

Correlation-Based Cluster Analysis Using Mixture of Constrained PCAs

Taku Yoshioka

taku-y@is.aist-nara.ac.jp

Natsuko Kawase

natsu-ka@is.aist-nara.ac.jp

Shin Ishii

ishii@is.aist-nara.ac.jp

Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, Nara, Japan

Keywords: cluster analysis, mixture of PCAs, time-series gene expression data

1 Introduction

In a cluster analysis, similarity measure must be appropriately specified. When dealing with gene expression time-series data, it is important for similarity measure to put more emphasis on the temporal behaviors than on the scale (magnitude) of the data. The correlation coefficient is one of the most popular similarity measure, which is insensitive to the scale. It is often used in hierarchical clustering. However, there is no theoretical way to determine the number of clusters from results from hierarchical clustering. In this report, we propose a probabilistic model for correlation-based cluster analyses. The probabilistic formulation makes the determination of the cluster number a model selection problem, and the problem is statistically solved within the Bayesian framework.

2 Method and Results

In a situation where gene expression data are taken at d time points, an expression pattern of a single gene is represented by a d -dimensional vector \mathbf{y} . We assume that an expression pattern vector \mathbf{y} is produced by a noisy linear model: $\mathbf{y} = \mathbf{w}x + \epsilon$, where \mathbf{w} is a d -dimensional vector. A scalar x obeys a one-dimensional normal distribution $\mathcal{N}(x|0, 1)$ whose mean and variance are 0 and 1, respectively. Noise ϵ obeys a d -dimensional normal distribution $\mathcal{N}(\epsilon|\mathbf{0}, \sigma^{-1}\mathbf{I}_d)$. In this linear model, \mathbf{w} and x are regarded as a representative vector and a scale of the expression pattern, respectively. Since this model is a constrained version of probabilistic PCA, we refer it as to a constrained PPCA (CPCA) model. In order to deal with cluster structures, we define a mixture of CPCAs (MCPCA) (see [2] for more detail).

In the Bayesian framework, the marginal likelihood can be used for comparing models with different number of clusters. However, application of Bayesian inference to MCPCA is intractable. Then, a variational Bayes (VB) method is used for approximately conducting Bayesian inference. Since the logarithm of the marginal likelihood is approximated by the free energy in the VB method, the free energy can be used as a criterion for determining an appropriate number of clusters. After a posterior distribution of the model parameters is obtained by the VB method, the probability that expression pattern \mathbf{y} belongs to the j -th cluster $P(j|\mathbf{y})$ ($1 \leq j \leq m$) can be calculated.

Our colleagues measured expression level of 4010 genes of *Bacillus subtilis* during a sporulation process at 19 time points. We chose a subset of genes whose average expression level is larger than a certain threshold. As a result, we obtained a 19-dimensional expression pattern vector for each of the chosen 617 genes. Our clustering method was applied to this data set. We randomly prepared 50 initial conditions for each number of clusters between 1 to 14. MCPCA models with these component numbers and initial conditions were trained by the VB method.

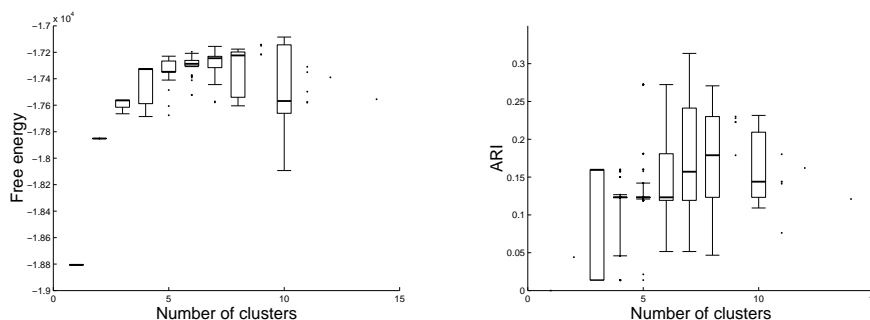


Figure 1: Evaluation of clustering results. (left) The number of clusters versus free energy. (right) The number of clusters versus ARI.

A number of genes have been biologically partitioned according to the time when they are activated during the sporulation process. To assess the proposed method, clustering results were compared to the above partition of genes by using adjusted Rand index (ARI) (see [1] for more detail). Namely, we assume ARI provides biological evaluation for clustering results.

The left (right) panel of Figure 1 shows the distribution of the free energy (ARI) for different numbers of clusters. Each thick line in a box denotes a median. The median of both of the free energy and ARI have a peak at eight clusters. This result implies that a biologically-meaningful number of clusters can be determined automatically based on a statistical criterion (i.e., free energy).

Although the free energy is correlated with the biological evaluation with respect to the number of clusters, the correlation is not strong enough to extract biological information only from a single clustering result with the largest free energy. In order to cope with this problem, we present an approach that considers an ensemble of clustering results (see [2] for more detail).

In order to extract representative expression patterns from the ensemble of clustering results, we applied an analysis assuming a CPCA model; this is a feature extraction analysis and not a cluster analysis. The result is shown in Figure 2. The representative expression patterns in the figure are consistent with biological knowledge about the sporulation process. For examples, cluster 1 includes genes activated in an early stage of the sporulation process, while genes in cluster 6 are activated in a later stage. It should be noted that no biological knowledge was required to obtain these expression patterns.

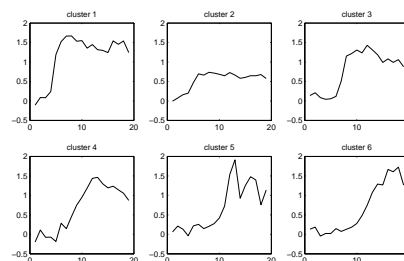


Figure 2: Representative expression patterns extracted by our ensemble procedure.

Acknowledgments

The authors are grateful to Dr. Kazuo Kobayashi and Prof. Naotake Ogasawara for providing the experimental data.

References

- [1] Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., and Ruzo, W.L., Model-based clustering and data transformations for gene expression data, *Bioinformatics*, 17(19):977–987, 2001.
- [2] Yoshioka, T. and Ishii, S., Clustering for time-series gene expression data by mixture of PCAs, *Proc. 9th Int. Conf. on Neural Information Processing*, in press.