MINING SCIENTIFIC RESULTS THROUGH THE COMBINED USE OF CLUSTERING AND LINEAR PROGRAMMING TECHNIQUES*

Andrea Tagarelli, Irina Trubitsyna, Sergio Greco
DEIS - University of Calabria
87030 Rende, Italy

Keywords: Data Mining, Clustering, DEA, Efficiency Measures.

Abstract: The paper proposes a technique based on a combined approach of data mining algorithms and linear programming methods for classifying organizational units, such as research centers. We exploit clustering algorithms for grouping information concerning the scientific activity of research centers. We also show that the replacement of an expensive efficiency measurement, based on the solution of linear programs, with a simple formula allows clusters of very good quality to be computed efficiently. Some initial experimental results, obtained from an analysis of research centers in the agro-food sector, show the effectiveness of our approach, both from an efficiency and a quality-of-results point of view.

1 INTRODUCTION

The high performance of organizational units, also known as *decision-making units*, relies on good decision support which can have a major impact on the achievement of the goals of the unit. On the other hand, the soundness of a decision usually reflects the quality of the activities of the unit. For instance, a decision made on a project in which a scientific research center is involved could lead to an increasing in the productivity of the research center itself, provided that such a project represents a relevant activity from a scientific point of view.

The process of evaluating and comparing the performances of organizational units is a challenging application, in principle, for several research disciplines. In particular, there is growing interest in measuring the efficiency of organizational units involved in similar activities, technologies and inputs. Moreover, evaluating the productivity of research centers is useful from the point of view of a careful deployment of financial resources to the centers themselves: intuitively, a research center with a high performance may gain more economic benefits rather than other research centers with lower quality scores.

*Work supported by a MURST grant under the project "Sistemi informatici integrati a supporto del bench-marking di progetti ed interventi ad innovazione tecnologica in campo agro-alimentare"

Traditional efficiency measures are often inadequate due to the presence of multiple inputs and outputs related to different resources, activities and environmental factors. In many productive fields, the methods of parametric and non-parametric evaluation seem to be preferred with respect to the combined use of traditional indicators. In fact, such methods provide a synthetic indicator of the productivity by simultaneously considering multiple inputs and outputs of the productive process. As a consequence, they allow the comparison of the efficiency of a given organizational unit with respect to the frontier of the possible efficient solutions for all the organizational units. The parametric methods (DFA, SFA) require the presumptive definition of the productive function, while the non-parametric ones (DEA, FDH) are able to determine the relative efficiency of organization units by means of linear programming techniques. This is an advantage, since the non-parametric methods permit us to evaluate the performance of organization units without any knowledge of their productive process.

The contribution of this paper is the definition of a methodology for the classification of research centers combining data mining techniques, such as clustering, and linear programming techniques. The expected result is a system capable of organizing research centers by considering information about the volume and the quality of their scientific activity. We study how to extract and represent both scientific results and perfor-

Tagarelli A., Trubitsyna I. and Greco S. (2004).
MINING SCIENTIFIC RESULTS THROUGH THE COMBINED USE OF CLUSTERING AND LINEAR PROGRAMMING TECHNIQUES.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 84-91
DOI: 10.5220/0002624000840091

mance information from research centers. Then, we exploit clustering algorithms to accomplish the task of organizing such information, and evaluate the corresponding accuracy of the proposed approach.

The remainder of this paper is organized as follows. The next section is a short overview of the clustering process in a suitable way to our purposes. Section 3 presents DEA, a linear programming based technique for measuring the efficiency of organizational units. Section 4 illustrates the overall architecture and the features of a system for the classification of research centers. Section 5 describes a methodology for organizing research centers based on models computing their efficiency. In Section 6 proposes an alternative way to compute the efficiency of research centers; this section ends reporting the experimental evaluation stating the effectiveness of our approach. Finally, Section 7 contains concluding remarks.

2 DATA CLUSTERING

Clustering is the task of organizing a collection of objects (whose classification is unknown) into meaningful or useful groups, called *clusters*, based on the interesting relationships discovered in the data. The goal is that the objects within a cluster will be highly similar to each other, but will be very dissimilar from objects in other clusters. The greater the homogeneity/heterogeneity within/between groups, the better the resulting partition of clusters.

