Available online at http://ijdea.srbiau.ac.ir

Int. J. Data Envelopment Analysis (ISSN 2345-458X)

Vol. 10, No. 3, Year 2022 Article ID IJDEA-00422, Pages 57-64
Research Article

International Journal of Data Envelopment Analysis

Science and Research Branch (IAU)

A combined machine learning algorithms and Interval DEA method for measuring predicting the efficiency

Hasan Babaei Keshteli, Mohsen Rostamy-Malkhalifih *

Department of Mathematics, Science and Research Branch, Islamic Azad University, Tehran, Iran

Received 25 January 2022, Accepted 29 April 2022

Abstract
One of the best methods for computing the efficiency of decision-Making Units (DMU) is Data Envelopment Analysis (DEA) that is useful for improving organizational performance. If we added a new unit to our observation sets, we have to run the model again. Nowadays, datasets from many organizations in the real world have been growing. So, we need a huge amount of computation for examining efficiency for new dataset. To overcome this problem, we combine Machine Learning (ML) and DEA. We consider organizations have interval data. According to we have interval data set, so we use interval DEA. Actually, we link between interval DEA and ML algorithms. First, we compute the efficiency score of these organizations by using Interval DEA. Second, we compute two scores that come in the first stage. Then, use these scores in ML. The empirical results show that the average accuracy of the predicted efficiency of DMUs is about 89%.

Keywords: Data Envelopment Analysis; Machine Learning; Efficiency; Interval Data

* Corresponding author: Email: mohsen_rostamy@yahoo.com

1. Introduction

Data Envelopment Analysis (DEA) is a non-parametric method to evaluate the performance of various decision-making units (DMUs) with multiple and similar inputs and outputs.

This method was introduced by Charnes et al. in 1978 (CCR model) [1]. Then, Banker et al. introduced the BCC model, which has the characteristic of variable returns to scale (VRS) [2]. (For further study, refer to [3,4]). There are many applications for DEA in different fields [5-13]. To evaluate performance using DEA models, each of the units under evaluation is considered a decision-making unit (DMU) and in order to calculate the efficiency score of each DMU, DEA models run once.

Now suppose that we added a new DMU to our under-evaluation DMUs, we must run the models again. Nowadays, datasets from many organizations in the real world have been growing. So, we need a huge amount of computation. For example, there are more than 73 million small and micro-sized companies in mainland China [14]. So, if we add a new DMU to this observation set and want to calculate the efficiency score, we have to run the models again. Therefore, we need a huge computer in terms of memory and CPU.

To overcome this problem, we try to attempt a linkage between DEA and machine learning (ML) algorithms.

Machine learning is a branch of artificial intelligence that allows the system to learn from previous data and while machine learning algorithms learn from past data they can produce more accurate predictive models [15]. This science is the study of algorithms and statistical models used by computer systems that use patterns and inferences to perform tasks instead of using clear instructions [16]. Machine learning allows computers to learn about a specific subject without the need for an explicit program. Also, ML helps a lot to save operational costs and improve the speed of data analysis. Therefore, as a subset of artificial intelligence, machine learning algorithms create a mathematical model based on sample data or training data to predict or make decisions without overt planning. The goal of machine learning techniques is to automatically recognize complex patterns in a data set, so inferences or predictions are possible on new data sets [15].

There are many successful applications of machine learning, including systems that analyze past sales data to predict customer behaviour, recognize faces or speech, and optimize robot behaviour so that a task can be completed with minimal resources [17] Also, ML has been used in various industries such as the oil and petrochemical industries and so on. So that the final optimal results are efficient and optimal. Therefore, due to the numerous applications of ML in various fields, researchers have conducted extensive studies on the use of ML in DEA.

Azadeh et al. (2011) presented a new method using a combined approach of DEA and NN to optimize the selection of a suitable location for a solar factory [18]. Also, the combination of clustering, Classification and Regression Trees (CART) with DEA was presented by Samoikenko and Osei-Bryson to increase the discriminatory power of the method [19]. In this regard, Wu et al used the combination of DEA, CART and data mining to evaluate the performance of the bank [20]. In another study, to evaluate health centres, the combination of DEA and CART was used by De Nicola et al [21]. Also, Random Forest Regression and DEA were used to assess Indian banks and also to investigate the impact of some government policies by Thaker et al [22]. In another study, DEA and clustering were used to investigate the impact of the coronavirus on the performance of different countries in responding to this crisis.

