ALGODAN

Algorithmic Data Analysis

POSTER SESSION March 18th 2012

Posters

1)

Title: **Mixture models from multiresolution 0-1 data**
Authors: Prem Raj Adhikari, Jaakko Hollmén
Description: This contribution proposes a multiresolution mixture model consisting of multiresolution mixture components whose structure are determined from the domain ontology. The individual mixture components provide functionality of Bayesian networks.

2)

Title: **Predicting the hardness of learning Bayesian networks**
Authors: Matti Järvisalo, Kustaa Kangas, Mikko Koivisto, Brandon Malone, Petri Myllymäki
Description: There are various algorithms for finding a Bayesian network structure that is optimal with respect to a given scoring function. It is a priori not clear which algorithm performs best on a given problem instance, given that no single algorithm dominates the others in speed and the running times are complicated functions of problem instances. We use machine learning techniques to train running time predictors based on a number of efficiently computable features of problem instances. Our results show that running times can be predicted with a reasonable accuracy and even a simple predictor will almost always choose the fastest algorithm.

3)

Title: **Lempel-Ziv Parsing in External Memory**
Authors: Juha Kärkkäinen, Dominik Kempa, Simon J. Puglisi
Description: For over three decades, the Lempel-Ziv factorization (or LZ77 parsing) has been a fundamental tool for data compression. More recently it has become the basis for several compressed text indexes that are particularly effective for massive, highly repetitive data sets. When computing the factorization for such large data sets, the space requirement of algorithms can become a problem. We escape the limitations of RAM by describing the first external memory LZ77 parsing algorithms. We provide an experimental comparison of several different approaches.

4)

Title: **Discovering significant patterns from sequential data**
Authors: Nikolaj Tatti
Description: Pattern discovery is a well-studied topic in data mining. One of the biggest setback in traditional pattern mining is that the obtained patterns are heavily redundant. Consequently, in order to improve the quality of the results, patterns are ranked or filtered based on some statistical model. Unlike with other pattern types, ranking such patterns requires sophisticated combinatorial arguments even for the very simplest models. In practice, ranking can be done efficiently and provides a great tool for improving the quality of patterns.

5)

Title: **Size matters: Finding the most informative set of window lengths**
Authors: Jefrey Lijffijt, Panagiotis Papapetrou, and Kai Puolamäki
Description: Event sequences often contain continuous variability at different levels of granularity and, when looking for local patterns in sequential data, it is often difficult to choose the right granularity (window length) for analysis. We study the problem of finding the best set of window lengths for analyzing discrete event sequences using algorithms with real-valued output. We propose to select the most informative set of window lengths by mapping the problem to a regression problem, introduce a corresponding optimisation problem, and show that the problem can be solved efficiently. We demonstrate the method on synthetic data and data from two domains: text and DNA sequences.

6)

Title: **Regression models for data streams with missing values**
Authors: Indrė Žliobaitė, Jaakko Hollmén
Description: We theoretically analyze effects of missing values to the accuracy of linear predictive models operating on streaming data. We derive the optimal least squares solution that minimizes the expected mean squared error given an expected rate of missing values. Based on this theoretically optimal solution we propose a recursive algorithm for producing and updating linear regression online, without accessing historical data.

7)

Title: **Mining the near infrared sky: star formation and embedded clusters**
Authors: Otto Solin, Lauri Haikala, Esko Ukkonen
Description: The aim of this research is to locate previously unknown stellar clusters from two near infrared surveys: the UKIDSS Galactic Plane Survey and the VISTA variables in the Vía Láctea survey. The cluster candidates were computationally searched from pre-filtered catalogue data using a recently proposed method that fits a mixture model of Gaussian densities and background noise using the expectation maximization algorithm. The pre-filtering of the data involves both removing data artefacts and searching for sources classified as non-stellar due to associated surface brightness thus directing the search to particularly embedded stellar clusters. The findings were further screened by visual inspection of images, and SIMBAD was used to study sources in the direction of the candidates. Our search resulted in 294 new cluster candidates.

8)

Title: **Algorithms for genome assembly**
Authors: Leena Salmela, Veli Mäkinen, Niko Välimäki, Esko Ukkonen
Description: The goal of de novo genome assembly is to reconstuct the genomic sequence of a previously unsequenced organism. We have been involved with the de novo assembly project of the Granville fritillary butterfly (Melitaea cinxia). In this project we have developed several new methods and tools for the various stages of genome assembly.

9)

Title: **Predicting quantitative binding interactions between drug compounds and protein kinases**

Authors: A. Cichonska, J. Tang, T. Aittokallio, J. Rousu

Description: Various computational methods have been developed to facilitate the process of determining interactions between drug compounds and their molecular targets. Among them, similarity-based machine learning algorithms are considered as state-of-the-art approaches. The assumption is that similar compounds are likely to interact with similar targets. Similarities between drugs are typically being computed based on their chemical structures and similarities between targets are being obtained by amino acid sequence alignments. The presented work was concentrated on predicting missing drug-target interaction affinities in the large-scale studies of selectivity profiles for kinase inhibitors, using Kronecker RLS machine learning algorithm The focus was on quantitative measurements instead of binary on-off relationships.