A first stage in a typical clustering task is the definition of a model to represent the objects, drawn from the same feature space. Typically, an object is represented as a multidimensional vector, where each dimension is a single feature. Formally, given an m-dimensional space, an object $\mathbf{x}$ is a single data point and consists of a vector of m measurements: $\mathbf{x} = (x_1, \ldots, x_m)$. A set of n objects $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ to be clustered is in the form of an object-by-attribute structure, i.e. an n-by-m matrix. The scalar components x_i of $\mathbf{x}$ are called features or attributes.

Many different clustering algorithms can be exploited (Jain and Dubes, 1988). Partitional and hierarchical clustering techniques are by far the most popular and important ones. In this work, we exploit the well-known *k-Means* partitional algorithm which has the main advantage of requiring $\mathcal{O}(n)$ comparisons and guarantees a good quality of clusters. The algorithm starts by randomly choosing k objects as the initial cluster centers. Then it, iteratively, reassigns each object to the cluster to which it is the closest, based on the proximity between the object and the cluster center until a convergence criterion is met.

The definition of a proximity measure between objects is crucial in the clustering. Object proximity is assessed on the basis of the attribute values describing the objects, and is usually measured by a distance function or metric. The most commonly used metric, at least for ratio scales and continuous features, is the *Minkowski* metric, defined as $d_M(\mathbf{x}_i, \mathbf{x}_j) = (\sum_{h=1}^{m} |x_{ih} - x_{jh}|^p)^{1/p} = \|(\mathbf{x}_i - \mathbf{x}_j)\|_p$, which is a generalization of the popular *Euclidean* distance, obtained when $p = 2$. Higher p values increase the influence of large differences at the expense of small differences and, from this point of view, the Euclidean distance represents a good trade-off. It works well when the objects within a collection are naturally clustered in compact and convex-shaped groups, and it is exploited to define the squared-error criterion, which is the most intuitive and frequently used criterion function in partitional clustering algorithms. The squared-error criterion computes the sum of the squared distance of each object from the center of the cluster, and tries to make the resulting clusters as compact and as separate as possible.

Quality in clustering deals with questions like how well a clustering scheme fits a given dataset, and how many groups partition the analyzed data. Three approaches are adopted to investigate cluster validity (Halkidi et al., 2002): external criteria, internal criteria, and relative criteria. A pre-specified structure, which reflects our intuition about the clustering structure of the dataset, is exploited by external criteria to evaluate a clustering. Internal criteria are defined over quantities that involve the representations of the data themselves (e.g. proximity matrix). The basic idea of the latter approach is instead the comparison of different clustering schemes resulting from the same algorithm but with different parameter values.

Our choice falls back on external criteria, since it is particularly convenient, for our purposes, to measure the degree to which a dataset confirms an a-priori specified scheme.

3 DEA TECHNIQUE

Data Envelopment Analysis (DEA) is a linear programming technique that has been frequently applied to assess the efficiency of *decision-making units* (hereinafter called *DMUs*), where the presence of multiple inputs, as well as outputs, makes comparisons difficult.

The measurement of relative efficiency was addressed in (Farrell, 1957) and developed in (Farrell and Fieldhouse, 1962), focusing on the creation of a hypothetical efficient unit, as a weighted average of efficient units, to act as a comparator for an inefficient unit. The first DEA model was introduced in (Charnes et al., 1978) and its extents were used for

measuring and comparing the efficiency of local authority departments, schools, hospitals, shops, bank branches and similar entities with homogeneous sets of units (Chung et al., 2000; Zhu, 2002; Charnes et al., 1994; Stern et al., 1994; Thanassoulis et al., 1987). In the Data Mining context, (Sohn and Choi, 2001) proposes using DEA in order to find the weights involved in multi-attribute performances of classifiers in a data ensemble algorithm. A recent bibliography of DEA including applications can be found in (Emrouznejad, 2001).

DEA is a non-parametric technique, in the sense that it does not require any assumption about the functional form relating the independent variables to the dependent variables. By contrast, the efficiency of each DMU is computed as the ratio of a weighted sum of outputs and a weighted sum of inputs, where the weight sets are different for distinct DMUs and have to be selected to maximize the efficiency of each DMU.

The selection of the attributes and their partition, as input and output parameters, play a crucial role in the definition of a DEA model. In other terms, a DEA model involves not only the choice of individual attributes, but also deciding whether an attribute will be treated as an input or an output parameter.