All classic DEA models assume that all inputs and outputs are precise and definitive. But this issue cannot always be true in real and practical issues in the world around us. In other words, in the real world, because decision-makers are faced with risk and uncertainty conditions, it is impossible to specify exact values for each input and output. Therefore, we will hesitate in the validation of the results obtained from classic models without considering imprecise data.

In general, uncertain data and imprecise data can be classified as interval numbers, fuzzy numbers, and random numbers. Therefore, how to evaluate and manage a set of decision-making units with interval, fuzzy or random data is one of the important research topics in the DEA field. All these cases indicate the need to develop DEA models in the field of theoretical foundations and real applications. So DEA models with imprecise data can be used to solve this shortcoming. Several studies have been done in this field, including [23].

This paper's main purpose is to link interval DEA and ML algorithms.

The rest of the paper is organized as follows. The interval-DEA method and ML algorithms are explained in Section 2. The research framework for combining the interval-DEA method with ML algorithms is discussed in Section 3. In Section 4, we ran the proposed method for a dataset, added some new data to our main dataset, and measured, predicted and compared with the efficiency scores obtained by the interval-DEA model. Finally, Section 5 concludes this paper and future studies.

2. Methodologies

2.1 DEA model

DEA is a non-parametric approach for measuring the efficiency of DMUs that use mathematical programming. The DEA-CCR model is the first model for computing efficiency that use the ratio of multi-outputs to multi-inputs of a DMU in terms of the similar ratios for all other DMUs to be less than or equal to one. The model does not need a priori weights on inputs and outputs [1].

Suppose that there is a set of $N=\{1,..,n\}$ DMUs, $M=\{1,..,m\}$ inputs, and $S=\{1,..,s\}$ outputs. Let $X_j=\{x_{1j},x_{2j},...,x_{mj}\}\in R^m$ and $Y_j=\{y_{1j},y_{2j},...,x_{sj}\}\in R^s$ be non-negative and non-zero vectors of inputs and outputs of DMU_j, $j\in N$. The DEA-CCR model for computing the efficiency of DMU_o is as follows.

$$
\begin{aligned}
\max \quad & \sum_{r=1}^{s} y_{ro}u_r \\
s.t \quad & \sum_{r=1}^{s} y_{rj}u_r - \sum_{r=1}^{m} x_{ij}v_i \le 0 \\
& \sum_{r=1}^{m} x_{io}v_i = 1 \qquad\qquad (1) \\
& v_i, u_r \ge \varepsilon, \qquad \forall i \in R
\end{aligned}
$$

In model (1), v_i and u_r are the weights of the i[th] input and r[th] output.

2.2 Interval DEA

Now, considering our data are an interval. We use models (2) and (3) for computing efficiency. Actually, we get an interval for efficiency.

$$
\begin{aligned}
\bar{\theta} = \max & \sum_{r=1}^{s} y_{ro}{}^{u}u_r \\
s.t \quad & \sum_{r=1}^{s} y_{rj}{}^{l}u_r - \sum_{r=1}^{m} x_{ij}{}^{u}v_i \le 0, \ \forall j \ne 0 \\
& \sum_{r=1}^{s} y_{ro}{}^{u}u_r - \sum_{r=1}^{m} x_{io}{}^{l}v_i \le 0 \\
& \sum_{r=1}^{m} x_{io}{}^{l}v_i = 1 \qquad\qquad (2) \\
& v_i, u_r \ge \varepsilon, \qquad \forall i \in R
\end{aligned}
$$

$$\underline{\theta} = \max \sum_{r=1}^{s} y_{ro}{}^{l} u_r$$

$$
\begin{aligned}
s.t \quad & \sum_{r=1}^{s} y_{rj}{}^{u} u_r - \sum_{r=1}^{m} x_{ij}{}^{l} v_i \leq 0, \ \forall j \neq 0 \\
& \sum_{r=1}^{s} y_{ro}{}^{l} u_r - \sum_{r=1}^{m} x_{io}{}^{u} v_i \leq 0 \\
& \sum_{r=1}^{m} x_{io}{}^{u} v_i = 1 \qquad (3) \\
& v_i, u_r \geq \varepsilon, \qquad \forall i \in R
\end{aligned}
$$

We consider $\frac{\bar{\theta} + \underline{\theta}}{2}$ as efficiency scores and show it by θ.

