Mixed Factors Analysis: Unsupervised Statistical Discrimination with Kernel Feature Extraction

Ryo Yoshida
Department of Statitical Modeling
Institute of Statistical Mathematics
Research Organization of Information and Systems
4-6-7 Minami-Azabu, Minato-ku, Tokyo, 106-8569, Japan
yoshidar@ism.ac.jp

ABSTRACT: We address the problem of clustering and feature extraction of exceedingly high-dimensional data where the dimensionality of the feature space is much higher than the number of training samples. For such a sparsely-distributed dataset, direct application of conventional model-based clustering might be impractical due to occurrence of an over-learning. In order to overcome the limit of application, we have developed the mixed factors model which was originally aimed at solving the over-learning problem in the unsupervised discriminant analysis of gene expression profiles. The idea is to extract the feature variables involved in the underlying group structure, and then, train an unsupervized discriminative classifier by using the extracted features which are projected onto the lower-dimensional factor space. By alternating projection and clustering, the method seeks an optimal direction of projection such that the overlap of the projected clusters is small. One main purpose of this research is to elucidate the statistical machineries of the feature extraction system offered by the mixed factors model. Particularly, we give the connection to Fisher's discriminant analysis and the principal component analysis. After showing some theoretical consequences, we also attempt to present a more generic approach of clustering within the framework of kernel machine learning. By this extension, we can deal with much more complicated shapes of clusters and clustering on the generic feature spaces.

Keywords: *Clustering*, *Feature Extraction*, *Gene Expression Profiling*, and *Kernel Learning*

References

Yoshida, R., Higuchi, T., and Imoto, S. (2004): A mixed factors model for dimension reduction and extraction of a group structure in gene expression data, *Proceedings of the third Computational Systems Bioinformatics*, 161–172.

Yoshida, R., Higuchi, T., Imoto, S., and Miyano, S. (2006): ArrayCluster: An analytic tool for clustering, data visualization and module finder on gene expression profiles, *Bioinformatics*, 22, 1538–1539.