

A kernel method for hierarchical and non-hierarchical clustering

Kenji Fukumizu

Institute of Statistical Mathematics

4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569 Japan

fukumizu@ism.ac.jp

Keywords: clustering, similarity, distance, reproducing kernel.

1. Introduction

In cluster analysis, choosing a distance measure among given data is not straightforward. For defining a similarity matrix in hierarchical clustering and some types of non-hierarchical clustering such as spectral clustering ([1]), several distance measures such as Pearson correlation, Euclidean distance, and non-parametric measures have been widely used. The result of clustering, however, often depends on a specific choice of the measure. If our purpose is to extract a group of variables, for which we cannot presume the type of dependence, a specific choice of measure may not capture the dependence of variables.

This paper proposes a novel approach to define a similarity matrix, which can incorporate any nonlinear correlation of variables. Our method uses kernel Hilbert spaces, which contain a sufficiently rich class of nonlinear functions of variables. We define the similarity of two variables based on the theory of kernel Hilbert spaces and covariance operators ([2]). This similarity represents the sum of all the nonlinear correlations in principle, and provides a more reasonable way for defining the similarity or dependence of two random variables than existing measures, which extract only partial information of variables.

2. Kernel Similarity Matrix for Clustering

Suppose we have n data of d -dimensional for clustering. Each of the d -dimensional data is represented by $X_i \in \mathbb{R}^d$ for $1 \leq i \leq n$. Let $k(x, y) = \exp(-\|x - y\|^2/\sigma^2)$ be a RBF kernel function. The *kernel similarity matrix* (D_{ij}) ($1 \leq i, j \leq n$) is defined by

$$D_{ij} = \text{Tr} [A_{ij}^T A_{ij}], \quad A_{ij} = \widehat{\Sigma}_{ii}^{-1/2} \widehat{\Sigma}_{ij} \widehat{\Sigma}_{jj}^{-1/2} \quad (1)$$

where $\widehat{\Sigma}_{ii}$ and $\widehat{\Sigma}_{ij}$ are the $d \times d$ matrices, which are the empirical estimates of the covariance operators ([2]) and defined by

$$\widehat{\Sigma}_{ii} = (G_i + \varepsilon I_d)^2, \quad \widehat{\Sigma}_{ij} = G_i G_j \quad (i \neq j). \quad (2)$$

The ε is a regularization coefficient. The $d \times d$ matrix G_i is the centralized Gram matrix defined by $G_i = Q(k(X_{ia}, X_{ib}))Q$, where Q is the orthogonal projection onto the $(d - 1)$ -dimensional subspace orthogonal to $(1, \dots, 1)^T$. The value of

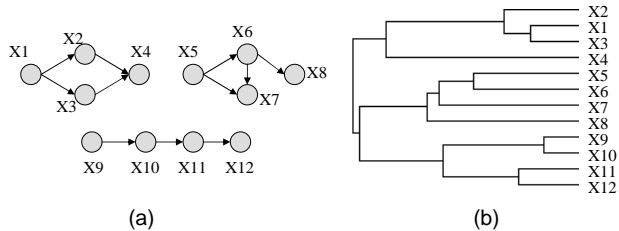


Figure 1: (a) Network for generating data. (b) Result of hierarchical clustering with the kernel similarity matrix.

D_{ij} is expected to represent all the nonlinear correlations of the d -dimensional random variables X_i and X_j ([3]).

The proposed similarity matrix can be easily used for any type of clustering methods based on a similarity matrix.

3. Experimental Result

The proposed method is applied for clustering of the synthesized data, which are 50 samples from the Bayesian network in Fig.1 (a). They are used for clustering of the 12 variables. Some variables are continuous and others are discrete, and the dependence between a node and its parents include nonlinear and non-monotonic relations. While the hierarchical clustering with Pearson correlation, Euclidean distance, and K-mean clustering do not give an appropriate result, the hierarchical clustering with the kernel similarity matrix provides correct clustering, as Fig.1 (b) shows. Also, the spectral clustering ([1]) with the kernel similarity matrix outputs the correct clusters. Application on real-world data sets are now under preparation.

References

- [1] Ng, A., M. Jordan, and Y. Weiss (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, 849–856
- [2] Fukumizu, K., F. Bach, and M. Jordan (2003). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. Tech report, Dept. Statistics, UC Berkeley.
- [3] Bach, F. R. and M. I. Jordan (2002). Kernel independent component analysis. *Journal of Machine Learning Research 3*, 1–48.