2.3 ML algorithm

One of the applications of ML algorithms is for prediction and decision. There are many algorithms in ML that have got different applications. In this study, we use a Genetic algorithm to measure and predict efficiency scores.

3. Research framework

We want to combine the DEA method with the ML algorithm. Actually, we will Combine Interval DEA with a Genetic algorithm.

First, we use the interval DEA model to measure the efficiency of each DMU in the training datasets. We consider θ as the target variable, and the input and output indicators of the DEA model are feature variables). Then we mark the units that they a have score efficiency equal one.

For learning the rules from these units, we use ML algorithms to parse these DMUs marked by DEA efficiency.

We added some units to our data sets.

Finally, their efficiency of them can be predicted through the trained ML model. A discussed ML-DEA algorithm, i.e., Genetic-DEA is used to predict the DEA efficiency of new DMU.

4. Case study and analysis

To illustrate the ML-DEA algorithm discussed, we run it for our data set. Suppose that the data of 25 companies (i.e. samples) are collected. Here, 20 samples are randomly chosen and used them as training datasets, and other 5 samples are used as the testing datasets. The result of 20 units are as follow:

Table 1: Result

DMU_j	$\bar{\theta}_{CCR}$	$\underline{\theta}_{CCR}$
1	1	1
2	1	0.371
3	1	0.52
4	1	1
5	0.76	0.61
6	1	0.917
7	1	0.72
8	1	1
9	1	1
10	1	1
11	1	1
12	0.49	0.32
13	0.70	0.45
14	0.72	0.26
15	1	0.41
16	1	0.22
17	1	1
18	0.95	0.26
19	1	0.99
20	1	0.18

Now, we compute the average $\bar{\theta}_{CCR}$ and $\underline{\theta}_{CCR}$ show it by θ. We consider θ as a goal in the ML algorithm.

We compute the efficiency of five unites in our data set by Interval CCR, calculate θ for them, and M–DEA Model and compare them to each other.

The results show that the average accuracy of the predicted efficiency of DMUs is about 89%.

5. Conclusions

The aim of this study is to link the interval DEA method and ML algorithm. We proposed a way that we can find efficiency scores when we added new DMU to our datasets and it did not need to run the model again. Also, we can use another algorithm of ML. For future study, we want to link between ML algorithm and fuzzy and stochastic DEA.

References

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

[2] Banker, R.D., Charnes, A., & Cooper, W.W. (1984). Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. Management Science, 30, 1078-1092.

[3] Cooper, W.W., Seiford, L.M., & Tone, K. (2006), Introduction to Data Envelopment Analysis and Its Uses: With DEA-Solver software and references, Springer, US.

[4] Ray, S.C. (2004), Data Envelopment Analysis: Theory and Techniques for Economics and Operations Research, Cambridge: Cambridge University Press.

[5] Emrouznejad, A., & Yang, G. (2018), A survey and analysis of the first 40 years of scholarly literature in DEA: 1978-2016. Journal of Socio-Economic Planning Sciences, 61 (1): 1-5.

[6] Jahanshahloo, G. R., Hosseinzadeh Lotfi, F., Rostamy-Malkhalifeh, M., & Ghobadi, S., "Using Enhanced Russell Model to Solve Inverse Data Envelopment Analysis Problems", The Scientific World Journal, vol. 2014, Article ID 571896, 10 pages, 2014. https://doi.org/10.1155/2014/571896

[7] Jahanshahloo, G. R., M.Soleimani-damaneh, M., & Rostamy-Malkhalifeh, M., An enhanced procedure for estimating returns-to-scale in DEA, Applied Mathematics and Computation, Volume 171, Issue 2, 15 December 2005, Pages 1226-1238.

