Doctoral dissertations

Data exploration process based on the self-organizing map

Juha Vesanto

Dissertation for the degree of Doctor of Science in Technology on 16 May 2002.

External examiners:
Jari Kangas (Nokia Research Center)
Jouko Lampinen (Helsinki University of Technology)
Opponent:
Alfred Ultsch (Philipps-University of Marburg, Germany)

Abstract:

With the advances in computer technology, the amount of data that is obtained from various sources and stored in electronic media is growing at exponential rates. Data mining is a research area which answers to the challange of analysing this data in order to find useful information contained therein. The Self-Organizing Map (SOM) is one of the methods used in data mining. It quantizes the training data into a representative set of prototype vectors and maps them on a low-dimensional grid. The SOM is a prominent tool in the initial exploratory phase in data mining.

The thesis consists of an introduction and ten publications. In the publications, the validity of SOM-based data exploration methods has been investigated and various enhancements to them have been proposed. In the introduction, these methods are presented as parts of the data mining process, and they are compared with other data exploration methods with similar aims.

The work makes two primary contributions. Firstly, it has been shown that the SOM provides a versatile platform on top of which various data exploration methods can be efficiently constructed. New methods and measures for visualization of data, clustering, cluster characterization, and quantization have been proposed. The SOM algorithm and the proposed methods and measures have been implemented as a set of Matlab routines in the SOM Toolbox software library.

Secondly, a framework for SOM-based data exploration of table-format data - both single tables and hierarchically organized tables - has been constructed. The framework divides exploratory data analysis into several sub-tasks, most notably the analysis of samples and the analysis of variables. The analysis methods are applied autonomously and their results are provided in a report describing the most important properties of the data manifold. In such a framework, the attention of the data miner can be directed more towards the actual data exploration task, rather than on the application of the analysis methods. Because of the highly iterative nature of the data exploration, the automation of routine analysis tasks can reduce the time needed by the data exploration process considerably.

Probabilistic models of early vision

Patrik Hoyer

Dissertation for the degree of Doctor of Science in Technology on 15 November 2002.

External examiners:
Mikko Lehtokangas (Tampere University of Technology)
Pentti Laurinen (University of Helsinki)
Opponent:
Eero Simoncelli (University of New York)

Abstract:

How do our brains transform patterns of light striking the retina into useful knowledge about objects and events of the external world? Thanks to intense research into the mechanisms of vision, much is now known about this process. However, we do not yet have anything close to a complete picture, and many questions remain unanswered. In addition to its clinical relevance and purely academic significance, research on vision is important because a thorough understanding of biological vision would probably help solve many major problems in computer vision.

A major framework for investigating the computational basis of vision is what might be called the probabilistic view of vision. This approach emphasizes the general importance of uncertainty and probabilities in perception and, in particular, suggests that perception is tightly linked to the statistical structure of the natural environment. This thesis investigates this link by building statistical models of natural images, and relating these to what is known of the information processing performed by the early stages of the primate visual system.

Recently, it was suggested that the response properties of simple cells in the primary visual cortex could be interpreted as the result of the cells performing an independent component analysis of the natural visual sensory input. This thesis provides some further support for that proposal, and, more importantly, extends the theory to also account for complex cell properties and the columnar organization of the primary visual cortex. Finally, the application of these methods to predicting neural response properties further along the visual pathway is considered.

Although the models considered account for only a relatively small part of known facts concerning early visual information processing, it is nonetheless a rather impressive amount considering the simplicity of the models. This is encouraging, and suggests that many of the intricacies of visual information processing might be understood using fairly simple probabilistic models of natural sensory input.

Unsupervised pattern recognition methods for exploratory analysis of industrial process data

Esa Alhoniemi

Dissertation for the degree of Doctor of Science in Technology on 13 December 2002.

External examiners:
Heikki Hyötyniemi (Helsinki University of Technology)
Jussi Parkkinen (University of Joensuu)
Opponents:
Hannu Koivisto (Tampere University of Technology)
Jussi Parkkinen (University of Joensuu)

Abstract:

The rapid growth of data storage capacities of process automation systems provides new possibilities to analyze behavior of industrial processes. As existence of large volumes of measurement data is a rather new issue in process industry, long tradition of using data analysis techniques in that field does not yet exist. In this thesis, unsupervised pattern recognition methods are shown to represent one potential and computationally efficient approach in analysis of such data.

