Doctoral dissertations

Data exploration with self-organizing maps in environmental informatics and bioinformatics

Mikko Kolehmainen

Dissertation for the degree of Doctor of Science in Technology on 27 February 2004.

External examiners:
Enso Ikonen (University of Oulu)
Erkki Pesonen (University of Kuopio)
Opponent:
Jussi Parkkinen (University of Joensuu)

Abstract:

The aim of this thesis was to evaluate the usability of self-organizing maps and some other methods of computational intelligence in analysing and modelling problems of environmental informatics and bioinformatics. The concepts of environmental informatics, bioinformatics, computational intelligence and data mining are first defined. There follows an introduction to the data processing chain of knowledge discovery and the methods used in this thesis, namely linear regression, self-organizing maps (SOM), Sammon's mapping, U-matrix representation, fuzzy logic, c-means and fuzzy c-means clustering, multi-layer perceptron (MLP), and regularization and Bayesian techniques. The challenges posed by environmental processes and bioprocesses are then identified, including missing data problems, complex lagged dependencies among variables, non-linear chaotic dynamics, ill-defined inverse problems, and large search space in optimization tasks.

The works included in this thesis are then evaluated and discussed. The results show that the combination of SOM and Sammon's mapping has great potential in data exploration, and can be used to reveal important features of the measurement techniques (e.g. separability of compounds), reveal new information about already studied phenomena, speed up research work, act as a hypothesis generator for traditional research, and supply clear and intuitive visualization of the environmental phenomenon studied. The results of regression studies show, as expected, that the MLP network yields better estimates in predicting future values of airborne pollutant concentration of NO_2 compared with SOM based regression or the Least Squares approach using periodic components. Additionally, the use of local MLP models is shown to be slightly better for estimating future values of episodes compared with one MLP model only. However, it can be concluded in general that the architectural issues tested are not able to solve solely model performance problems.

Finally, recommendations for future work are laid out. Firstly, the data exploration solution should be enhanced with methods from signal processing to enable the handling of measurements with different time scale and lagged multivariate time-series. The main suggestion, however, is to create an integrated environment for testing different hybrid schemes of computational intelligence for better time-series forecasting in environmental informatics and bioinformatics.

Exploratory source separation in biomedical systems

Jaakko Särelä

Dissertation for the degree of Doctor of Science in Technology on 29 October 2004.

External examiners:
Te-Won Lee (University of California at San Diego)
Ole Jensen (Radboud University Nijmegen)
Opponent:
Lars Kai Hansen (Technical University of Denmark)

Abstract:

Contemporary science produces vast amounts of data. The analysis of this data is in a central role for all empirical sciences as well as humanities and arts using quantitative methods. One central role of an information scientist is to provide this research with sophisticated, computationally tractable data analysis tools.

When the information scientist confronts a new target field of research producing data for her to analyse, she has two options: She may make some specific hypotheses, or guesses, on the contents of the data, and test these using statistical analysis. On the other hand, she may use general purpose statistical models to get a better insight into the data before making detailed hypotheses.

Latent variable models present a case of such general models. In particular, such latent variable models are discussed where the measured data is generated by some hidden sources through some mapping. The task of *source separation* is to recover the sources. Additionally, one may be interested in the details of the generation process itself.

We argue that when little is known of the target field, *independent component analysis* (ICA) serves as a valuable tool to solve a problem called *blind source separation* (BSS). BSS means solving a source separation problem with no, or at least very little, prior information. In case more is known of the target field, it is natural to incorporate the knowledge in the separation process. Hence, we also introduce methods for this incorporation. Finally, we suggest a general framework of *denoising source separation* (DSS) that can serve as a basis for algorithms ranging from almost blind approach to highly specialised and problem-tuned source separation algoritms. We show that certain ICA methods can be constructed in the DSS framework. This leads to new, more robust algorithms.

It is natural to use the accumulated knowledge from applying BSS in a target field to devise more detailed source separation algorithms. We call this process *exploratory source separation* (ESS). We show that DSS serves as a practical and flexible framework to perform ESS, too.

Biomedical systems, the nervous system, heart, etc., constitute arguably the most complex systems that human beings have ever studied. Furthermore, the contemporary physics and technology have made it possible to study these systems while they operate in near-natural conditions. The usage of these sophisticated instruments has resulted in a massive explosion of available data. In this thesis, we apply the developed source separation algorithms in the analysis of the human brain, using mainly magnetoencephalograms (MEG). The methods are directly usable for electroencephalograms (EEG) and with small adjustments for other imaging modalities, such as (functional) magnetic resonance imaging (fMRI), too.