[8] Sharafi, H., Rostamy-Malkhalifeh, M., Salehi, A., & Izadikhah, M. (2015). Efficiency Evaluation and Ranking DMUs in the Presence of Interval Data with Stochastic Bounds. International Journal of Data Envelopment Analysis, 3(1), 617-626.

[9] Sanei, M., Rostami-Malkhalifeh, M., & Saleh, H. (2009). A new method for solving fuzzy DEA models. International Journal of Industrial Mathematics, 1(4), 307-313.

[10] Peykani, P., Mohammadi, E., Sadjadi, S. J., & Rostamy-Malkhalifeh, M. (2018, May). A robust variant of radial measure for performance assessment of stock. In 3th International Conference on Intelligent Decision Science, Iran.

[11] Peykani, P., Seyed Esmaeili, F. S., Rostamy-Malkhalifeh, M., & Hosseinzadeh Lotfi, F. (2018). Measuring productivity changes of hospitals in Tehran: the fuzzy Malmquist productivity index. International Journal of Hospital Research, 7(3), 1-16.

[12] Banihashemi, S. A., Khalilzadeh, M., Shahraki, A., Malkhalifeh, M. R., & Ahmadizadeh, S. S. R. (2021). Optimization of environmental impacts of construction projects: a time–cost–quality trade-off approach. International Journal of Environmental Science and Technology, 18(3), 631-646.

[13] Razipour-GhalehJough, S., Hosseinzadeh Lotfi, F., Jahanshahloo, G., Rostamy-Malkhalifeh, M., & Sharafi, H. (2020). Finding closest target for bank branches in the presence of weight restrictions using data envelopment analysis. Annals of Operations Research, 288(2), 755-787.

[14] Zhang, M. (2017). The number of small and micro-sized enterprises in China has exceeded 73 million, China Central Television: http://news.cctv.com/2017/09/03/ARTIJoH0BPCYJ33FJ1EqCxm0170903.shtml.

[15] Alghanem, H., Shanna, M., Salloum, S., Shaalan, K. (2020). The Role of KM in Enhancing AI Algorithms and Systems. Advances in Science, Technology and Engineering Systems Journal, 4(5), 388-396. https://dx.doi.org/10.25046/aj0504245.

[16] Duda, R. O., Hart, P. E., Stork, D. G. (2001). Pattern Classification. New York: Wiley, 2(24), 305-307.https://doi.org/10.1007/s00357-007-0015-9.

[17] Athanassopoulos, A. D., Curram, S. P. (1996). A Comparison of Data Envelopment Analysis and Artificial Neural Networks as Tools for Assessing the Efficiency of Decision-Making Units. Journal of the Operational Research Society, 47 (8), 100-1016.

[18] Azadeh, A., Sheikhalishahi, M., Asadzadeh, S.M. (2011a). A flexible neural network-fuzzy data envelopment analysis approach for location optimization of solar plants with uncertainty and complexity. Renewable Energy, 36 (12), 3394-3401.

[19] Samoilenko, S.; Osei-Bryson, K.-M. Increasing the Discriminatory Power of DEA in the Presence of the Sample Heterogeneity with Cluster Analysis and Decision Trees. Expert Syst. Appl. 2008, 34, 1568–1581.

[20] Wu, D. Supplier Selection: A Hybrid Model Using DEA, Decision Tree and Neural Network. Expert Syst. Appl. 2009, 36, 9105–9112.

[21] De Nicola, A.; Gitto, S.; Mancuso, P. Uncover the Predictive Structure of Healthcare Efficiency Applying a Bootstrapped Data Envelopment Analysis. Expert Syst. Appl. 2012, 39, 10495–10499.

[22] Thaker, K.; Charles, V.; Pant, A.; Gherman, T. A DEA and Random Forest Regression Approach to Studying Bank Efficiency and Corporate Governance. J. Oper. Res. Soc. 2022, 73, 1258–1277.

[23] Aydin, N.; Yurdakul, G. Assessing Countries' Performances against COVID-19 via WSIDEA and Machine Learning Algorithms. Appl. Soft Comput. 2020, 97, 106792.