

# Clustering Gene Expression Data with Stepwise Data Envelopment Analysis

Masako Hoshino

m-hoshino@ina-lab.it.aoyama.ac.jp

Hiroshige Inazumi

hiro@ina-lab.it.aoyama.ac.jp

Graduate school of Science and Engineering, Aoyama Gakuin University,  
5-10-1 Fuchinobe, Sagamihara, Kanagawa 229-8558, Japan

**Keywords:** Data Envelopment Analysis (DEA), Gene Expression Analysis, Clustering

## 1 Introduction

DNA microarray technology has now made it possible to monitor the expression levels of thousands of genes simultaneously during important biological processes and across collections of related samples. Usually, gene expression matrix has several particular macroscopic phenotypes of samples. However, this matrix has few samples, and vast amounts of genes. This feature makes it difficult to classify samples correctly. A first step toward addressing this difficulty is the use of clustering techniques. In this paper considering Data Envelopment Analysis (DEA)[1], we present a new clustering algorithm for gene expression data using Stepwise DEA.

## 2 Method and Discussion

Data Envelopment Analysis (DEA) solves optimization problems with multiple input/output models, which is commonly used to evaluate the efficiency of a number of Decision Making Units, DMUs, by comparing against a peer directly. The heart of the analysis lies in finding the *efficient* DMUs, e.g., companies, organization and so on. DMUs providing outputs with less inputs are defined to be *efficient*, where the set of weights for inputs/outputs assigned to objective DMU is assured to be most favorable for itself. The procedure of evaluating each  $DMU_j, j = 1, 2, \dots, n$ , can be formulated as a linear programming, where  $DMU_j$  is characterized by 2 parameters,  $\theta_j$  and  $\lambda_j$ . A parameter  $\theta_j, 0 < \theta_j \leq 1$ , is the measure of  $DMU_j$ 's efficiency, and a parameter  $\lambda_j = (\lambda_{j1}, \lambda_{j2}, \dots, \lambda_{jn})$  is a vector describing the percentages of reference to the efficient DMUs. Let  $DMU_i$  be efficient and  $DMU_j$  be inefficient,  $\theta_i = 1, 0 < \theta_j < 1$ , and  $\lambda_{ii} = 1, \lambda_{jj} = 0, \forall k \neq i, \lambda_{ik} = 0, \forall k \neq j, 0 \leq \lambda_{jk} < 1$ . By using such parameters, DMUs can be classified based on the characteristics of input-output relations.

Applying DEA to a gene expression matrix of microarray, each cell sample is assumed to be the entity responsible for gene expression processes. Then, samples with typical gene combination patterns of higher expression values are defined to be *efficient*. Therefore, each sample is assigned to be either *efficient* or *inefficient*. Moreover, since references between them are derived, a set of samples with the same class is classified based on samples's efficiency from the view point of gene expression patterns.

However, it is difficult to cluster a set of samples with DEA in practice. This is because of some efficient samples with no references from inefficient samples, and inefficient samples with multiple references. For discovering the interesting subclusters with typical gene combination patterns, the problems to be solved are how to define strictly the relationship between samples by useful parameters,  $\theta$  and  $\lambda$ , and by "weights". We have proposed a heuristic or ad hoc methods previously[3]. In this paper, we assume that the samples' relationships should be defined from not only reference's

parameter,  $\lambda$ , but also efficiency parameter,  $\theta$ . Moreover, extract the efficient samples with tight references from others. Then, DEA analysis is allowed to be repeatedly applied to the subset of samples after excluding them. We propose such kind of DEA applying to gene expression analysis as a new clustering method, "Stepwise DEA". A sketch of the method is as follows:

For the first step, DEA derives efficient samples, which show remarkable expression pattern, and inefficient samples, of which expression pattern corresponds to some efficient samples. Group of inefficient samples which corresponds to an efficient sample is called its "reference group". Any sample belonging to the reference group is characterized by the group's efficient sample. Then,  $f_1(\theta, \lambda)$  is defined as relationship for each inefficient sample to its corresponding efficient samples. If the efficient sample has the reference group, where inefficient samples satisfy the condition,  $f_1(\theta, \lambda) > threshold$ , such sample is defined to be the efficiency frontier of 1st level.

At the next step, DEA is carried out for the set of samples from which the efficiency frontier's samples are excluded. This derives new efficiencies and references. So,  $f_2(\theta, \lambda)$  is calculated with new  $\theta$  and  $\lambda$  for each inefficient sample. As with the first step, efficient sample having the reference group at this step is defined to be the efficiency frontier of 2nd level. After repeating these operations, every sample is able to be characterized by sum of evaluation functions of each step. As a result, the whole samples are classified into clusters, and each cluster is specified by some typical efficient samples.

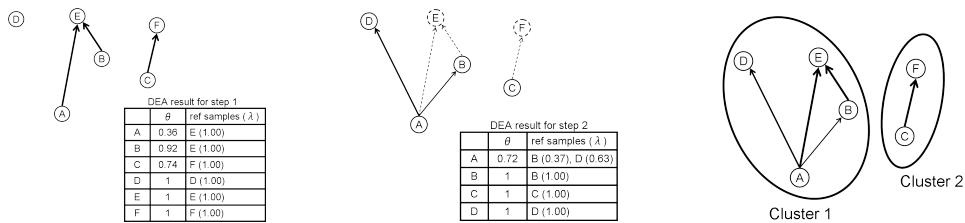


Figure 1: step 1

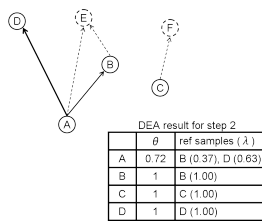


Figure 2: step 2

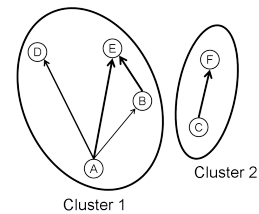


Figure 3: Clusters

A simple example shown in Figs, describes how six samples, A, B, ..., F, are clustered in Stepwise DEA. The values of  $\theta$  and  $\lambda$  are shown in tables, and reference relations derived from DEA are shown with arrows. As shown in Fig.1, sample D does not have a reference group at step1. However, sample A turns to refer sample D at the next step, because sample E and F was excluded from the set of samples.(Fig.2) According to the result of step2, only sample A is left for the next sample subset. Then the process is terminated and the clustering will be carried out with the results of step1 and 2. As shown in Fig.3, two clusters are defined. Each cluster's character will be specified by sample E and F respectively. Also, sample D can be considered to support specifying Cluster1.

For discovering useful subclusters, e.g., T-cell and B-cell, we applied this method to the well known Leukemia data, 48 ALL samples and 25 AML samples[2]. Genes used for clustering were selected with also DEA based method proposed previously. Stepwise DEA has suggested the existence of subclusters for each class. Since the efficient sample at step1 has the most prominence pattern, we may mention it as a "core sample" of each subcluster. Core samples can be identified easily, so subclusters suggested by our method can show the typical pattern for each cluster.

## References

- [1] Charnes, A., Cooper, W. W. and Rhodes, E, Measuring the Efficiency of Decision Making Units, *European Journal of Operational Research*, 2(6):429–444, 1978.
- [2] Golub, T.R., Slonim, D.K., *et al.*, Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science*, 286:531–537, 1999.
- [3] Hoshino, M. and Inazumi, H., A Step Towards Substructure Exploration from Gene Expression Patterns, *The Fifteenth International Conference on Genome Informatics 2004*, P043, 2004.