From insights to innovations: data mining, visualization, and user interfaces

Johan Himberg

Dissertation for the degree of Doctor of Science in Technology on 5 November 2004.

External examiners:
Sami Khuri (San José State University)
Olli Silvén (University of Oulu)
Opponent:
Juha Röning (University of Oulu)

Abstract:
This thesis is about data mining (DM) and visualization methods for gaining insight into multidimensional data. Novel, exploratory data analysis tools and adaptive user interfaces are developed by tailoring and combining existing DM and visualization methods in order to advance in different applications.

The thesis presents new visual datamining (VDM)methods that are also implemented in software toolboxes and applied to industrial and biomedical signals: First, we propose a method that has been applied to investigating industrial process data. The self-organizing map (SOM) is combined with scatterplots using the traditional color linking or interactive brushing. The original contribution is to apply color linked or brushed scatterplots and the SOM to visually survey local dependencies between a pair of attributes in different parts of the SOM. Clusters can be visualized on a SOM with different colors, and we also present how a color coding can be automatically obtained by using a proximity preserving projection of the SOM model vectors. Second, we present a new method for an (interactive) visualization of cluster structures in a SOM. By using a contraction model, the regular grid of a SOM visualization is smoothly changed toward a presentation that shows better the proximities in the data space. Third, we propose a novel VDM method for investigating the reliability of estimates resulting from a stochastic independent component analysis (ICA) algorithm. The method can be extended also to other problems of similar kind. As a benchmarking task, we rank independent components estimated on a biomedical data set recorded from the brain and gain a reasonable result.

We also utilize DMand visualization for mobile-awareness and personalization. We explore how to infer information about the usage context from features that are derived from sensory signals. The signals originate from a mobile phone with on-board sensors for ambient physical conditions. In previous studies, the signals are transformed into descriptive (fuzzy or binary) context features. In this thesis, we present how the features can be transformed into higher-level patterns, contexts, by rather simple statistical methods: we propose and test using minimum-variance cost time series segmentation, ICA, and principal component analysis (PCA) for this purpose. Both time-series segmentation and PCA revealed meaningful contexts from the features in a visual data exploration.

We also present a novel type of adaptive soft keyboard where the aim is to obtain an

ergonomically better, more comfortable keyboard. The method starts from some conventional keypad layout, but it gradually shifts the keys into new positions according to the user's grasp and typing pattern.

Related to the applications, we present two algorithms that can be used in a general context: First, we describe a binary mixing model for independent binary sources. The model resembles the ordinary ICA model, but the summation is replaced by the Boolean operator OR and the multiplication by AND. We propose a new, heuristic method for estimating the binary mixing matrix and analyze its performance experimentally. The method works for signals that are sparse enough. We also discuss differences on the results when using different objective functions in the FastICA estimation algorithm. Second, we propose "global iterative replacement" (GIR), a novel, greedy variant of a merge-split segmentation method. Its performance compares favorably to that of the traditional top-down binary split segmentation algorithm.

Data exploration with learning metrics

Jaakko Peltonen

Dissertation for the degree of Doctor of Science in Technology on 17 November 2004.

External examiners:
Hannu Toivonen (University of Helsinki)
Kari Torkkola (Motorola Labs)
Opponent:
John Shawe-Taylor (University of Southampton)

Abstract:

A crucial problem in exploratory analysis of data is that it is difficult for computational methods to focus on interesting aspects of data. Traditional methods of unsupervised learning cannot differentiate between interesting and noninteresting variation, and hence may model, visualize, or cluster parts of data that are not interesting to the analyst. This wastes the computational power of the methods and may mislead the analyst.

In this thesis, a principle called “learning metrics” is used to develop visualization and clustering methods that automatically focus on the interesting aspects, based on auxiliary labels supplied with the data samples. The principle yields non-Euclidean (Riemannian) metrics that are data-driven, widely applicable, versatile, invariant to many transformations, and in part invariant to noise.

Learning metric methods are introduced for five tasks: nonlinear visualization by Self-Organizing Maps and Multidimensional Scaling, linear projection, and clustering of discrete data and multinomial distributions. The resulting methods either explicitly estimate distances in the Riemannian metric, or optimize a tailored cost function which is implicitly related to such a metric. The methods have rigorous theoretical relationships to information geometry and probabilistic modeling, and are empirically shown to yield good practical results in exploratory and information retrieval tasks.