A DEA model can hence be formally stated as follows. Given N DMUs with I inputs and O outputs, let x_{ij} and y_{oj} be, respectively, the i-th input and the o-th output of DMU j, and let v_{ij} and w_{oj} be the corresponding weights, where $j \in \{1,\ldots,N\}$, $i \in \{1,\ldots,I\}$, $o \in \{1,\ldots,O\}$. The efficiency E_j of a given DMU j can be obtained by solving the following linear program:

$$
\begin{array}{ll}
max & E_j = \frac{\sum_{o=1}^{O} w_{oj} y_{oj}}{\sum_{i=1}^{I} v_{ij} x_{ij}} \\
subject\ to & \frac{\sum_{o=1}^{O} w_{oj} y_{ol}}{\sum_{i=1}^{I} v_{ij} x_{il}} \leq 1 \\
 & w_{oj}, v_{ij} \geq \varepsilon \\
where & l \in \{1 \cdots N\}, i \in \{1 \cdots I\}, o \in \{1 \cdots O\}.
\end{array}
$$

The variables of the above problem are the weights that have been chosen to maximize the efficiency of a given DMU j. The first constraint represents the upper bound for the efficiency of all DMUs computed with the current weights. The second constraint, where ε is a positive value close to 0, avoids that an input or an output is totally ignored in determining the efficiency.

If $E_j = 1$ then DMU j is efficient with respect to other DMUs, otherwise there is some other more efficient DMU, even if the weights have been chosen in favor of DMU j. In fact, the solution technique attempts to make the efficiency E_j as large as possible. The search procedure stops when some DMU hits the upper bound of 1. Thus, for an inefficient DMU at least another unit will be efficient with the given set of weights.

The flexibility in the choice of weights is both a weakness and a strength of this approach. It is a weakness because in some cases the evaluation can be more affected by the choice of the weights than by the attribute values of DMUs; on the other hand, the independence of the weights is a strength because the evaluation of DMUs' inefficiency is definitive as the most valuable weights have been chosen.

4 A SYSTEM FOR CLASSIFYING DMUs

We present a system for the classification of research centers based on different parameters involving scientific results and efficiency indicators. For this purpose, the system combines clustering algorithms and linear programming techniques. It takes in input aggregate information, stored in the source database, concerning the scientific activity of research centers and, in particular, aggregate data involving any product concerning scientific activities, such as publications, projects, citations, and patents. As the number of scientific publications and citations are absolute values, not actually useful without a comprehensive point of reference, some *scientometric indicators* (see Section 4.1) need to be taken into account.

The global classification process is reported in Figure 1 and consists of three main steps implemented by the following modules:

1. *Indicator computation* – This module takes in input the source aggregate information about research centers and computes some scientometric indicators on the volume and quality of the scientific activity of research centers. The output of this module is merged with the source database.
2. *Efficiency evaluation* – The efficiency evaluation is based on a given model which exploits both source aggregate information and scientometric indicators. Such a model is usually defined as a DEA problem. In this case, the efficiency is computed as the result of the objective function of a DEA linear program. For each research center, the computed efficiency value is merged with the scientometric indicators and the source information.
3. *Clustering* – This module provides an organization of DMUs into homogeneous groups according to both source and derived information.

Note that in the computation of the efficiency of DMUs we also used a model selecting from the set of attributes the input parameters and the output parameters. In Section 5, we will show how different models (i.e. different selections of attributes) lead to different behavior which could lead to different classifications of research centers.

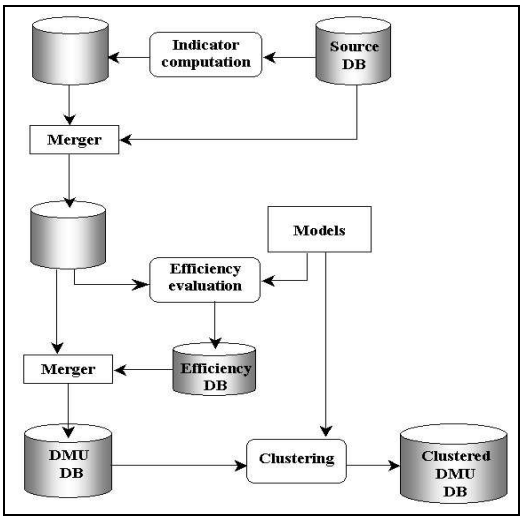


Figure 1: The research center analysis system.

