日本オペレーションズ・リサーチ学会
2022年 秋季研究発表会

An approximate method for large-scale DEA problem

05001516 Osaka University *Zhuang Qianwei
01604524 Osaka University Morita Hiroshi

1. Background

Data Envelopment Analysis (DEA) is a linear programming tool to evaluate the relative performance of a group of Decision Making Units(DMUs) with multiple inputs and outputs. If the number of DMUs turn to be large(eg. $n \geq 10000$), the corresponding time consumption to solve the model will increase remarkably[1].

In the previous research[2], we provided a framework which works very efficiently. In this study, a brand new approximate method to find benchmark DMUs is investigated by partitioning DMUs into groups. The required running-time for the large scale DEA problem will be expected further decrease at the cost of relatively small accuracy losing.

2. Theoretical basis

We consider the envelopment form of radial input-oriented variable return-to-scale(VRS) DEA model, which contains $n+1$ decision variables, $m+s+1$ constraints and n non-negativity restrictions, where n, m and s are the number of DMUs, inputs and outputs, respectively.

$$\min \varphi_l - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right)$$

$$s.t.$$

$$\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = \varphi_l x_{il}, i = 1,2,\dots,m,$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = y_{rl}, r = 1,2,\dots,s,$$

$$\sum_{j=1}^{n} \lambda_j = 1,$$

$$\lambda_j, s_i^-, s_r^+ \geq 0, \forall i, r, j$$

Definition:

DMU_l is efficient if $\varphi_l^* = 1$. DMU_l is strong efficient if $\varphi_l^* = 1$ and $\forall i,j,\ s_i^{-*} = s_r^{+*} = 0$. Otherwise if $\varphi_l < 1$ it is inefficient.

Lemma:

If DMU_l is the unique DMU that satisfies $\sum_{i=1}^{m} w_i x_{il} - \sum_{r=1}^{s} w_r y_{rl} = min_j\{\sum_{i=1}^{m} w_i x_{ij} - \sum_{r=1}^{s} w_r y_{rl}\}$, where $w_i \geq 0, w_r \geq 0\ and\ \sum_{i=1}^{m} w_i + \sum_{r=1}^{s} w_r > 0$ then DMU_l is VRS efficient.

The proof is not listed here since it is trivial.

3. Method

The above lemma provides a convenient way to find out efficient DMUs in a very simple arithmetic way instead of running the model. Thus, we designed the following framework to search for benchmark DMUs in a quick way.

3.1 Framework

Denote the input and output data of DMUs as matrix A. The following framework is proposed:

---

1. Get original matrix $A_{n:m+s}$,
2. $G_0 \leftarrow$ Select DMUs satisfies the preceding lemma as a hull,
3. Eliminate G_0 and calculate combination matrix $A_{n:p}^C$,
4. Group the rest DMUs into H groups: $G_1, \dots, G_H$ according to $A_{n:p}^C$,
5. For $k = 1, \dots, H$ do:
   $E_k \leftarrow$ Find exterior DMUs respect to G_0 in G_k,
   If $E_k = \{\}$ do:
   record k,
   go to 6
6. End.

In step3 p is a hyperparameter where $p \leq 2^{m+s} - 1$ in case of combination explosion happens. In step4 we do the select-eliminate operation iteratively to group DMUs base on each column's minimum value in $A_{n:p}^C$. H is dependent on the cardinality and dimensions.

3.2 Discussion about the framework

Every DMU in the preceding groups are strong efficient to those in the posterior groups according to the above framework. In DEA, the production frontier is formed by efficient DMUs. And intuitively strong efficient ones are more possibly be selected as benchmarks for performance evaluation. With several conventional DEA computation result, we define $b_k = \sum_{l=1}^{n} \sum_{j \in G_k} {\lambda_{lj}}^*$ and the distribution of benchmarks respect to each group is as follows:

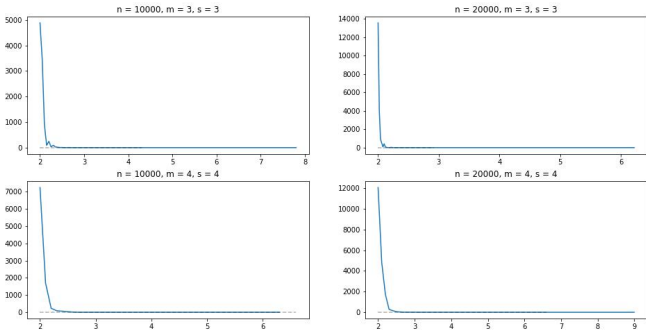

Fig.1. Benchmarked index of each group

Fig.1 explicitly shows that target DMUs majorly locate in the first several groups. Thus in step5 we can find groups contain most benchmarks.

4. Datasets and results

Each input and output of the dataset denoted by $m + s(n)$ here are generated by uniform distribution independently. Here we show part of the results.

Table1. Computational results

Dataset	Checked Groups	Accuracy
1+1(20000)	2/5034	0.9999
1+2(20000)	8/2033	0.9997
3+3(15000)	25/163	0.9998
4+3(10000)	14/67	0.9988
4+4(20000)	18/68	0.9998
4+4(30000)	27/93	0.9997
5+5(20000)	14/50	0.9940
5+5(30000)	17/60	0.9654
6+6(10000)	20/33	0.9997
7+7(10000)	20/29	0.9991

Step5 defines a terminate criteria and the final checked group's index is recorded in the second column of Table1 with the total number of groups from step4. The result shows that only of the groups got checked, which is supposed to be a big contrast to our previous framework[2] which has to loop over the whole dataset for searching benchmark DMUs.

The accuracy in the third column is the percent of DMUs' efficiency score with the difference less than 0.1 in comparison with conventional computation result. Each dataset shows relatively high accuracy proportion which means most of the DMUs are evaluated properly. Since the conventional computation is conducted on a different computer, the small differences can be neglected. However, there are some DMUs received higher efficiency scores than usual since this approximate framework can possibly leave several benchmarks in the posterior groups in step5(eg. dataset 5+5(20000)).

5. Conclusions and future work

We proposed an approximate framework for large scale DEA problem. Only part of DMUs are supposed to be examined to find out benchmarks, which is expected to result in remarkable time reduction especially for datasets with larger cardinality. This current method can be a good option if we only care about the efficiency score.

However, the targeted DMUs are also important issue for performance evaluation. In the future, we will try to improve the computation accuracy of this framework and investigate more about the variation of target DMUs. On the other hand, the number of groups from step 4 has dependence on the cardinality and dimensions of the dataset, which also deserves further study.

References

[1]K., D., Z., J., et al. (2019). Data envelopment analysis and big data. European Journal of Operational Research, 274(3), 1047-1054.

[2]Z., Q., M., H., Streamlined DEA computation in the big data context. 日本オペレーションズ・リサーチ学会 2022 年春季研究発表会.