Doctoral dissertations

Bayesian ensemble learning for nonlinear factor analysis

Harri Valpola

Dissertation for the degree of Doctor of Science in Technology on 10 November 2000.

External examiners:
Mark Girolami (University of Paisley)
Petri Myllymäki (University of Helsinki)
Opponent:
Zoubin Ghahramani (University College London)

Abstract:
An active research topic in machine learning is the development of model structures which would be rich enough to represent relevant aspects of the observations but would still allow efficient learning and inference.

Linear factor analysis and related methods such as principal component analysis and independent component analysis are widely used feature extraction and data analysis techniques. They are computationally efficient but are restricted to linear models. Many natural phenomena are nonlinear and therefore several attempts have been made to generalise the model by relaxing the linearity assumption. The suggested approaches have suffered from overfitting and the computational complexity of many of the algorithms scales exponentially with respect to the number of factors, which makes the application of these methods to high dimensional factor spaces infeasible.

This thesis describes the development of a nonlinear extension of factor analysis. The learning algorithm is based on Bayesian probability theory and solves many of the problems related to overfitting. The unknown nonlinear generative mapping is modelled by a multi-layer perceptron network. The computational complexity of the algorithm scales quadratically with respect to the dimension of the factor space which makes it possible to use a significantly larger number of factors than with the previous algorithms. The feasibility of the algorithm is demonstrated in experiments with artificial and natural data sets. Extensions which combine the nonlinear model with non-Gaussian and dynamic models for the factors are introduced.

Text mining with the WEBSOM

Krista Lagus

Dissertation for the degree of Doctor of Science in Technology on 11 December 2000.

External examiners:
Dieter Merkl (Technische Universität Wien)
Pasi Koikkalainen (University of Jyväskylä)
Opponent:
Risto Miikkulainen (University of Texas at Austin)

Abstract:
The emerging field of text mining applies methods from data mining and exploratory data analysis to analyzing text collections and to conveying information to the user in an intuitive manner. Visual, map-like displays provide a powerful and fast medium for portraying information about large collections of text. Relationships between text items and collections, such as similarity, clusters, gaps and outliers can be communicated naturally using spatial relationships, shading, and colors.

In the WEBSOM method the self-organizing map (SOM) algorithm is used to automatically organize very large and high-dimensional collections of text documents onto two-dimensional map displays. The map forms a document landscape where similar documents appear close to each other at points of the regular map grid. The landscape can be labeled with automatically identified descriptive words that convey properties of each area and also act as landmarks during exploration. With the help of an HTML-based interactive tool the ordered landscape can be used in browsing the document collection and in performing searches on the map.

An organized map offers an overview of an unknown document collection helping the user in familiarizing herself with the domain. Map displays that are already familiar can be used as visual frames of reference for conveying properties of unknown text items. Static, thematically arranged document landscapes provide meaningful backgrounds for dynamic visualizations of for example time-related properties of the data. Search results can be visualized in the context of related documents.

Experiments on document collections of various sizes, text types, and languages show that the WEBSOM method is scalable and generally applicable. Preliminary results in a text retrieval experiment indicate that even when the additional value provided by the visualization is disregarded the document maps perform at least comparably with more conventional retrieval methods.

Self-organizing maps for signal and symbol sequences

Panu Somervuo

Dissertation for the degree of Doctor of Science in Technology on 17 December 2000.

External examiners:
Olli Ventä (Technical Research Centre of Finland)
Jukka Heikkonen (Helsinki University of Technology)
Opponent:
Kari Torkkola (Motorola, USA)

Abstract:
Temporal sequences arise from various kinds of sources in the nature. Sensory elements transform the events into measurements and corresponding feature vectors. Methods for comparing sequences are needed in retrieval, error correction, and recognition tasks. Due to the durational differences in the feature sequences and the variation and noise in the feature vectors, both temporal and spatial fluctuations must be tolerated in the comparison.

The topic of this thesis is the application of the Self-Organizing Map (SOM) and Learning Vector Quantization (LVQ) algorithms to the processing of data sequences with numeric or symbolic elements. Usually the models of the data are fixed-dimensional feature vectors in the SOM and LVQ. In this work also other models have been experimented such as hidden Markov models, feature vector sequences, and symbol sequences. The goal has been to utilize the principles of the unsupervised learning of the SOM and the error corrective learning of the LVQ while also taking the sequential nature of the input data into account.

The developed methods have been applied to the unsupervised segmentation of speech, forming of templates of variable-length feature vector sequences, improvement of the performance of a speech recognizer based on hidden Markov models, construction of a data-driven pronunciation dictionary for speech recognition, and clustering of all currently known protein sequences.

User profiling and classification for fraud detection in mobile communications networks

Jaakko Hollmén

Dissertation for the degree of Doctor of Science in Technology on 19 December 2000.

