

A Step Towards Substructure Exploration from Gene Expression Patterns

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-shi, Kanagawa 229-8558, Japan

Keywords: Data Envelopment Analysis (DEA), substructures-clustering, gene expression analysis

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. Within a gene expression matrix, there are usually several particular macroscopic phenotypes of samples. Cluster analysis seeks to partition a given data set into groups based on specified features. The goal of sample-based clustering, which regards the samples as the objects and the genes as the features, is to find the phenotype structures or substructures of the samples.

In this paper, as a viewpoint of substructures-clustering, we propose a new clustering technique based on Data Envelopment Analysis, DEA,[1] where a coherent gene expression pattern characterizes the common trend of expression levels for a group of co-expressed genes. Note that this technique is applied to the compressed gene expression matrix after the step of informative gene selection through supervised classification.

2 Method

Data Envelopment Analysis 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. If DMU_A shows the same output with less values of inputs than any other DMU, then DMU_A is defined to be *efficient*. Otherwise, DMU_A is *inefficient*. Although there exist some efficient DMUs simultaneously in accordance with combination of each input-value.

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, θ , and λ_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 inputs-output relations.

Applying DEA to a gene expression matrix, a data set is necessarily to be represented as a set of samples, of which features are specified by the reciprocal gene expression values as inputs, and constant value as output. Then, samples with relative small inputs, i.e., some higher gene expression patterns, are defined to be *efficient*. Let w_{ij} be j th gene expression value of i th sample, an input value, z_{ij} , is derived from the reciprocal of w_{ij} , i.e., $z_{ij} = 2 - (w_{ij} - \min_j) / (\max_j - \min_j)$, where \max_j and \min_j are the maximum and minimum value of j th gene expression. Therefore, a set of samples are labeled to either *efficient* or *inefficient* through DEA.

Samples are classified and grouped through following two steps.

1. For each inefficient sample, link it to some efficient samples according to reference percentage, λ , over threshold t ($0 \leq t < 1$) as a similarity measure of expression patterns. Samples linked in this step are regarded as strongly relevant one. Then, any group is constructed by a set of some linked samples.
2. For efficient samples with no link, "independent samples", Assign them to feasible group applying k-nearest neighbor algorithm. Link each of them to weak relevant efficient sample, and reconstruct any group with all linked samples including weak linkage.

As the result of DEA, each set of samples is grouped based on expression patterns, where efficient samples without independent samples are assumed respectively as a core of each group, and show a typical trend of expression level of each group.

3 Discussion

We discuss the DEA-based substructures of the well known Leukemia data[2], comparing with ordinary clustering results. The data consisted of 48 acute lymphoblastic leukemia (ALL) samples and 25 acute myeloid leukemia (AML) samples. After the step of informative gene selection through supervised classification[3], substructures-clustering is proceeded. Comparison between the results of K-means clustering algorithm and DEA-based clustering algorithm is shown in Fig.1. Each row of Fig.1 describes one group made by DEA, and each column describes one cluster made by K-means. The sample ID is made up with class alphabet and its number. Each class alphabet, "L" and "M", is for ALL and AML respectively. While a bold type describes efficient sample, an italic type describes inefficient sample.

The result suggests that, assuming almost the same accuracy as ordinary clustering algorithm, DEA-based substructure-clustering shows the new aspect of common trend of expression level.

		K-means						
		# 1	# 2	# 3	# 4	# 5	# 6	# 7
DEA + kNN	A	<i>L19, M22</i>				<i>M1, M2, M5, M6, M7, M9, M10, M11, M14, M17, M19, M21</i>	M3, M8, M23	M4, M13, M15, M18
	kNN A					L10, M20, M24	M12, M25	M16
	B	<i>L2, L3, L4, L5, L6, L9, L11, L12, L17, L30, L31, L32, L35, L36, L41, L42, L46, L47</i>	<i>L14, L33, L38</i>	L13	<i>L20, L27</i>	<i>L45</i>		
	kNN B	L1, L28, L29	L7, L8, L34, L37, L43, L48					
	C	<i>L15, L16, L40</i>	L18					
	kNN C							
	D	<i>L39</i>	L44					
	kNN D							
	E				<i>L22, L23, L24, L25, L26</i>			
	kNN E				L21			

Figure 1: Example Result of K-means clustering algorithm and DEA-based clustering algorithm

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 gene expression data analysis using the method of Data Envelopment Analysis(DEA), *SIG-KBS-A304*, 151–157, 2004. (in Japanese)