This thesis consists of an introduction and six publications. The introduction contains a survey on process monitoring and data analysis methods, exposing the research which has been carried out in the fields so far. The introduction also points out the tasks in the process management framework where the methods considered in this thesis – self-organizing maps and cluster analysis – can be benefited.

The main contribution of this thesis consists of two parts. The first one is the use of the existing and development of novel SOM-based methods for process monitoring and data analysis purposes. The second contribution is a concept where cluster analysis is used to extract and identify operational states of a process from measured data. In both cases, the methods have been successfully applied in analysis of real data from processes in the wood processing industry.

Adaptive methods for on-line recognition of isolated handwritten characters

Vuokko Vuori

Dissertation for the degree of Doctor of Science in Technology on 14 December 2002.

External examiners:
Tapio Seppänen (University of Oulu)
Jukka Heikkonen (Helsinki University of Technology)
Opponents:
Tapio Seppänen (University of Oulu)
Louis Vuurpijl (Nijmegen University, Netherlands)

Abstract:

The main goal of the work presented in this thesis has been the development of an on-line handwriting recognition system which is able to recognize handwritten characters of several different writing styles and is able to improve its performance by adapting itself to new writing styles. The recognition method should be applicable to hand-held devices of limited memory and computational resources. The adaptation process should take place during normal use of the device, not in some specific training mode. For the usability aspect of the recognition system, the recognition and adaptation processes should be easily understandable to the users.

The first part of this thesis gives an introduction to the handwriting recognition. The topics considered include: the variations present in personal handwriting styles; automatic grouping of similar handwriting styles; the differences between writer-independent and writer-dependent as well as on-line and off-line handwriting recognition problems; the different approaches to on-line handwriting recognition; the previous adaptive recognition systems and the experiments performed with them; the recognition performance requirements and other usability issues related to on-line handwriting recognition; the current trends in on-line handwriting recognition research; the recognition results obtained with the most recent recognition systems; and the commercial applications.

The second part of the thesis describes an adaptive on-line character recognition system and the experiments performed with it. The recognition system is based on prototype matching. The comparisons between the character samples and prototypes are based on the Dynamical Time Warping (DTW) algorithm and the input characters are classified according to the k Nearest Neighbors (k-NN) rule. The initial prototype set is formed by clustering character samples collected from a large number of subjects. Thus, the recognition system can handle various writing styles. This thesis work introduces four DTW-based clustering algorithms which can be used for the prototype selection. The recognition system adapts to new writing styles by modifying its prototype set. This work introduces several adaptation strategies which add new writer-dependent prototypes into the initial writer-independent prototype set, reshape the existing prototypes with a Learning Vector Quantization (LVQ)-based algorithm, and inactivate poorly performing prototypes. The adaptations are carried out on-line in a supervised or self-supervised fashion. In the former case, the user explicitly labels the input characters which are used

as training samples in the adaptation process. In the latter case, the system deduces the labels from the recognition results and the user's actions. The latter approach is prone to erroneously labeled learning samples.

The different adaptation strategies were experimented with and compared with each other by performing off-line simulations and genuine on-line user experiments. In the simulations, special attention has been paid to the various erroneous learning situations likely to be encountered in real world handwriting recognition tasks. The recognition system is able to improve its recognition accuracy significantly on the basis of only a few additional character samples per class. Recognition accuracies acceptable in real world applications can be attained for most of the test subjects.

This work also introduces a Self-Organizing Map (SOM)-based method for analyzing personal writing styles. Personal writing styles are represented by high-dimensional vectors, the components of which indicate the subjects' tendencies to use certain prototypical writing styles for isolated characters. These writing style vectors are then visualized by a SOM which enables the detection and analysis of clusters of similar writing styles.

Using visualization, variable selection and feature extraction to learn from industrial data

Sampsa Laine

Dissertation for the degree of Doctor of Science in Technology on 19 September 2003.

External examiners:
Olli Saarela (Keskuslaboratorio Oy)
Petri Vasara (Jaakko Pöyry Group Oyj)
Opponent:
John Klus (University of Wisconsin-Madison)

Abstract:

Although the engineers of industry have access to process data, they seldom use advanced statistical tools. Why this reluctance? I believe engineers do not have adequate statistical skills, and that inexpert use of statistics leads to useless results. For example, failure to correctly identify and remove outliers disturbs those tools that assume Gaussian distribution of data. Also, failure to correctly parameterize the used algorithm leads to poor results, as an example, a process engineer may find it difficult to find the best structure of an artificial neural network. Failures of statistical tools lead the engineer to disregard statistics, and resort to visual study of manually selected data.