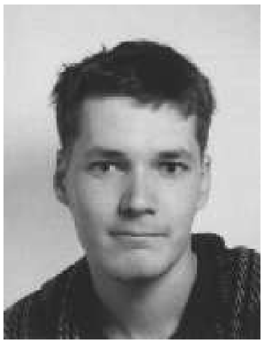
External examiners:
Jyrki Joutsensalo (University of Jyväskylä)
Henry Tirri (University of Helsinki)
Opponent:
Klaus Obermayer (Technische Universität Berlin)

Abstract:
The topic of this thesis is fraud detection in mobile communications networks by means of user profiling and classification techniques. The goal is to first identify relevant user groups based on call data and then to assign a user to a relevant group. Fraud may be defined as a dishonest or illegal use of services, with the intention to avoid service charges. Fraud detection is an important application, since network operators lose a relevant portion of their revenue to fraud. Whereas the intentions of the mobile phone users cannot be observed, it is assumed that the intentions are reflected in the call data. The call data is subsequently used in describing behavioral patterns of users. Neural networks and probabilistic models are employed in learning these usage patterns from call data. These models are used either to detect abrupt changes in established usage patterns or to recognize typical usage patterns of fraud. The methods are shown to be effective in detecting fraudulent behavior by empirically testing the methods with data from real mobile communications networks.

Attribute trees as adaptive object models in image analysis

Markus Peura

Dissertation for the degree of Doctor of Science in Technology on 23 February 2001.



External examiners:
Pekka Kilpeläinen (University of Kuopio)
Heikki Kälviäinen (Lappeenranta University of Technology)
Opponent:
Pekka Orponen (University of Jyväskylä)

Abstract:
This thesis focuses on the analysis of irregular hierarchical visual objects. The main approach involves modelling the objects as unordered attribute trees. A tree presents the overall organization, or topology, of an object, while the vertex attributes describe further visual properties such as intensity, color, or size. Techniques for extracting, matching, comparing, and interpolating attribute trees are presented, principally aiming at systems that can learn to recognize objects. Analysis of weather radar images has been the pilot application for this study, but the main ideas are applicable in structural pattern recognition more generally.

The central original contribution of this thesis is the Self-Organizing Map of Attribute Trees (SOM-AT) which demonstrates how the proposed tree processing techniques - tree indexing, matching, distance functions, and mixtures - can be embedded in a learning system; the SOM-AT is a variant of the Self-Organizing Map (SOM), the neural network model invented by Prof. Teuvo Kohonen. The SOM is especially suited to visualizing distributions of objects, classifying objects and monitoring changes in objects. Hence, the SOM-AT can be applied in the respective tasks involving hierarchical objects. More generally, the proposed ideas are applicable in learning systems involving comparisons and interpolations of trees.

The suggested heuristic index-based tree matching scheme is another independent contribution. The basic idea of the heuristic is to divide trees to subtrees and match the subtrees recursively by means of topological indices. Given two attribute trees, the larger of which has N vertices, and the maximal child count (out-degree) is D vertices, the complexity of the scheme is only $O(N \log D)$ operations while exact matching schemes seem to have at least quadratic complexity: $O(N^{2.5})$ operations in checking isomorphisms and $O(N^3)$ operations in matching attribute trees. The proposed scheme is efficient also in terms of its "hit rate", the number of successfully matched vertices. In matching two random trees of $N \leq 10$ vertices, the number of heuristically matched vertices is on average 99 % of the optimum, and the accuracy for trees of $N \leq 500$ vertices is still above 90 %.

The feasibility of the proposed techniques is demonstrated by database querying, monitoring, and clustering of weather radar images. A related tracking scheme is outlined as well. In addition to weather radar images, a case study is presented on Northern light (*Aurora Borealis*) images. Due to the generic approach, there are also medical and geographical applications in view.

Auroral monitoring network: from all-sky camera system to automated image analysis

Mikko Syrjäsuo

Dissertation for the degree of Doctor of Science in Technology on 28 November 2001.

External examiners:
Ari Visa (Tampere University of Technology)
Jøran Moen (University of Oslo)
Opponents:
Jøran Moen (University of Oslo)
Jussi Parkkinen (University of Joensuu)

Abstract:
Auroras occur in the auroral zones encircling the magnetic poles of the Earth. Precipitating particles collide with atmospheric atoms and molecules in the ionosphere and produce light. In the northern hemisphere, we call this light aurora borealis or the Northern Lights. The solar wind, the magnetosphere and the ionosphere are electromagnetically coupled, which means that auroral activity is a direct consequence of the dynamic plasma processes that take place in the near-Earth space. Thus, the aurora provides us a way to monitor the Earth s plasma environment.

This thesis focuses on developing auroral imaging technology and data processing. The contribution to auroral instrumentation is a unique state-of-the-art network of all-sky camera stations that acquire images of the aurora in Fennoscandia and Svalbard. All stations operate autonomously and can also be remote controlled. The computer-controlled imagers utilise fish-eye lenses and image intensifiers to capture the whole sky into one digital image nominally at 20-s intervals during dark time. Also, calibration efforts have been initiated, which results in higher quality data than earlier.

The auroral data has traditionally been analysed manually, which is not only tedious but also prone to errors and is not easily repeatable. This thesis contributes the very first steps in combining contemporary image processing and machine vision techniques in analysing auroral images automatically. Techniques based on sample auroral images, shape skeletons and attribute trees are developed to classify the inherently noisy and ambiguous images. Time-series of images are analysed by utilising an auroral arc tracker.

The feasibility of the proposed techniques is demonstrated by comparing their performance to that of a human expert. The amount of required manual labour was vastly reduced by using the machine vision methods in searching for auroral arcs for a statistical study. Also, the auroral occurrence for one year was determined based on classifying 180000 images completely automatically.