What do data sets have in common? The yeast stress case

Samuel Kaski, University of Helsinki, Dept of Computer Science
Janne Nikkilä, Helsinki University of Technology, Neural Networks Research Centre
Christophe Roos, Medicel Inc.

Introduction

We develop methods for mining data sets for properties that are unknown *a priori,* but known to be common to all sets. Data comes in tuples or pairs (**x**,**y**), with **x** and **y** coming from different sets.

Yeast stress is a prototype case study. A group of yeast genes, called common environmental response (CER) or environmental stress response (ESR) genes, responds to stress. Their behavior should, by definition, have common properties across stress treatments (the data sets).

Objectives

Introduce general-purpose methods for extracting common properties from data sets:

- canonical correlation (CCA)-based depenency-preserving dimensionality reduction
- bootstrapped associative clustering for exploring dependencies between data sets

- Extract stress-related gene expression by the dimensionality reduction method
- Apply associative clustering to explore (regulatory) dependencies between the stress-related expression and transcription factor binding data

Methods

Dimensionality reduction by gCCA

Standard Canonical Correlation Analysis (CCA) finds pairs or components, one from **x** and one from **y**, such that the components correlate maximally. For Gaussian data the components maximize mutual information between the sets.

CCA can be computed by whitening both data sets separately, and computing principal components (eigenvectors) of the concatenated data **z**=[**x y**].

CCA can be interpreted as dimensionality reduction as follows: Whitening removes the non-interesting data set-specific variation, and only between-data variation remains. Dimensionality reduction by principal components analysis then tries to preserve this interesting variation.

This whole procedure can be generalized (gCCA) directly to several data sets: Whiten all data sets and compute principal components of the concatenated data.

Associative clustering (AC)

AC clusters two data sets such that statistical dependency between the two clusterings is maximized. The clusters detect regularities and exceptions in co-occurrences of **x** and **y**.

Clusters are defined by K-means-type prototypes: **x** belongs to cluster *i* if **x** is closest to $\mathbf{m}_i$

The two clusterings form a *contingency table* of co-occurrence counts n_{ij}. The clusters are optimized by maximizing the Bayes factor between two models, in which the contingency table is thought to arise from independent or dependent margins, respectively.

Contingency table cells are colored according to their deviance from independence. Yellow: unexpectedly many genes (implies regularities), Blue: unexpectedly few (implies outliers)

Search for reliable groups by bootstrap

We search for sets of genes that (1) occur in the same cluster pair (contingency table cell) reliably and that (2) signify dependencies between the data sets.

(1) Is taken care of by bootstrap: For each pair of genes, compute the probability that the genes occur in the same contingency table cell. This is a similarity measure.

(2) Is taken care of by only considering yellow (regular) contingency table cells.

Finally, the similarities are summarized. Here we used hierarchical clustering, and selected the set of N=51 most homogeneous clusters for further analysis.

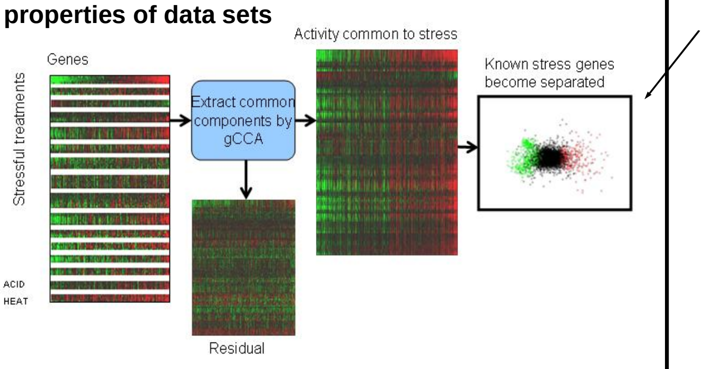
Data

Expression data: Short time series from altogether 16 stress treatments, a total of 106 time points. The data was collected from (Causton *et al*., Mol Biol Cell, 2001) and (Gasch *et al*., Mol Biol Cell, 2000). Each treatment is used as a separate data set.

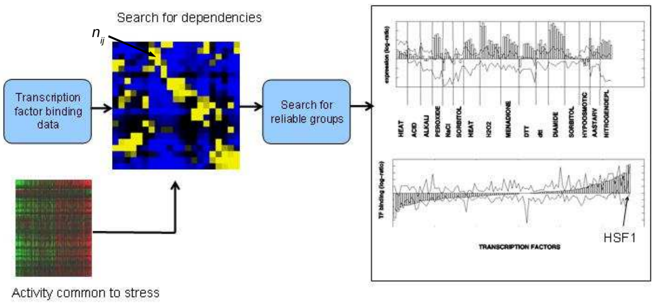
Transcription factor (TF) binding data: Binding profiles of 113 transcription factors, in the promoter region of each gene, had been measured by Lee *et al.* (Science 2002).

Log ratios for 5998 genes, profiles normalized with respect to zero point of time series (or other control). Missing values imputed by genewise averages within each data set.

Dimensionality reduction by gCCA extracts common properties of data sets


Genes
Stressful treatments
ACID
HEAT
Extract common components by gCCA
Activity common to stress
Residual
Known stress genes become separated

Dependency exploration by Associative Clustering: hunting for hints of stress regulation



Results

Dimensionality reduction by gCCA separates known stress genes

The number of canonical components (=12) was chosen to maximize mutual information in cross-validation.

These components showed significant association (Wilcoxon rank sum test, p<0.01) with known environmental stress genes (ESR, Gasch et al., Mol Biol Cell, 2000).

Already two components separate nicely up-regulated (red) and down-regulated (green) stress genes from the mass.

AC finds dependencies between stress-related expression and TF-binding data

Compared to independent K-means clusterings of both data sets, AC found significantly higher statistical dependency between the data sets (p<0.01, paired t-test in 10-fold cross-validation).

Moreover, known ESR genes are enriched in the reliable (judged based on bootstrap) clusters: upregulated ESR genes in 14 out of N=51 clusters and downregulated ESR genes in 12.

Interpretations of the dependencies

EASE (Hosack et al., Genome Biol, 2003) gives interpretations: 14 remaining even after the ultra-conservative Bonferroni correction.

Profiles of TF binding suggest regulatory interactions.

Discussion

Our exploratory models are complementary to the popular graphical models of regulation (a la Friedman and others): They are simpler-to-use in novel tasks and in that sense more general-purpose (have been used in modeling dependencies of gene expression in mouse and man, for instance).

Technical difference: We model *only the common properties* of data sets, instead of all data-specific details.

Conclusions

We have introduced methods for dependency-preserving dimensionality reduction (gCCA-based) and dependency exploration (Associative Clustering-based).

The methods work: They find stress-related genes and possibly interesting regulatory interactions.

More work is needed for interpretations.

More information: http://www.cis.hut.fi/projects/mi

samuel.kaski@cs.helsinki.fi