMU3	0.6437	0.6437	0.6452	0.6452
DMU4	0.5187	0.6447	0.4124	0.4834
DMU5	1.0000	1.0000	1.0000	0.6621
DMU6	0.8098	0.5039	0.4614	0.6376
DMU7	0.7866	0.7866	0.7867	0.7867
DMU8	0.9976	1.0000	0.9743	0.9562
DMU9	1.0000	1.0000	1.0000	1.0000
DMU10	0.4939	0.4939	0.5205	0.3337
DMU11	1.0000	1.0000	1.0000	1.0000
DMU12	0.6655	0.6655	0.6655	0.6655
DMU13	1.0000	1.0000	1.0000	1.0000
DMU14	0.2291	0.2291	0.2291	0.0842
DMU15	0.1136	0.0583	0.0328	0.1136
DMU16	0.0666	0.1025	0.0665	0.0286
DMU17	1.0000	1.0000	1.0000	1.0000
DMU18	0.2460	0.2460	0.2460	0.2460
DMU19	1.0000	1.0000	1.0000	1.0000
DMU20	1.0000	1.0000	1.0000	1.0000

Results in Table 4 show that CCR model to evaluate efficiencies had been resolved with high amount of input and output components based on reducts. Also, it is noteworthy to say that there exist meaningless differences between the related efficiency of each reduct.

4-Conclusion:

In this paper, considering the properties of Rough set theory in the criteria classification and distinguishing the criteria minimal sets, along with providing an example the implementation of RST in input and output reduction is proposed. By this approach the necessary and unnecessary inputs and outputs can be discovered. This paper contains elementary arguments between RST and DEA. RST can be used to assess the performance of units that include qualitative inputs and outputs. Studying this topic is suggested for future research.

ACKNOWLEDGMENT

Authors would like to thank anonymous referees. This research was supported by a grant from the Shahre-Rey Branch, Islamic Azad University.

REFERENCES

Bowlin, W.F., 1998. Measuring Performance: An Introduction to Data Envelopment Analysis (DEA). Journal of Cost Analysis, 3(1): 3-28.

Charnes, A., W.W. Cooper and E. Rhodes, 1978. Measuring the Efficiency of Decision Making Units. European Journal of Operational Research, 2(6): 429-444.

Pawlak, Z., 1982. Rough Sets. International Journal of Computer Information Science, 11: 341-356.

Pawlak, Z., 1991. Rough Sets: Theoretical Aspects of Reasoning about Data. Dordrecht: Kluwer.

Polkowski, L. and A. Skowron, 1998. Rough Sets in Knowledge Discovery I & II. Heidelberg: Physica-Verlag.

Polkowski, L., 2002. Rough Sets: Mathematical Foundations. Heidelberg: Physica-Verlag.

Raab, R. and R. Lichty, 2002. Identifying Sub-areas that Comprise a Greater Metropolitan Area: The Criterion of County Relative Efficiency. Journal of Regional Science, 42: 579-594.

Walczak, B. and D.L. Massart, 1999. Rough sets theory. Chemometrics and Intelligent Laboratory Systems, 47: 1-16.