유전자발현데이터의 군집분석을 위한 나무 의존 성분 분석

김종경° 최승진
포항공과대학교 컴퓨터공학과
{blkimjk°, seungjin}@postech.ac.kr

Tree-Dependent Components of Gene Expression Data for Clustering

Jong Kyoung Kim° Seungjin Choi
Department of Computer Science
Pohang University of Science and Technology

요 약

Tree-dependent component analysis (TCA) is a generalization of independent component analysis (ICA), the goal of which is to model the multivariate data by a linear transformation of latent variables, while latent variables fit by a tree-structured graphical model. In contrast to ICA, TCA allows dependent structure of latent variables and also consider non-spanning trees (forests). In this paper, we present a TCA-based method of clustering gene expression data. Empirical study with yeast cell cycle-related data, yeast metabolic shift data, and yeast sporulation data, shows that TCA is more suitable for gene clustering, compared to principal component analysis (PCA) as well as ICA.

1. INTRODUCTION

Clustering genes expression data into biologically relevant groups, is a valuable tool for finding characteristic expression patterns of a cell and for inferring functions of unknown genes [1]. On one hand, classical clustering methods such as k-means, hierarchical clustering, self-organizing map (SOM), have widely been used in bioinformatics. On the other hand, linear latent variables models were recently used in the task of gene clustering. The underlying assumption in linear latent variable models, is that gene expression profiles are generated by a linear combination of linear modes (corresponding to prototype biological processes) with weights (encoding variables or factors) determined by latent variables. Clustering gene profiles can be carried out by investigating the significance of latent variables and representative biological functions directly come from linear modes of latent variable models [2]. Tree-dependent component analysis (TCA) is a generalization of ICA, the goal of which is to seek a linear transform with latent variables well-fitting by a tree-structured graphical model, in contrast to ICA which restricts latent variable to be statistically independent [3]. In this paper, we present a method of gene clustering based on TCA. We compare the performance of TCA to PCA and ICA, for three yeast data sets, evaluating the enrichment of clusters through the statistical significance of *Gene Ontology* (GO) annotations [4].

2. LINEAR LATENT VARIABLE MODELS

Gene expression patterns measured in microarray experiments, result from unknown generative processes contributed by diverse biological processes such as the binding of transcription factors and environmental change outside a cell [5]. Genome-wide gene expression involves a very complex biological system and the characteristics of biological processes is hidden to us. A promising way to model such a generative process, is to consider a linear latent variable model such as PCA and ICA. The linear generative model assumes that a gene profile $x_t \in R^m$ (the elements of x_t represent the expression levels of gene t at m samples or m time points) is assumed to be generated by

$$x_t = As_t + \epsilon_t, \quad t = 1,...,N, \tag{1}$$

where $A = [a_1 \cdots a_n] \in R^{m \times n}$ contains linear modes in its columns and $s_t \in R^n$ is a latent variable vector with each element s_{it} associated with the contribution of the linear mode a_i to the gene profile x_t. The noise vector $\epsilon_t \in R^m$ takes the uncertainty in the model into account and it is assumed to be statistically independent of s_t. Then the linear generative model (1) can be written in a compact form:

$$X = AS, \tag{2}$$

where $X = [X_{it}] \in R^{m \times N}$ is the data matrix with each element X_{it} associated with the expression level of gene t at sample i (or time i). The latent variable matrix $S \in R^{n \times N}$ contains s_t for $t = 1, \ldots, N$. Given a data matrix X, latent variables S are determined by $S = WX$, where the linear transformation W is estimated by a certain optimization method. Depending on restrictions or assumptions on A and S, various methods including PCA, ICA, and TCA have been developed.

3. TCA

TCA is a generalization of ICA, where instead of seeking a linear transformation W that makes components $\{s_i\}$ independent (s_i is the ith-element of $s = Wx$), it searches for a linear transform W such that components (latent variables) $\{s_i\}$ well-fit by a tree-structured graphical model [3]. In TCA, s_i are referred to as *tree-dependent components*. In contrast to ICA, TCA allows the components s_i to be dependent and its dependency is captured by a tree-structured graphical model. Thus, it is expected that TCA will be more suitable for gene clustering than ICA, since it is more realistic in seeking hidden biological processes. Incorporating with a non-spanning tree in TCA, allows us to model inter-cluster independence, while providing a rich but tractable model for intra-cluster dependence. This is desirable for clustering since an exact graphical model for clusters of variables would have no edges between nodes that belong to different clusters and would be fully connected within a cluster. In order for non-spanning trees to be allowed, the prior term (penalty term), $\zeta(T) = \log p(T)$, was considered in [3]. The objective function involves the calculation of entropy, which requires the probability distribution of s that is not available in advance. Several empirical contrast functions were considered in [3]. These include: (1) kernel density estimation (KDE); (2) Gram-Charlier expansion; (3) kernel generalized variance; (4) multivariate Gaussian stationary process-based entropy rate. A brief overview of TCA is given below, and see [3] for more details.

4. PROPOSED METHOD FOR CLUSTERING

The procedures of TCA-based clustering are summarized below.

Algorithm Outline: TCA-Based Clustering

Step 1 [Preprocessing] The gene expression data matrix X is preprocessed such that each element is associated with $X_{it} = \log_2 R_{it} - \log_2 G_{it}$ where R_{it} and G_{it} represent the red and green intensity of cDNA microarray, respectively. Genes whose profiles have missing values more than 10% are discarded. Missing values in X are filled in by applying the *KNNimpute* [6], a method based on k-nearest neighbors. The data matrix is centered such that each row vector has zero mean. In the case of high-dimensional data, PCA could be applied to reduce the dimension, but it is not always necessary.