Linear space–time modulation in multiple–antenna channels

Ari Hottinen

Dissertation for the degree of Doctor of Science in Technology on 25 November 2004.

External examiners:
Jyrki Joutsensalo (University of Jyväskylä)
Tapani Ristaniemi (Tampere University of Technology)
Opponent:
David Gespert (Institut Eurécom)

Abstract:

This thesis develops linear space–time modulation techniques for (multi-antenna) multi-input multi-output (MIMO) and multiple-input single-output (MISO) wireless channels. Transmission methods tailored for such channels have recently emerged in a number of current and upcoming standards, in particular in 3G and ‘’beyond 3G” wireless systems. Here, these transmission concepts are approached primarily from a signal processing perspective.

The introduction part of the thesis describes the transmit diversity concepts included in the WCDMA and cdma2000 standards or standard discussions, as well as promising new transmission methods for MIMO and MISO channels, crucial for future high data-rate systems. A number of techniques developed herein have been adopted in the 3G standards, or are currently being proposed for such standards, with the target of improving data rates, signal quality, capacity or system flexibility.

The thesis adopts a model involving matrix-valued modulation alphabets, with different dimensions usually defined over *space* and *time*. The symbol matrix is formed as a linear combination of symbols, and the space-dimension is realized by using multiple transmit and receive antennas. Many of the transceiver concepts and modulation methods developed herein provide both spatial multiplexing gain and diversity gain. For example, full-diversity full-rate schemes are proposed where the symbol rate equals the number of transmit antennas. The modulation methods are developed for open-loop transmission. Moreover, the thesis proposes related closed-loop transmission methods, where space–time modulation is combined either with automatic retransmission or multiuser scheduling.

Multiagent reinforcement learning: asymmetric and symmetric approaches

Ville Könönen

Dissertation for the degree of Doctor of Science in Technology on 3 December 2004.

External examiners:
Petri Koistinen (University of Helsinki)
Kary Främling (Helsinki University of Technology)
Opponent:
Ann Nowé (Vrije Universiteit Brussel)

Abstract:

Modern computing systems are distributed, large, and heterogeneous. Computers, other information processing devices and humans are very tightly connected with each other and therefore it would be preferable to handle these entities more as agents than stand-alone systems. One of the goals of artificial intelligence is to understand interactions between entities, whether they are artificial or natural, and to suggest how to make good decisions while taking other decision makers into account. In this thesis, these interactions between intelligent and rational agents are modeled with Markov games and the emphasis is on adaptation and learning in multiagent systems.

Markov games are a general mathematical tool for modeling interactions between multiple agents. The model is very general, for example common board games are special instances of Markov games, and particularly interesting because it forms an intersection of two distinct research disciplines: machine learning and game theory. Markov games extend Markov decision processes, a well-known tool for modeling single-agent problems, to multiagent domains. On the other hand, Markov games can be seen as a dynamic extension to strategic form games, which are standard models in traditional game theory. From the computer science perspective, Markov games provide a flexible and efficient way to describe different social interactions between intelligent agents.

This thesis studies different aspects of learning in Markov games. From the machine learning perspective, the focus is on a very general learning model, i.e. reinforcement learning, in which the goal is to maximize the long-time performance of the learning agent. The thesis introduces an asymmetric learning model that is computationally efficient in multiagent systems and enables the construction of different agent hierarchies. In multiagent reinforcement learning systems based on Markov games, the space and computational requirements grow very quickly with the number of learning agents and the size of the problem instance. Therefore, it is necessary to use function approximators, such as neural networks, to model agents in many real-world applications. In this thesis, various numeric learning methods are proposed for multiagent learning problems.

The proposed methods are tested with small but non-trivial example problems from different research areas including artificial robot navigation, simplified soccer game, and automated pricing models for intelligent agents. The thesis also contains an extensive literature survey on multiagent reinforcement learning and various methods based on Markov games.

Extensions of independent component analysis for natural image data

Mika Inki

Dissertation for the degree of Doctor of Science in Technology on 10 December 2004.

External examiners:
Michael Lewicki (Carnegie Mellon University)
Heikki Hyötyniemi (Helsinki University of Technology)
Opponent:
Gustavo Deco (Universitat Pompeu Fabra)

Abstract:

An understanding of the statistical properties of natural images is useful for any kind of processing to be performed on them. Natural image statistics are, however, in many ways as complex as the world which they depict. Fortunately, the dominant low-level statistics of images are sufficient for many different image processing goals. A lot of research has been devoted to second order statistics of natural images over the years.

