DEA implementation and clustering analysis using the K-Means algorithm

C. A. A. Lemos, M. P. E. Lins & N. F. F. Ebecken
COPPE/Universidade Federal do Rio de Janeiro, Brazil

Abstract

Nowadays, problems that involve efficiency analysis and decision support systems inside a company need special attention and a number of tools have been developed to support managers. DEA – Data Envelopment Analysis is one of these tools and its use is increasing in research and in new developments. The problem is how to improve the quality of DEA analysis when the DMU (decision-making unit) it analyzes is considered efficient, and how to guarantee the analysis if the input and output parameters that contain a lot of zeros? Probably these parameters have not been considered in how to visualize the inputs and outputs in n-dimensional space? This paper proposes combining another tool with DEA based in data mining, CLUSTERING, to evaluate the efficiency analyses made for DEA tools, and visualize groups which have inefficient DMUs, based on the K-Means algorithm, and apply over a telecommunication database that contains an indicator of efficiency of the telephone installation in the Brazilian market.
Keywords: Data Envelopment Analysis, clustering, data mining, telecommunication quality indicator, decision support system.

1 Introduction

Problems that involve efficiency analysis inside a company need to have special attention. Tools are being development to support managers. Some companies use complex formulations based on traditional statistical methods and others are using new environments based on computational intelligence and others tools. DEA [2] is one of these tools that obtain relative efficiency between two or more companies, departments or groups. The problem in DEA is how to improve the quality of analysis when the DMU (decision-making unit) it analyzes is



considered efficient. In this paper we will present and discuss one possibility to improve DEA analysis making a pre-processing in data using intelligent computational toll based on clustering.

2 DEA: Data Envelopment Analysis

DEA uses a linear programming approach to identify the efficient DMUs (decision making units), those units that make the most efficient use of inputs to produce outputs. The efficiency units consist of a frontier among all DMUs. The efficiencies of the DMUs are measured by projecting to this frontier. The DEA model in its original form represents the performance of efficiency of the DMU as the ratio of weighted outputs to weight inputs [3]. To date, the DEA literature has developed numerous models and detailed discussion can be found in [2,3]. Essentially, various models for DEA seek to establish which subset of DMUs determines an envelopment surface and address how to characterize each DMU by an efficiency score. There are two basic models: CRS – constant returns to scale and VRS – variable returns to scale. Both are presented below [2].

2.1 The constant returns to scale (CRS) DEA Model

This method was proposed by Charnes, Cooper and Rhodes (CCR models - 1978) where the term DEA – data envelopment analysis, was first used. This first approach uses input orientation and assumes constant return to scale. Later, others papers have considered alternatives sets of assumptions. Suppose N data points (DMUs) are to be evaluated. Assume there are data on K inputs and M outputs for each DMU. For the i-th DMU they are represented by column vector x_i and y_i, respectively. The K x N input matrix, X and M x N output matrix, Y, represent the data for all DMUs. An intuitive way to introduce DEA is via the ratio form. For each DMU, we would like to obtain a measure of the ratio of all outputs over all inputs, such as $u'y_i/v'x_i$, where u is a M x 1 vector of output weights and v is a K x 1 vector of input weights.

The optimal weights are obtained by solving the mathematical programming problem:

$$\begin{aligned} &\max_{u,v}(u'y_i / v'x_i), \\ &st \quad u'y_j / v'x_j, \qquad j = 1,2,....,N \\ &\qquad u, v \geq 0. \end{aligned} \tag{1}$$

This involves finding values for u and v, such that the efficiency measure for the i-th firm is maximised, subject to the constraints that all efficiency measure must be less than or equal to one.

2.1.1 The problems of slacks

The piece-wise linear form of nom-parametric frontier in DEA can cause few difficulties in efficiency measurement. The problem arises because the sections

of the piece-wise linear frontier that run parallel to the axes. This problem can give us incorrect analysis (inefficient Pareto frontier). The CRS model assumption is only appropriate when all firms are operating at an optimal scale. Imperfect competition, constrains on finance, etc., may cause a DMU to be not operating at optimal scale.

