Decomposition into local data set-specific and shared effects by a mixture of probabilistic CCAs

Arto Klami and Samuel Kaski
Helsinki Institute for Information Technology and
Adaptive Informatics Research Centre;
Laboratory of Computer and Information Science
Helsinki University of Technology
P.O. Box 5400, FI-02015 TKK, FINLAND
samuel.kaski@tkk.fi, http://www.cis.hut.fi/projects/mi

Correlation and canonical correlation analysis (CCA) are classical methods for studying dependencies between two or more data sets of co-occurring (paired) samples (x, y). CCA maximizes the mutual information between representations it extracts from x and y (for normally distributed data), which makes it a predecessor of the variety of more recent methods that specialize in finding dependencies between sets of variables, including various forms of co- and discriminative clustering, information bottleneck and Imax, and kernelized CCA.

A main problem with these models is that they overfit very easily to small data sets. In practical data analysis tasks, such as data fusion in bioinformatics, the number of dimensions may even exceed the number of data points, and the existing regularization methods (such as ridge penalties for kernel-CCA) are insufficient. The recent finding [1] that CCA can be interpreted as a generative model was very promising since it opened the road to Bayesian treatments, and hence to rigorous ways of including both prior knowledge and complexity control.

The generative model of CCA assumes normally distributed data and the components are linear, which are very restrictive assumptions in practical data analysis tasks. It would make sense to make these assumptions locally, however, and search for local dependencies between data sets. This naturally requires very good complexity control methods since the effective number of data points per CCA will decrease. In addition, locality should be defined in the sense of the covariance matrix the CCA introduces, not in the original metric of the data space.

CCA in effect finds features that are shared (in the sense of being statistically dependent) by several data sets. In practical data analysis it would be even better to find both what is shared by the data sets, and what is specific to each, that is, to decompose each data set into common and data set-specific components. It is possible to extend the generative model of CCA to achieve this [3].

We introduce a fully Bayesian mixture of CCAs which chooses the decomposition to shared and data set-specific components locally using an Automatic Relevance Determination prior, and the number of mixture components using a Dirichlet process prior. The posterior is computed by a split-merge procedure based on Gibbs sampling [2], and the quantities used in data analysis, such as degree of local dependency, probability of samples to share the same dependencies, loadings of the components relevant for the dependencies, probability of samples to belong to the same clusters, etc, are computed from the posterior.

References

[1] F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.

[2] S. Jain and R. Neal. Splitting and merging components of a non-conjugate dirichlet process mixture model. Technical Report 0507, Department of statistics, University of Toronto, 2005.

[3] A. Klami and S. Kaski. Generative models that discover dependencies between data sets. In *Proceedings of Machine Learning for Signal Processing XVI*, pages 123–128, 2006.

Topic: learning algorithms
Preference: oral