4.1 Scientometric indicators

Scientometric indicators (Schubert, 1988; Galante et al., 1998; Okubo, 1997) aim at measuring the output of scientific and technological research through data derived not only from scientific literature but from patents as well. We used two scientometric indicators concerning scientific publications and citations and defined as follows.

Definition 1 *Let S be a set of scientific publications, r be a research center, and c be a scientific discipline. The* Activity Index *of r with respect to a category c is defined as $AI_r^c = \frac{P_r^c \,/\, P^c}{P_r \,/\, P}$, where P_r^c is the number of publications of r belonging to category c, P^c is the total number of publications belonging to category c, P_r is the number of publications of r, and P is the total number of publications in S.* □

Definition 2 *Let r be a research center, y be a fixed year, and S be a set of scientific publications related to r in the year y. The* Relative Citation Rate (RCR) *of r in the year y is defined as $RCR = \frac{Q \,/\, J}{F \,/\, J}$, where J is is the number of publications contained in S, Q is the number of citations received by publications in S in the years y, $y+1$, $y+2$, and F is the sum of Impact Factors of journals publishing each item in S_r in y.* □

The journal Impact Factor is a measure of the frequency with which the "average article" in a journal has been cited during a given year. As a consequence, RCR provides a measure of the incoming citations for all items in S with respect to the expected citations. The above two indicators, together with information contained in the source database, will be used to compute the efficiency.

4.2 Efficiency Evaluation

As described in Section 3, a suitable way to compute the efficiency of DMUs is to solve a system of DEA linear programs (one for each DMU) according to a given model stating the relevance of source aggregate information and indicators. The results of the DEA problems consist of the values assigned to the weights which maximize the objective functions (i.e. efficiency of DMUs). In the following, we will define different DEA models each of which is based on different selections of attributes that will be used, respectively, as input and output parameters.

In order to apply linear programming methods, a DEA problem needs to be converted into a linear form. This can be obtained by setting the denominator of the objective function equal to a constant (e.g. 1) and maximizing its numerator. The resultant DEA problem for a given DMU j is defined as follows:

$$\begin{array}{ll} max & E_j = \sum_{o=1}^{O} w_{oj} y_{oj} \\ subject\ to & \sum_{i=1}^{I} v_{ij} x_{ij} = 1 \\ & \sum_{o=1}^{O} w_{oj} y_{ol} - \sum_{i=1}^{I} v_{ij} x_{il} \leq 0 \\ & w_{oj}, v_{ij} \geq \varepsilon \\ where & l \in \{1 \cdots N\}, i \in \{1 \cdots I\}, o \in \{1 \cdots O\}. \end{array}$$

Note that the introduction of the first constraint, that normalizes the weighted sum of inputs, leads to the transformation of the problem in linear form.

4.3 Clustering of DMUs

Clustering of DMUs aims at identifying homogeneous groups of DMUs similar from the scientific activity point of view. Formally, the problem can be stated as follows: given a set $\mathcal{U} = \{u_1, \ldots, u_N\}$ of DMUs, find a suitable partition $\mathcal{P} = \{C_1, \ldots, C_k\}$ of $\mathcal{U}$ in k groups such that each group contains a homogeneous subset of DMUs.

In our context, the notion of homogeneity can be measured by exploiting, as attributes of DMUs, the information previously presented. Each DMU is represented as a multidimensional vector (Baeza-Yates and B. Ribeiro-Neto, 1999). Moreover, to our purposes it is particularly convenient to adopt a Euclidean metric, since all the attributes have numeric values. However, if the Euclidean metric is used directly, some attributes (such as the ones corresponding to absolute indicators) can exhibit a dominant effect over other ones that have a smaller scale of measurement. In order to avoid this, for each DMU j we normalize all the attribute values to fall within the range [0,1]. For each attribute z_{pj}, the corresponding attribute with normalized value is defined as $a_{pj} = \frac{z_{pj} - min(\mathbf{z}_p)}{max(\mathbf{z}_p) - min(\mathbf{z}_p)}$, where $z_{pj} \in \{x_{1j}, \ldots, x_{Ij}, y_{1j}, \ldots, y_{Oj}, E_j\}$ is the actual value of the p-th attribute of DMU j, $\mathbf{z}_p = \{z_{p1}, \ldots, z_{pN}\}$ is

the set of values assigned to the same attribute of distinct DMUs, and $max(\mathbf{z}_p)$ and $min(\mathbf{z}_p)$ compute, respectively, the maximum and the minimum value over all DMUs.