This thesis looks for algorithms that serve the common process engineer. I prefer three properties in an algorithm: supervised operation, robustness and understandability. Supervised operation allows and requires the user to explicate the goal of the analysis, which allows the algorithm to discover results that are relevant to the user. Robust algorithms allow engineers to analyse raw process data collected from the automation system of the plant. Understandability is the most important criterion: the user must understand how to parameterize the model, what is the principle of the algorithm, and know how to interpret the results.

These criteria are used to assess algorithms for visualization, variable selection and feature extraction. The objective of this thesis was to create a tool set the reliably and understandably provides the user with information that is related to a problem that he/she has defined interesting.

The tools and the criteria are illustrated by analysing an industrial case: the concentrator of the Hitura mine. This case illustrates how to define the problem using off-line laboratory data; and how to study the on-line data to find solutions. Statistical tools demonstratedly improve the efficiency of process study: my early results required approximately six man months of work; the algorithms proposed by this thesis produced comparable results in few weeks.

Interactive image retrieval using self-organizing maps

Markus Koskela

Dissertation for the degree of Doctor of Science in Technology on 14 November 2003.

External examiners:
Irwin King (Chinese University of Hong Kong)
Timo Ojala (University of Oulu)
Opponent:
Moncef Gabbouj (Tampere University of Technology)

Abstract:

Digital image libraries are becoming more common and widely used as visual information is produced at a rapidly growing rate. Creating and storing digital images is nowadays easy and getting more affordable all the time as the needed technologies are maturing and becoming eligible for general use. As a result, the amount of data in visual form is increasing and there is a strong need for effective ways to manage and process it. In many settings, the existing and widely adopted methods for text-based indexing and information retrieval are inadequate for these new purposes.

Content-based image retrieval addresses the problem of finding images relevant to the users' information needs from image databases, based principally on low-level visual features for which automatic extraction methods are available. Due to the inherently weak connection between the high-level semantic concepts that humans naturally associate with images and the low-level visual features that the computer is relying upon, the task of developing this kind of systems is very challenging. A popular method to improve retrieval performance is to shift from single-round queries to navigational queries where a single retrieval instance consists of multiple rounds of user-system interaction and query reformulation. This kind of operation is commonly referred to as relevance feedback and can be considered as supervised learning to adjust the subsequent retrieval process by using information gathered from the user's feedback.

In this thesis, an image retrieval system named PicSOM is presented, including detailed descriptions of using multiple parallel Self-Organizing Maps (SOMs) for image indexing and a novel relevance feedback technique. The proposed relevance feedback technique is based on spreading the user responses to local SOM neighborhoods by a convolution with a kernel function. A broad set of evaluations with different image features, retrieval tasks, and parameter settings demonstrating the validity of the retrieval method is described. In particular, the results establish that relevance feedback with the proposed method is able to adapt to different retrieval tasks and scenarios.

Furthermore, a method for using the relevance assessments of previous retrieval sessions or potentially available keyword annotations as sources of semantic information is presented. With performed experiments, it is confirmed that the efficiency of semantic image retrieval can be substantially increased by using these features in parallel with the standard low-level visual features.

Learning metrics and Discriminative Clustering

Janne Sinkkonen

Dissertation for the degree of Doctor of Philosophy on 21 November 2003.

External examiners:
Petri Myllymäki (University of Helsinki)
Kari Torkkola (Motorola)
Opponent:
Naftali Tisby (Hebrew University of Jerusalem)

Abstract:

In this work methods have been developed to extract relevant information from large, multivariate data sets in a flexible, nonlinear way. The techniques are applicable especially at the initial, explorative phase of data analysis, in cases where an explicit indicator of relevance is available as part of the data set.

The unsupervised learning methods, popular in data exploration, often rely on a distance measure defined for data items. Selection of the distance measure, part of which is feature selection, is therefore fundamentally important.