Step 2 [Decomposition] We apply TCA algorithm to the preprocessed data matrix to estimate the demixing matrix W and the encoding variable matrix S.

Step 3 [Gene clustering] In the case of ICA, row vectors are statistically independent. Thus clustering is carried out for each row vector of S (associated with each linear mode that is the column vector of A) In other words, for each row vector of S, genes with strong positive and negative values of associated independent components, are grouped into two clusters, each of which is related to induced and repressed genes, respectively. On the other hand, TCA reveals a dependency structure in the row vectors of S. Hence, the row vectors of S associated with a spanning tree undergo a weighted sum. These resulting row vectors (the number of these row vectors is equal to the number of spanning trees in the forest) are used for grouping genes into up-regulated and down-regulated genes. Denote by C_i the cluster associated with an isolated spanning tree determined by TCA. The up-regulated (C_i^u) and down-regulated (C_i^d) genes are grouped by the following rule:

$$C_i^u = \left\{ \text{gene } j \mid \sum_{k \in C_i} \|a_k\|_2^2 \, \text{sign}(\overline{a_k}) \, S_{kj} \geq c\sigma \right\},$$
$$C_i^d = \left\{ \text{gene } j \mid \sum_{k \in C_i} \|a_k\|_2^2 \, \text{sign}(\overline{a_k}) \, S_{kj} \leq -c\sigma \right\}, \tag{3}$$

where σ denotes the standard deviation of $\sum_{k \in C_i} \|a_k\|_2^2 sign(\overline{a_k}) S_{k,:}$, where $\overline{a_k}$ is the average of a_k and $S_{k,:}$ is the kth row vector of S. In our

experiment, we chose c=1.5.

5. NUMERICAL EXPERIMENTS

We used three publicly available gene expression time series data sets, including yeast sporulation (D1), metabolic shift (D2), and cell cycle-related data (D3) [7,8]. We have developed a software called *GOComparator* which calculates p values of GO annotations and compares the two clustering results visually by plotting the minimum p-values shared in both. It is freely available at *http://home.postech.ac.kr/~blkimjk/software.html*. We compared the performance of TCA-based clustering with PCA and ICA by using the three yeast datasets. The method of clustering with the two algorithms is very similar to TCA except that decomposition is performed by PCA and ICA, respectively. In addition, the weighted summation of tree-dependent components in the gene clustering step is not done as there are no clusters of hidden variables in the two algorithms. We also compared TCA algorithms with different empirical contrast functions: CUM, KGV, KDE, and STAT. The TCA algorithm based on Gaussian stationary process (STAT) outperforms the others for each dataset. The performance of TCA with a non-spanning tree was better than that of a spanning tree. The comparison results of three datasets are shown in Fig. 1. It confirms that TCA-based clustering outperforms PCA and ICA based-clustering. By applying PCA, we reduced the number of hidden variables in PCA and ICA based-clustering to the chosen number of clusters of TCA. Because of the computational cost of TCA, we reduced the dimension of the data vector to 10 by applying PCA for the dataset D3. For each dataset, the edge prior ζ_{ij}^{0} was chosen to $\frac{8\log N}{N}$, where N is the total number of genes.

6. CONCLUSIONS

In this paper, we have presented a method of TCA-based clustering for gene expression data. Empirical comparison to PCA and ICA, with three different yeast data sets, has shown that the TCA-based clustering is more useful for grouping genes into biologically relevant clusters and for finding underlying biological processes. The success of TCA-based clustering has confirmed that a tree-structured graph (a forest consisting of Chow-Liu trees) for latent variables is a more realistic and richer model for modelling hidden biological processes.

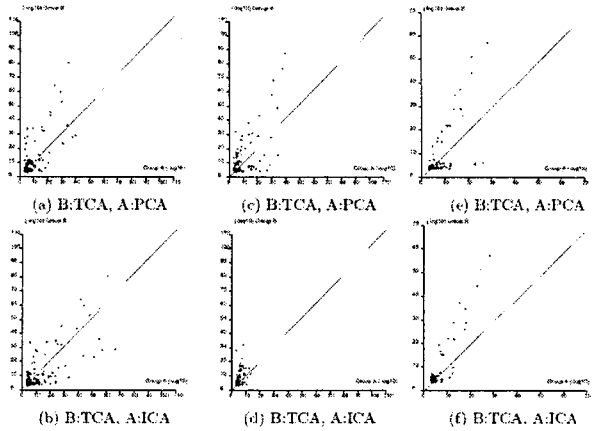

Fig.1. Comparison of TCA based clustering to PCA and ICA on three yeast datasets. For each dataset, TCA has more points above the diagonal, which indicates that TCA has more significant GO annotaions. (a),(b):D1, (c),(d):D2, (e),(f):D3

7. REFERENCES

[1] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. Nature Genetics 22:281-285, 1999.

[2] S. Lee and S. Batzoglou. ICA-based clustering of genes from microarray expression data. In Advances in Neural Information Processing Systems. 16, MIT Press, 2004.

[3] F. R. Bach and M. I. Jordan. Beyond independent components: Trees and clusters. Journal of Machine Learning Research 4:1205-1233, 2003.

[4] M. Ashburner, et al. Gene ontology: Tool for the unification of biology. Nature Genetics 25:25:29, 2000.

[5] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. Bioinformatics 18:51-60, 2002

[6] O. Troyanskaya, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 17:520-525, 2001.

[7] S. Chu, et al. The transcriptional program of sporulation in budding yeast. Science 282:699-705, 1998.

[8] J. L. DeRisi, et al. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680-686, 1997.

[9] P. T. Spellman et al. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hibridization. Molecular Biology of the Cell 9:3273-3297, 1998.