5 CLASSIFICATION OF RESEARCH CENTERS

Data Description

Our source database is composed of data related to research centers in the agro-food sector. In particular, we have collected more than 3600 projects and 8800 scientific publications, covering the period 1983-2000. We have also collected 2000 European or international patents, mostly those of 1999. Information about patents come from the PATLIB Center, an Italian information center for patents, whereas information about projects and scientific publications has been retrieved mostly through the CORDIS (Community Research and Development Information Service) site. In addition, we obtained information on about 15000 scientific publications with their bibliographic references, related to the years 1998, 1999 and 2000.

For each research center r we extracted and stored information which comprise the attributes described in Table 5.

Table 1: Attributes of research centers.

attribute	description
NPrj	Nr. of projects in which r is involved
NPub	Nr. of scientific publications financed by r
NPat	Nr. of patents financed by r
NCit	Nr. of incoming citations of publications financed by r
AI	AI value for r
RCR	RCR value for r

DEA models

In order to measure the efficiency of research centers we defined different DEA models, by considering different combinations of input and output attributes. The models used in our experiments are reported in Table 5, where we considered related attributes once (e.g. we considered either *NPub* or AI and either *NCit* or RCR). Observe that two models (M_7 and M_8) take in input the efficiency computed by other models (M_1 and M_2).

It is worth noticing that the models differently define the input and the output parameters used in the DEA linear programs. For instance, in the first model (M_1), we measured the efficiency of the research centers that have been involved in projects, evaluating their productivity in terms of patents and scientific publications. In the last two models we tried to assess efficiency variations of organizations during the time period by using a global efficiency measure (e.g. $E(M_1)$ and $E(M_2)$) and the parameters related to the number of citations (e.g. $NCit$ and RCR).

Table 2: Models for efficiency evaluation.

model	input param.	output param.
M_1	[NPrj]	[NPub, NPat]
M_2	[NPrj]	[AI, NPat]
M_3	[NPrj]	[NPat, NPub, NCit]
M_4	[NPrj]	[NPat, AI, RCR]
M_5	[NPrj]	[NCit, NPub]
M_6	[NPrj]	[RCR, AI]
M_7	[E(M_1), NPub]	[NCit]
M_8	[E(M_2), AI]	[RCR]

Clustering results

DMUs could be clustered on the basis of their efficiency computed using the DEA technique. DEA usually provides good results because it assesses the relative efficiency values by choosing the favorite weight sets for each DMU. However, in some cases, the evaluation can be more affected by the choice of the weights than by the attribute values of DMUs. Consider, for instance, two clusters based on the efficiency values calculated by model M_6 reported in Table 5. Observe that the partition is quite good, but the first cluster, which is characterized by high efficiency values, contains an outlier, DMU 7, whose scientific features are very close to the second cluster. In this case, very low input values (for the attribute $NPrj$) misleadingly result in a high efficiency value.

Table 3: Classification of DMUs based on M_6 model.

DMU	NPrj	RCR	AI	E
...	...	...	...	...
70	0.001	0.501	0.410	0.722
96	0.001	0.257	0.562	0.828
9	0.001	0.600	0.501	0.877
39	0.001	1.000	0.480	1.000
7	0.001	0.098	0.740	1.000
...	...	...	...	...
42	0.023	0	0.794	0.267
65	0.015	0	0.659	0.296
...	...	...	...	...

The above observation suggest a different classification of DMUs where the clustering algorithm takes into account, other than the efficiency computed by

means of DEA technique, also source aggregate data and scientometric indicators. We performed the k-Means algorithm on several experiments trying different k combinations for each model defined previously. As an example, in the portion of data reported in Table 5 we considered all the attributes together with the efficiency value. The clustering of DMUs, under the model M_6, assigned DMU 7 to cluster 7 instead of cluster 4. This solution is more appropriate as DMU 7 is very close to the other DMUs in cluster 7, whereas the degree of similarity between DMU 7 and DMUs belonging to cluster 4 is very low.