2.2 The variable returns to scale (VRS) DEA model

Banker, Charnes and Cooper 1984 (BCC model - 1984), suggest an extension of the CRS DEA model to account for variable returns to scale situation. The CRS linear programming problem (eq. 2) can be easily modified to account for VRS by adding the convexity constraint N1′λ=1:

$$\begin{aligned} &\min_{\Theta,\lambda} \Theta, \\ &\qquad -yi + Y\lambda \geq 0, \\ &st \qquad \Theta xi - X\lambda \geq 0, \\ &\qquad N1'\lambda = 1, \\ &\qquad \lambda \geq 0. \end{aligned} \tag{2}$$

where: N1 is an N x 1 vector of ones.

This approach forms a convex hull of intersecting planes which envelope the data points more tightly than the CRS conical hull and thus provides technical efficiency scores which are greater than or equal to those obtained using the CRS model. The VRS specification has been the most commonly used specification in the 1990s [2].

3 Clustering: K-means algorithm

Clustering is a toll to data mining used to classify things that have similar characteristics, and the output takes the form of a diagram that shows how the instances are inside into cluster. In the simplest case this involves associating a cluster number with each instance, which might be depicted by laying the instances out in two dimensions and partitioning the space to show each cluster.

Some clustering algorithms allow one instance to belong to more than one cluster, so the diagram might lay the instances out in two dimensions and draw overlapping subnets representing each cluster. Others, associate instances with clusters probabilistically rather than categorically. In this case, for every instance there is a probability or a degree of membership with which it belongs to each cluster (fuzzy clustering). Some algorithms produce a hierarchical structure of cluster [6].

There are a lot of applications of the K-mean Clustering, from unsupervised learning of Neural Network, Pattern Recognitions, Classification Analysis,

Artificial Intelligent, Image Processing, etc… In principle, if you have several objects and each object has attributes and you want to classify the objects based on the attributes, then you can apply this algorithm.

3.1 K-means algorithm

3.1.1 How K-means clustering works

If the number of data is less than the number of clusters then we assign each data as the centroid of the cluster. Each centroid will have a cluster number. If the number of data is bigger than the number of cluster, for each data, we calculate the distance to all centroid and get the minimum distance. This data is said to belong to the cluster to another that has minimum distance from this data.

Since we are not sure about the location of the centroid, we need to adjust the centroid location based on the current update data. Then we assign all the data to this new centroid. This process is repeated until no data is moving to another cluster anymore. Mathematically, this loop can be proved to be convergent. The ref. [8] has an example to k-mean algorithm in Visual Basic code.

3.1.2 Weakness of K-mean clustering

Similar to other algorithms, K-mean clustering has many weaknesses:

- When the number of data are not so many, initial grouping will determine the cluster significantly;
- The number of cluster, K, must be determined before hand;
- We never know the real cluster, using the same data. If it is input in a different way it may produce a different cluster if the number of data is few;
- We never know which attribute contributes more to grouping process since we assume that each attribute has the same weight.

4 The databases

DEA needs a data base where found inputs and outputs about specific DMU. In our research about the telecom manager indicator, we created a specific database to test and compare the methodologies proposal in this paper. Table 1 shows the database implemented by date from ref. [9] and [10]:

DMUs:
- Number of Decision Making Units: 34
- DMUs are telecommunications operation company in Brazil, acting in fixed telephony service.

INPUTs:
- POPulation – Number of inhabitants per region or state [POP]
- Cities NUmber –Inside the state or region [CNU]
- Total Area: state or Region - (Km^2) [TAR]
- Index of Urban Concentration [IUC]

OUTPUT:
- Number of Fix Telephone per state or region [NFT]

Table 1: Database: DEA efficiency.