The learning metrics principle is introduced to complement manual feature selection by enabling automatic modification of a distance measure on the basis of available relevance information. Two applications of the principle are developed. The first emphasizes relevant aspects of the data by directly modifying distances between data items, and is usable, for example, in information visualization with the self-organizing maps. The other method, discriminative clustering, finds clusters that are internally homogeneous with respect to the interesting variation of the data. The techniques have been applied to text document analysis, gene expression clustering, and charting the bankruptcy sensitivity of companies.

In the first, more straightforward approach, a new local metric of the data space measures changes in the conditional distribution of the relevance-indicating data by the Fisher information matrix, a local approximation of the Kullback-Leibler distance. Discriminative clustering, on the other hand, directly minimizes a Kullback-Leibler based distortion measure within the clusters, or equivalently maximizes the mutual information between the clusters and the relevance indicator. A finite-data algorithm for discriminative clustering is also presented. It maximizes a partially marginalized posterior probability of the model and is asymptotically equivalent to maximizing mutual information.

Computational models relating properties of visual neurons to natural stimulus statistics

Jarmo Hurri

Dissertation for the degree of Doctor of Science in Technology on 5 December 2003.

External examiners:
Pentti Laurinen (University of Helsinki)
Jukka Heikkonen (Helsinki University of Technology)
Opponent:
Laurenz Wiskott (Humboldt-Universität zu Berlin)

Abstract:

The topic of this thesis is mathematical modeling of computations taking place in the visual system, the largest sensory system in the primate brain. While a great deal is known about how certain visual neurons respond to stimuli, a very profound question is *why* they respond as they do. Here this question is approached by formulating models of computation which might underlie the observed response properties. The main motivation is to improve our understanding of how the brain functions. A better understanding of the computational underpinnings of the visual system may also yield advances in medical technology or computer vision, such as development of visual prostheses, or design of computer vision algorithms.

In this thesis several models of computation are examined. An underlying assumption in this work is that the statistical properties of visual stimuli are related to the structure of the visual system. The relationship has formed through the mechanisms of evolution and development. A model of computation specifies this relationship between the visual system and stimulus statistics. Such a model also contains free parameters which correspond to properties of visual neurons. The experimental evaluation of a model consists of estimation of these parameters from a large amount of natural visual data, and comparison of the resulting parameter values against neurophysiological knowledge of the properties of the neurons, or results obtained with other models.

The main contribution of this thesis is the introduction of new models of computation in the primary visual cortex. The results obtained with these models suggest that one defining feature of the computations performed by a class of neurons called simple cells, is that the output of a neuron consists of periods of intense neuronal activity. It also seems that the activity levels of nearby simple cells are positively correlated over short time intervals. In addition, the probability of the occurrence of such regions of intense activity in the joint space of time and cortical area seems to be small. Another contribution of the thesis is the examination of the relationship between two previous computational models, namely independent component analysis and local spatial frequency analysis. This examination suggests that results obtained with independent component analysis share some important properties with wavelets, in the way their localization in space and frequency depends on their average spatial frequency.

Advances in Independent Component Analysis with applications to data mining

Ella Bingham

Dissertation for the degree of Doctor of Science in Technology on 12 December 2003.

External examiners:
Thomas Hoffman (Brown University)
Helena Ahonen-Myka (University of Helsinki)
Opponent:
Mark Plumbley (Queen Mary University of London)

Abstract:

This thesis considers the problem of finding latent structure in high dimensional data. It is assumed that the observed data are generated by unknown latent variables and their interactions. The task is to find these latent variables and the way they interact, given the observed data only. It is assumed that the latent variables do not depend on each other but act independently.

A popular method for solving the above problem is independent component analysis (ICA). It is a statistical method for expressing a set of multidimensional observations as a combination of unknown latent variables that are statistically independent of each other. Starting from ICA, several methods of estimating the latent structure in different problem settings are derived and presented in this thesis. An ICA algorithm for analyzing complex valued signals is given; a way of using ICA in the context of regression is discussed; and an ICA-type algorithm is used for analyzing the topics in dynamically changing text data. In addition to ICA-type methods, two algorithms are given for estimating the latent structure in binary valued data. Experimental results are given on all of the presented methods.

Another, partially overlapping problem considered in this thesis is dimensionality reduction. Empirical validation is given on a computationally simple method called random projection: it does not introduce severe distortions in the data. It is also proposed that random projection could be used as a preprocessing method prior to ICA, and experimental results are shown to support this claim.

This thesis also contains several literature surveys on various aspects of finding the latent structure in high dimensional data.