Table 4: Clustering of DMUs based on M_6 model.

DMU	NPrj	RCR	AI	E		cluster
...	...	...	...	...		...
70	0.001	0.501	0.410	0.722		4
96	0.001	0.257	0.562	0.828		4
9	0.001	0.600	0.501	0.877		4
39	0.001	1.000	0.480	1.000		4
...	...	...	...	...		...
42	0.023	0	0.794	0.267		7
65	0.015	0	0.659	0.296		7
7	0.001	0.098	0.740	1.000		7
...	...	...	...	...		...

To sum up, the proposed technique for classifying research centers on the basis of their performance consists in two main steps: *i) efficiency evaluation*, which is performed using DEA based techniques, and *ii)* clustering of DMUs, which considers the efficiency values together with other model attributes. The first step provides a value that expresses the relative performance for each research center, while the second one acts as a further refinement through the classification of research centers so that DMUs with similar efficiency values can be assigned to different clusters. In some sense, this process is similar to the identification of relevant web pages (corresponding to DMUs with high efficiency values) and the identification of web communities (clusters of web pages with high numbers of co-citations[1]). Obviously, if we derive large clusters, the clustering process can be further refined by applying the algorithm to the distinct clusters.

6 APPROXIMATE EFFICIENCY MEASURE

The problem in measuring the efficiency with the above approach is that the DEA technique can be computationally expensive and cannot be applied to large datasets such as those currently used in Data Mining. In fact, the computation of the efficiency of DMUs consists in the resolution of N DEA linear programs whose solutions give us a suitable combination of weights that maximizes the objective function. DEA is good at estimating the "relative" efficiency but not the "absolute" efficiency of DMUs; it can tell you how well you are doing compared to your peers but not compared to a "theoretical maximum".

As said before, a crucial issue in DEA problems is the computational complexity. To address such an issue, we propose an alternative way to compute the efficiency of DMUs. Our idea is to define an approximation of the DEA-efficiency measure, by simply considering the objective function of a DEA model (provided that suitable weights are given), and then normalizing all the attributes as explained in Section 4.3. Formally, our approximate efficiency measure is defined as:

$$\eta_j = \frac{\sum_{o=1}^{O} w_o y_{oj}}{\sum_{i=1}^{I} v_i x_{ij}}$$

In order to minimize $|E'_j - \eta'_j|$, where E'_j and η'_j denote the normalized values of E_j and η_j respectively, suitable weight sets for the computation of η_j have to be found.

6.1 Weight assignments

For each model M, obtained by selecting a set of I input attributes and a set of O output attributes, we defined the *input assignment set*, denoted by V, as the list of values assigned to the weights of the input attributes; in an analogous way, we defined the *output assignment set*, denoted by W.

Note that, since η has a fractional form and η' denotes the normalized value of η, some weight assignments can provide the same values of η'. In such a case we say that the two weight assignments are equivalent.

Definition 3 *Two weight assignments $\Phi_1 = [V_1, W_1]$ and $\Phi_2 = [V_2, W_2]$, used to compute the approximate efficiency measures η_1 and η_2 respectively, are equivalent if $\eta'_1 = \eta'_2$.* □

Moreover, a sufficient condition to assess the equivalence of two assignments is the proportionality respectively between input and output weight values. Formally, this can be stated by the following proposition:

Proposition 1. *Two weight assignments Φ_1 and Φ_2 are equivalent if $\frac{V_{1_i}}{V_{2_i}} = c_1, \forall i \in \{1, \ldots, I\}$ and $\frac{W_{1_o}}{W_{2_o}} = c_2, \forall o \in \{1, \ldots, O\}$, where c_1 e c_2 are constants.* □

[1]Two web pages are "similar" if there is a significant number of pages containing links to both of them.

As a consequence, we have the subsequent corollary:

Corollary 1 *For each assignment* $\Phi = [V, W]$ *there exists a corresponding equivalent assignment* $\hat{\Phi} = [\hat{V}, \hat{W}]$ *such as* $\hat{V}[1] = 1$ *and* $\hat{W}[1] = 1$. □

From a practical point of view, the above corollary means that we can perform a comparative analysis by setting an element of $\hat{V}$ and an element of $\hat{W}$ to 1, and then trying different combinations for the remaining attribute weights. Thus, the number of parameters is reduced to $I + O - 2$.