DMU	Ref	Region	State	INPUT				OUTPUT
				POP	CNU	TAR	ICU	NFT
1	RJ	Region I	RJ	14.879.118	92	43.696,05	0,9289	4.943.943
2	MG	Region I	MG	18.553.312	759	522.010,18	0,7901	3.674.335
3	MG	Region I	MG	2.742.705	94	64.518,11	0,7901	571.920
4	ES	Region I	ES	3.250.219	78	46.077,52	0,7576	816.836
5	BA	Region I	BA	13.435.612	417	564.692,67	0,6527	2.139.280
6	SE	Region I	SE	1.874.613	75	21.910,35	0,6788	280.378
7	AL	Region I	AL	2.917.664	102	27.767,66	0,6577	321.076
8	PE	Region I	PE	8.161.862	185	98.311,62	0,7419	1.227.888
9	PB	Region I	PB	3.518.595	223	56.439,84	0,6947	433.921
10	RN	Region I	RN	2.888.058	167	52.796,79	0,7039	405.167
11	CE	Region I	CE	7.758.441	184	148.825,60	0,6850	990.652
12	PI	Region I	PI	2.923.725	222	251.529,19	0,6117	309.378
13	MA	Region I	MA	5.873.655	217	331.983,29	0,5726	514.887
14	PA	Region I	PA	6.574.993	143	1.247.689,52	0,6269	731.814
15	AP	Region I	AP	534.835	16	142.814,59	0,7985	87.255
16	AM	Region I	AM	3.031.068	62	1.570.745,68	0,6965	440.078
17	RR	Region I	RR	357.302	15	224.298,98	0,6936	65.734
18	SC	Region II	SC	5.607.233	293	95.346,18	0,7522	1.589.672
19	PR	Region II	PR	9.906.866	374	187.356,01	0,7856	2.577.954
20	PR	Region II	PR	619.179	25	11.958,84	0,7856	160.787
21	MS	Region II	MS	2.169.688	75	347.601,65	0,8051	527.230
22	MS	Region II	MS	37.408	2	9.523,31	0,8051	9.140
23	MT	Region II	MT	2.651.335	139	903.357,91	0,7499	521.986
24	GO	Region II	TO	6.536.640	239	333.284,97	0,7021	1.408.454
25	GO	Region II	GO	5.306.459	7	6.801,73	0,8279	36.864
26	DF	Region II	DF	2.189.789	1	5.801,94	0,8957	888.183
27	RO	Region II	RO	1.455.907	52	237.576,17	0,6077	254.920
28	AC	Region II	AC	600.595	22	152.581,39	0,6181	98.560
29	RS	Region II	RS	10.510.992	472	267.422,34	0,7914	2.695.301
30	RS	Region II	RS	500.523	24	14.326,20	0,7914	126.692
31	SP	Region III	SP	38.709.320	570	219.347,87	0,8935	11.908.830
32	SP	Region III	SP	774.186	12	4.617,85	0,8935	237.315
33	SP	Region III	SP	770.000	11	4.233,03	0,8935	236.477
34	SP	Region III	SP	3.519.029	52	20.010,68	0,8935	1.083.590

5 Experiments and results

If you look at the numbers in Table 1, it is possible to see a great variation between the lowest and the biggest values. Therefore, the fist thing is to normalize the database. After this we put the data in EMS software [4] and calculate the efficiency score using the basics DEA models. After that we convert the data base to ARFF format file and clustering using WEKA software [5]. The experiment follows the flowchart indicated in Figure 1.

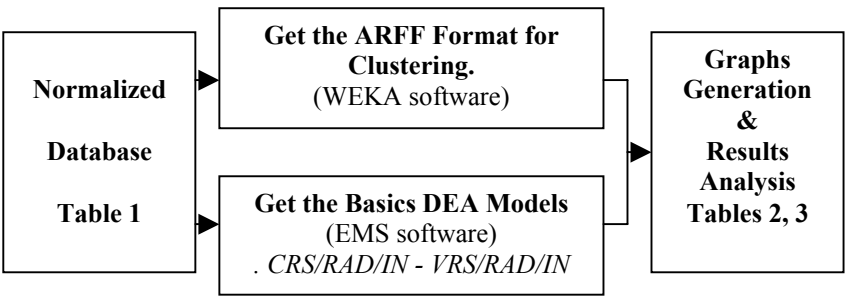


Figure 1: DEA x clustering.