Independent component analysis is a statistical tool for analyzing higher than second order statistics of data sets. It attempts to describe the observed data as a linear combination of independent, latent sources. Despite its simplicity, it has provided valuable insights of many types of natural data. With natural image data, it gives a sparse basis useful for efficient description of the data. Connections between this description and early mammalian visual processing have been noticed.

The main focus of this work is to extend the known results of applying independent component analysis on natural images. We explore different imaging techniques, develop algorithms for overcomplete cases, and study the dependencies between the components by using a model that finds a topographic ordering for the components as well as by conditioning the statistics of a component on the activity of another. An overview is provided of the associated problem field, and it is discussed how these relatively small results may eventually be a part of a more complete solution to the problem of vision.

Advances in variational Bayesian nonlinear blind source separation

Antti Honkela

Dissertation for the degree of Doctor of Science in Technology on 13 May 2005.

External examiners:
Fabian Theis (University of Regensburg)
Aki Vehtari (Helsinki University of Technology)
Opponent:
Tom Heskes (Radboud University Nijmegen)

Abstract:

Linear data analysis methods such as factor analysis (FA), independent component analysis (ICA) and blind source separation (BSS) as well as state-space models such as the Kalman filter model are used in a wide range of applications. In many of these, linearity is just a convenient approximation while the underlying effect is nonlinear. It would therefore be more appropriate to use nonlinear methods.

In this work, nonlinear generalisations of FA and ICA/BSS are presented. The methods are based on a generative model, with a multilayer perceptron (MLP) network to model the nonlinearity from the latent variables to the observations. The model is estimated using variational Bayesian learning. The variational Bayesian method is well-suited for the nonlinear data analysis problems. The approach is also theoretically interesting, as essentially the same method is used in several different fields and can be derived from several different starting points, including statistical physics, information theory, Bayesian statistics, and information geometry. These complementary views can provide benefits for interpretation of the operation of the learning method and its results.

Much of the work presented in this thesis consists of improvements that make the nonlinear factor analysis and blind source separation methods faster and more stable, while being applicable to other learning problems as well. The improvements include methods to accelerate convergence of alternating optimisation algorithms such as the EM algorithm and an improved approximation of the moments of a nonlinear transform of a multivariate probability distribution. These improvements can be easily applied to other models besides FA and ICA/BSS, such as nonlinear state-space models. A specialised version of the nonlinear factor analysis method for post-nonlinear mixtures is presented as well.

Exploratory cluster analysis of genomic high-throughput data sets and their dependencies

Janne Nikkilä

Dissertation for the degree of Doctor of Science in Technology on 1 December 2005.

External examiners:
Olli Yli-Harja (Tampere University of Technology)
Matej Oresic (Technical Research Centre of Finland)
Opponent:
Alvis Brazma (European Bioinformatics Institute)

Abstract:

This thesis studies exploratory cluster analysis of genomic high-throughput data sets and their interdependencies. In modern biology, new high-throughput measurements generate numerical data simultaneously from thousands of molecules in the cell. This enables a new perspective to biology, which is called systems biology. The discipline developing methods for the analysis of the systems biology data is called bioinformatics. The work in this thesis contributes mainly to bioinformatics, but the approaches presented are general purpose machine learning methods and can be applied in many problem areas.

A main problem in analyzing genomic high-throughput data is that the potentially useful new findings are hidden in a huge data mass. They need to be extracted and visualized to the analyst as overviews.

This thesis introduces new exploratory cluster analysis methods for extracting and visualizing findings of high-throughput data. Three kinds of methods are presented to solve progressively better-focused problems. First, visualizations and clusterings using the self-organizing map are applied to genomic data sets. Second, the recently developed methods for improving the visualization and clustering of a data set with auxiliary data are applied. Third, new methods for exploring the dependency between data sets are developed and applied. The new methods are based on maximizing the Bayes factor between the model of independence and the model of dependence for finite data.

The methods outperform their alternatives in numerical comparisons. In applications they proved capable of confirming known biological findings, which validates the methods, and also generated new hypotheses. The applications included exploration of yeast gene expression data, yeast gene expression data in a new metric learned with auxiliary data, the regulation of yeast gene expression by transcription factors, and the dependencies between human and mouse gene expression.