6.2 Experimental results

To evaluate the effectiveness of our approximate efficiency measure, we carried out a comparative analysis trying different combinations for the attribute weights. We performed experiments on two different datasets, containing respectively 540 and 134 research centers. We have used the models M_1 and M_2 for the largest dataset and the other models for the smallest dataset. Table 6.2 shows two different value assignments for the attribute weights, for each model. The vectorial notation matches the list of attributes selected for each model (see Table 5).

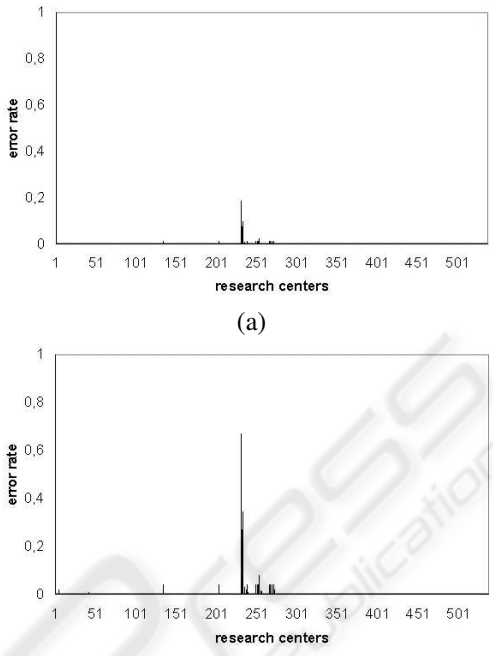
Table 5: Best settings of attribute weights.

model	Φ_1	Φ_2
M_1	[1], [1, 1]	[1], [1, 0.01]
M_2	[1], [1, 1]	[1], [1, 0.001]
M_3	[1], [1, 1, 1]	[1], [1, 0.1, 0.1]
M_4	[1], [1, 1, 1]	[1], [1, 0.1, 0.1]
M_5	[1], [1, 1]	[1], [1, 0.001]
M_6	[1], [1, 1]	[1], [1, 0.1]
M_7	[1, 1], [1]	[1, 0.001], [1]
M_8	[1, 1], [1]	[1, 0.001], [1]

Figure 2 shows a comparison of the η measure with respect to DEA-efficiency measure (i.e. $|E'_j - \eta'_j|$) relative to the model M_1. As we can see, high error peaks are very few, whereas most of the error values are below 0.2 and such a behavior is also confirmed for the remaining models. Thus, the η measure works as a good approximation of the DEA-efficiency. Moreover, we can take advantage of the fact that an approximate efficiency measure, such as η, allows an optimal trade-off between accuracy and efficiency, since its computation is not as expensive as solving a DEA problem.

6.3 Clustering quality results

To evaluate the outcome of a clustering process, it is important to check whether the computed clusters can



(a)

(b)

Figure 2: Error rates of η measure with respect to DEA-efficiency measure ($|E'_j - \eta'_j|$), according to Φ_1 (a) and Φ_2 (b) weight combinations.

be considered as of good quality. This can be done by comparing the clusters with an ideal categorization of DMUs. In our context, an ideal partition is defined as the result of the clustering algorithm applied to a given set of DMUs whose attributes include the DEA-efficiency measure together with source aggregate information and scientometric indicators.

In the experiments, our aim was to compare the ideal categorization $\Pi = \{\gamma_1, \ldots, \gamma_h\}$, of a set $\mathcal{U}$ of DMUs, to a clustering scheme $\mathcal{P} = \{C_1, \ldots, C_k\}$ of a set $\mathcal{U}'$, where $\mathcal{U}'$ was derived from $\mathcal{U}$ by replacing all DEA-efficiency values with the corresponding η efficiency values. The quality of $\mathcal{P}$ with respect to Π can be evaluated by exploiting several quality measures. In this work, we used the standard *F-measure* (Baeza-Yates and B. Ribeiro-Neto, 1999): higher values of the measure mean higher quality of clusters. Values close to the range $[0.7, 1]$ are typical of good clusters.