5.1 Clustering database

Figure 2 and Figure 3 show the results of cluster analysis using WEKA software.

In Table 2 we can see the DMUs and the clusters they belong, since the output of software is colored.

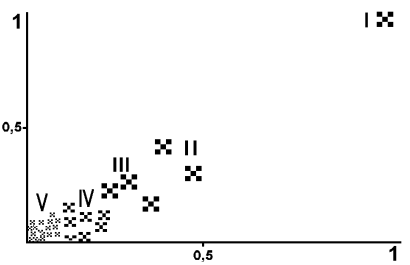


Figure 2: Clusters: population (x axis) x Number Fix Phone (y axis).

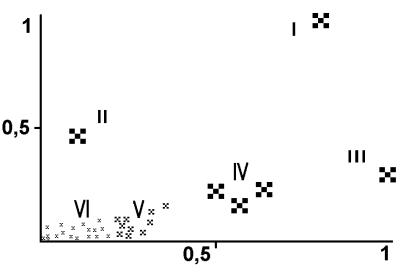


Figure 3: Clusters: Number Cities (x axis) x Number Fix Phone (y axis).

Table 2: DEA efficiency and clustering.

DMU	Ref	Region	State	DEA CRS Efficiency	DEA VRS Efficiency	Clusters	
						Figure 2	Figure 3
1	RJ	Region I	RJ	**100,00%**	**100,00%**	II	II
2	MG	Region I	MG	61,90%	87,30%	II	III
3	MG	Region I	MG	54,90%	87,60%	V	VI
4	ES	Region I	ES	68,20%	95,00%	V	VI
5	BA	Region I	BA	49,20%	94,70%	III	IV
6	SE	Region I	SE	37,80%	**100,00%**	V	VI
7	AL	Region I	AL	30,00%	**100,00%**	V	VI
8	PE	Region I	PE	44,50%	89,30%	IV	V
9	PB	Region I	PB	34,20%	94,00%	V	V
10	RN	Region I	RN	37,90%	93,60%	V	V
11	CE	Region I	CE	37,80%	93,70%	IV	V
12	PI	Region I	PI	29,10%	97,80%	V	V
13	MA	Region I	MA	25,80%	**100,00%**	IV	V
14	PA	Region I	PA	32,80%	96,00%	IV	V
15	AP	Region I	AP	40,20%	87,10%	V	VI
16	AM	Region I	AM	39,50%	88,40%	V	VI
17	RR	Region I	RR	45,40%	**100,00%**	V	VI
18	SC	Region II	SC	81,60%	94,90%	IV	V
19	PR	Region II	PR	77,50%	86,60%	III	IV
20	PR	Region II	PR	64,00%	**100,00%**	III	VI
21	MS	Region II	MS	61,10%	82,80%	V	VI
22	MS	Region II	MS	60,20%	**100,00%**	V	VI
23	MT	Region II	MT	52,00%	82,60%	V	V
24	GO	Region II	TO	63,00%	89,40%	IV	V
25	GO	Region II	GO	3,80%	**100,00%**	IV	VI
26	DF	Region II	DF	**100,00%**	**100,00%**	V	VI
27	RO	Region II	RO	43,20%	**100,00%**	V	VI
28	AC	Region II	AC	40,50%	**100,00%**	V	VI
29	RS	Region II	RS	76,50%	85,60%	V	IV
30	RS	Region II	RS	62,40%	99,30%	V	VI
31	SP	Region III	SP	**100,00%**	**100,00%**	I	I
32	SP	Region III	SP	75,60%	99,40%	V	VI
33	SP	Region III	SP	75,70%	**100,00%**	V	VI
34	SP	Region III	SP	82,60%	92,60%	V	VI

5.2 Analysis

In Table 2 we can see the result of EMS software (DEA Efficiency) and the result of WEKA software (CLUSTERING). We put in bold letters the efficiency 100%, in both DEA basic models: CRS and VRS.

