MA02. Supervised Statistical Data Mining and Mining of Mutual Dependency

Research group: Samuel Kaski, Jaakko Peltonen, Janne Nikkilä et al.
Website: www.cis.hut.fi/projects/mi/

We develop statistical machine learning methods for extracting (mining)useful regularities from large, high-dimensional data sets. Unsupervised mining cannot distinguish relevant from irrelevant variation in a data set. We study two solutions to this problem.

SUPERVISED MINING

In **supervised mining** one data set supervises the mining of the other. We have introduced a principle of *learning metrics*, where the supervision is learned as a distance metric. We have introduced new visualization, clustering and projection methods based on this formalism.

(1) Self-Organizing Map (SOM) in learning metrics visualizes data focusing on relevant differences (e.g. changes in financial indicators of companies that affect bankruptcy risk).

(2) Multidimensional Scaling (MDS) in learning metrics finds low-dimensional data representations and preserves distances based on relevant differences.

(3) Relevant Component Analysis finds linear projections of data that are relevant for classes of the data (e.g. projections of sound samples that discriminate different phonemes).

(4) Supervised Clustering finds clusters of data that are relevant for a variable of interest. We have a version that clusters vectorial data supervised by classes of the data (it finds e.g. clusters of gene expression profiles that are relevant to gene function), and a version that clusters discrete items supervised by co-occurring items (it finds e.g. clusters of documents that are relevant to co-occurring words).

MINING OF MUTUAL DEPENDENCY

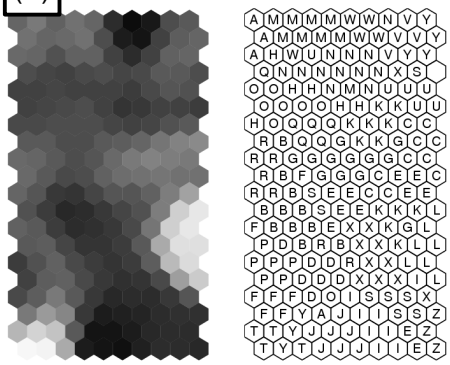
In **mining of mutual dependency** the supervision is symmetric and the task is to find dependencies *between* data sets. The data sets have shared samples but different features that supervise each other. We have introduced new clustering and component models for this task.

(5) Associative Clustering clusters two data sources and maximizes the dependency between the clusterings (e.g. maximizes dependency between clusters of human gene expression profiles and clusters of mouse gene expression profiles).

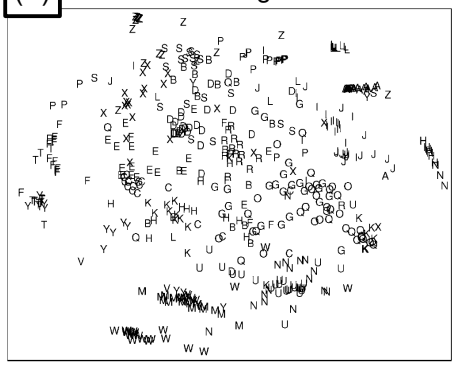
(6) Data Fusion by Canonical Correlation Analysis (CCA) integrates data from multiple sources, preserving their mutual statistical dependencies and discarding data set-specific effects (e.g. extracts common effects from multiple stress related yeast gene expression data sets to one data set that represents yeast's stress response).

(7) Non-parametric Dependent Components extends canonical correlation analysis to find linear projections that capture more general dependencies than just correlation.

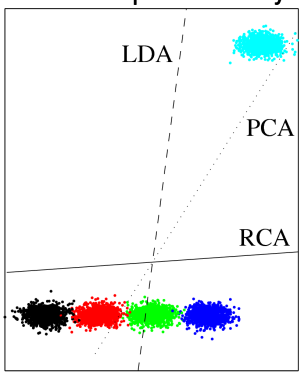
(1) SOM in learning metrics



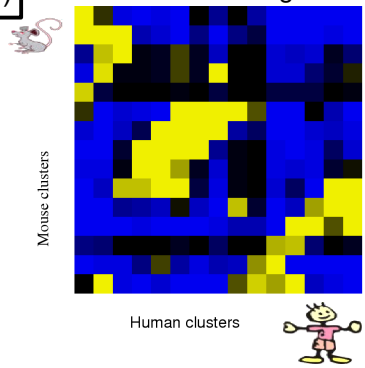
(2) MDS in learning metrics



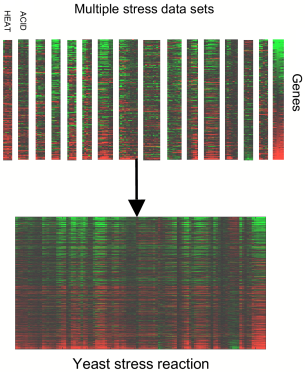
(3) Relevant Component Analysis



(5) Associative Clustering



(6) Data Fusion by CCA



(1), (2) J. Peltonen, A. Klami, and S. Kaski. Improved Learning of Riemannian Metrics for Exploratory Analysis. *Neural Networks* 17: 1087-1100, 2004.