We performed several experiments for each model with a different number of clusters. Figure 3 contains the summarized information for the case of 20 clusters. The high values of F-measure suggest that our η efficiency measure is a good approximation of DEA-efficiency for all the models. Moreover, there exists a model, M_7, such that the approximated technique provides the same results and this behavior is valid for any number of clusters. This means that the DEA

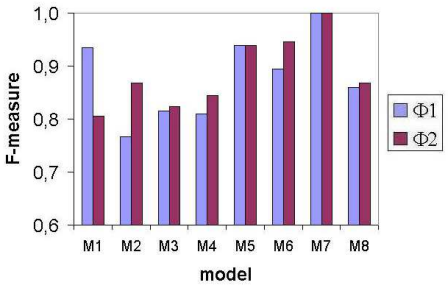


Figure 3: Clustering quality results.

efficiency measure can be substituted with the approximate measure, that improves the performance of our technique. This is particularly important in the case of large datasets.

It is important to note that while DEA techniques are non-parametric (i.e. the weight of parameters is computed by solving linear systems), in the computation of the approximate efficiency we have to assign a weight to the parameters. Our experiments have shown that the assignment of arbitrary weight values (selected without knowing the productive function), for some models, gives a good approximation of DEA (e.g. model M_7). In any case, in order to choose a good set of values for the weights, we can compare DEA and the approximate technique on small datasets.

7 CONCLUSIONS

We have presented a technique for the classification of organizational units, such as research centers, according to information on the volume and the quality of their scientific activity. Such information involves aggregate data and scientometric indicators and allows the computation of efficiency values for the productivity of research centers. We also proposed an alternative efficiency measure which exhibits a good approximation of DEA, but with the advantage of not requiring the resolution of N linear programs. The classification process, based on clustering algorithms, was tested in several experiments, showing a high degree of efficiency and effectiveness in the research center context.

REFERENCES

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. ACM Press Books, Addison Wesley.

Charnes, A., Cooper, W. W., Lewin, A. Y., and Seiford, L. M. (1994). *Data Envelopment Analysis: Theory, Methodology and Applications*. Kluwer Academic Publishers.

Charnes, A., Cooper, W. W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research* 2, 429–444.

Chung, S. H., Yang, Y. S., and Wu, T. -H. (2000). Evaluating the Efficiency of University via DEA approach. In *Proc. 5th Annual Int. Conf. on Industrial Engineering Theory, Applications and Practice*.

Emrouznejad, A. (2001). An Extensive Bibliography of Data Envelopment Analysis. Tech. Rep., Business School, Univ. of Warwick,

Farrell, M. J. (1957). The measurements of productive efficiency. *J.R. Statis Soc.*, Series A 120, 253–281.

Farrell, M. J., and Fieldhouse, M. (1962). Estimating efficient production functions under increasing returns to scale. *J.R. Statis Soc.*, Series A 125, 252–267.

Galante, E., Sala, C., and Lanini, I. (1998). *Valutazione della ricerca agricola*, Franco Angeli (ed.), Milano.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002). Cluster Validity Methods: *Sigmod Record* 31(2), 40-45.

Jain, A. K., and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall advanced reference series.

Okubo, Y. (1997). Bibliometric indicators and analysis of research systems: Methods and examples. OECD, WP#1.

Schubert, A., Glaenzel, W., and Braun, T. (1988). Against Absolute Methods: Relative Scientometrics Indicators and Relational Charts as Evaluation Tools. *Handbook of Quantitative Studies of Science and Technology*, Van Ran A. F. J. (ed.), North-Holland, Amsterdam.

Sohn, S. Y., and Choi, H. (2001). Ensemble Based on Data Envelopment Analysis. In *Proc. Aspects of Data Mining, Decision Support and Meta-Learning*, 129–137.

Stern, Z. S., Mehrez, A., and Barboy, A. (1994). Academic departments efficiency via DEA. *Computers and Operations Research* 21(5), 543–556.

Thanassoulis, E., Dyson, R. G., and Foster, M. J. (1987). Relative Efficiency Assessments using Data Envelopment Analysis: an Application to Data on Rates Departments. *J. Opl. Res. Soc.* 38, 397–412.

Viveros, M. S., Nearhos, J. P., and Rothman, M. J. (1996). Applying Data Mining Techniques to a Health Insurance Information System. In *22th VLDB Conf.*, 286–294.

Zhu, J. (2002). *Quantitative Models for Performance Evaluation and Benchmarking: Data Envelopment Analysis with Spreadsheets and DEA Excel Solver*. Kluwer Academic Publishers, Boston.