In Table 3 the DMUs are classified inside of the respective cluster that they had been found of the proper data.

Table 3: DMUs clustering.

	Cluster I	**Cluster II**	**Cluster III**	**Cluster IV**	**Cluster V**	**Cluster VI**
Graf 5.1.1	**DMU-31(*)**	**DMU-1(*)** DMU-2	DMU-5 DMU-19 **DMU-20**	DMU-8 DMU-11 **DMU-13** DMU-14 DMU-18 DMU-24 **DMU-25**	DMU-3 DMU-4 **DMU-6** **DMU-7** DMU-9 DMU-10 DMU-12 DMU-15 DMU-16 **DMU-17** DMU-21 **DMU-22** DMU-23 **DMU-26(*)** **DMU-27** **DMU-28** DMU-29 DMU-30 DMU-32 **DMU-33** DMU-34	-
Graf 5.1.2	**DMU-31(*)**	**DMU-1(*)**	DMU-2	DMU-5 DMU-19 DMU-29	DMU-8 DMU-9 DMU-10 DMU-11 DMU-12 **DMU-13** DMU-14 DMU-18 DMU-23 DMU-24	DMU-3 DMU-4 **DMU-6** **DMU-7** DMU-15 DMU-16 **DMU-17** **DMU-20** DMU-21 **DMU-22** **DMU-25** **DMU-26(*)** **DMU-27** **DMU-28** DMU-30 DMU-32 **DMU-33** DMU-34

(*) DMU´S who get 100% efficiency in basic DEA models CRS and VRS.

6 Conclusion

In DEA analysis with more than one input and one or two outputs, we have difficulty to visualize the behavior of data sets. The analysis of data set improves when a cluster algorithms is added. With the information obtained by clustering, we can return to DEA software and perform the analysis in a more homogeneous group. This prevents the problem of slacks mentioned before.

Using clustering software we can see the problem for different parameters and plot graphs to assist the analysis. Looking for DMU 31 we can identify "out-standard" DMU, probably a benchmark DMU. This DMU needs a specific analysis, and included in other cluster will be problem. We can do the same analysis for all groups and graphs and improve the DEA analysis.

Clustering analysis combined with DEA analysis is a very interesting tool, reducing the numbers of variables that decides if DMU is efficiency or not, improve the visualization of variables and making a coherent and a homogeneous comparison.

References

[1] Banker, R.D., A. Chanes, W, W Cooper, 1984. Some Models for estimating Technical and Scale Inefficiencies In Data Envelopment Analysis, Management Science.

[2] Coelli, T., Prasada Rao, George Battese. 1998 - An Introduction To Efficiency and productivity Analysis, Kluwer Academic Publishers, Boston.

[3] Cooper, W., Laurence Seiford, Kaoru Tone. 2002. Data Envelopmente Analysis: A comprehensive text with models, applications, references and DEA-solver software. Dordrecht, Netherlands: Kluwer Academic publishers.

[4] Scheel, Holger, A Guide for EMS Version 1.3: A Data Envelopment Analysis (Computer Program). University Dortmund – Germany. 2000

[5] Written, I. H, A Guide for WEKA – Wikato Environment for knowledge Analysis (Computer Program)– University of Waikato, New Zealand 1999-2004.

[6] Written, I. H. Data mining: practical machine learning tools and techniques with java implementations / Ian H. Witten, Eibe Frank.

[7] Dulá, J. H., Computation in DEA – School of Business Administrations – University of Mississippi – 2001.

[8] Teknomo, Kardi, K-Mean Clustering. http://www.planetsourcecode.com.

[9] ANATEL Brazilian Bureau of Telecommunication – www.anatel.gov.br.

[10] IBGE – Brazilian Institute of Geography and Statistics- www.ibge.gov.br.

