Linear Data Fusion with drCCA

Abhishek Tripathi[1], Arto Klami[2], and Samuel Kaski[2]

[1] Department of Computer Science, University of Helsinki
[2] Laboratory of Computer and Information Science, Helsinki University of Technology

We consider a data fusion problem of combining two or more data sources where each source consists of vector-valued measurements from the same object or entities but on different variables. The task is to include only those aspects which are *mutually* informative of each other. This task of including only shared aspects of data sources is motivated through two interrelated lines of thought. The first is noise reduction. If the data sources are measurements of the same entity corrupted by independent noise, discarding source-specific aspects will discard the noise and leave the shared properties that describe the shared entity. The second motivation is to analyze what is interesting in the data. One example is the study of activation profiles of yeast genes in several stressful treatments in the task of defining yeast stress response [1]. In this example what is in common in the sources is what we are really interested in. The "noise" may be very structured; its definition is simply that it is source-specific.

It turns out that a very simple and fast procedure can be used to preprocess the data such that only the shared aspects remain; it can be shown that the method is equivalent to applying generalized canonical correlation analysis, which is known to maximize the mutual information for normally distributed data, and combining the resulting components into a shared representation. The procedure is to whiten each data set to avoid detecting variation specific to a single data set, and then combine all data sets into a single low-dimensional representation using PCA on the concatenation of the whitened data sets.

Experiments on gene expression collections are used to illustrate how even such simple linear data fusion technique works effectively. We classify cell cycle regulated genes in yeast [3] and identify differentially expressed genes in leukemia [2]. Implementation of the method in R is available at `http://www.cis.hut.fi/projects/mi/software/drCCA/`.

References

1. J. Nikkilä, C. Roos, E. Savia, and S. Kaski. Explorative modeling of yeast stress response and its regulation with gCCA and associative clustering. *International Journal of Neural Systems*, 15(4):237–246, 2005.
2. M. E. Ross, X. Zhou, G. Song, S. A.Shurtleff, K. Girtman, W. K. Williams, H-C. Liu, R. Mahfouz, S. C. Raimondi, N. Lenny, A. Patel, and J. R. Downing. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, 102(8):2951–2959, 2003.
3. P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–97, 1998.