10)

Title: **Supervised and unsupervised biological network inference from multiple genomic data**

Authors: Jana Kludas, Fitsum Tamene, Juho Rousu

Description: Our goal is to investigate and ultimately improve biological network reconstruction based on integrating proteomic and genomic data. First, this work investigates the descriptive power of different features extracted from the protein sequences and genes for predicting biological networks of protein-protein interactions and metabolic networks. Three fundamentally different methods for graph inference have been implemented: (I) classification and integration of local interaction models, (II) classification of a global interaction model and (III) an unsupervised method based on estimating the inverse covariance matrix. The results show that for PPI and metabolic network prediction different data sources and different learning strategies are effective. For predicting the metabolic network the integration of proteomic and genomic data sources performs best. On the other hand, protein-protein interactions are best predicted by classification of sequence alignment scores.

11)

Title: **Molecular fingerprint prediction with multiple kernel learning**
Authors: Huibin Shen, Kai Duhrkop, Sebastian Bocker, Juho Rousu
Description: We combine fragmentation tree computations with kernel-based machine learning to predict molecular fingerprints and identify molecular structures. We introduce a family of kernels capturing the similarity of fragmentation trees, and combine these kernels using recently proposed multiple kernel learning approaches. Experiments on two large reference datasets show that the new methods significantly improve molecular fingerprint prediction accuracy. These improvements result in better metabolite identification.

12)

Title: **Modeling and predicting regulatory areas**
Authors: Jarkko Toivonen and Esko Ukkonen
Description: We have created models describing the binding of transciption factors to DNA. Combination of these simpler models can be used to model and predict the regulatory areas, the areas that affect the expression of genes.

13)

Title: **Geospatial Data Analysis and Processing**
Authors: Mikko Nikkilä, Valentin Polishchuk, Topi Talvitie
Description: We developed a new shape reconstruction tool and applied it in two domains: to traces of seismic events in Fennoscandia (results to appear in Geophysical Journal International, 2014) and to Helsinki public transport data (results published in IEEE Transactions on Visualization and Computer Graphics, 2013). In another research direction, we gave algorithms to coordinate flight of searchers over a terrain during rescue operations (results presented at ACM SIGSPATIAL GIS 2013). Further on the flight planning frontier, we designed an "Air Traffic Controller Game" and invite the participants to try themselves as controllers.

14)

Title: **Multilabel Classification through Random Graph Ensemble**
Authors: Hongyu Su, Juho Rousu
Description: We present new methods for multilabel classification, relying on ensemble learning on a collection of random output graphs imposed on the multilabel and a kernel-based structured output learner as the base learner. Diversity of base classifiers arises from the different random output structures, a different approach from boosting or bagging.

15)

Title: **Treedy: A Heuristic for Counting and Sampling Subsets**
Authors: Teppo Niinimäki, Mikko Koivisto
Description: Consider a collection of weighted subsets of a ground set N. We present a tree-based greedy heuristic, Treedy, that for a given query subset Q of N and a tolerance d approximates the weighted sum over all subsets of Q within relative error d. It also enables approximate sampling of subset of Q proportionally to the weights within total variation distance d. Experimental results show that approximations yield dramatic savings in running time compared to exact computation, and that Treedy typically outperforms a previously proposed sorting-based heuristic.

16)

Title: **Analysis of Environmental Proxies and Dendrochronological Series**
Authors: Mikko Korpela and Jaakko Hollmén
Description: We have studied temperature reconstruction using environmental proxy variables. Dendrochronological series such as tree-ring widths can also be used for this purpose. Our latest contribution to dendrochronology software package dplR is the adaptation and improvement of an uneven sampling spectral analysis method, facilitated by a public domain implementation available from its original authors.

17)

Title: **Expandable String Representation for Music Features Using Limited-Size Alphabets**
Authors: Simo Linkola, Lari Rasku, Teppo E. Ahonen
Description: We present an extandable string representation for symbolic polyphonic music. The high-level music features are turned into string representations with efficiently limited alphabet sizes. This allows using any common similarity metric for measuring similarity between pieces of music, and also makes adding new features into the representation easy. The representation is evaluated with a retrieval experiment using a dataset of classical music variations.

18)

Title: Poetry Engine: Designing an Interactive Tool for Poetry Co-Creation at School
Authors: Anna Kantosalo, Jukka M. Toivanen, Hannu Toivonen
Description: The poster presents our ongoing work on creating a poetry co-creation tool targeted at children attending the Finnish comprehensive school. The tool is a serious real life application for the computational poetry composition methods developed at the University. We are using a user-centered development approach to designing the tool in order to ensure it's suitability for the target audience.

19)

Title: **Efficient and simple algorithms for time-scaled and time-warped music search**
Authors: Antti Laaksonen
Description: The poster presents new algorithms for time-scaled and time-warped search in a symbolic music database. The algorithms are efficient both in theory and practice, and they are also easy to implement.

20)

Title: **Text summarization based on word associations**
Authors: Oskar Gross, Antoine Doucet, Hannu Toivonen
Description: This paper proposes the association mixture method, a novel approach for (multi-) document summarization. In an unsupervised and language-independent fashion, this technique relies on the strength of word associations in the set of documents to be summarized. The summaries are generated by picking sentences which cover the most specific word associations of the document(s). Representative word associations are detected by contrasting their statistical strength in the document(s) against their strength in a stereotypical background corpus. The performance of the method is measured on the DUC 2007 dataset. With statistical significance, our experiments indicate that the association mixture model is the best-performing unsupervised summarization method in the state-of-the-art that makes no use of human-curated knowledge bases.

Software Demonstrations

21)

Title: **Recent improvements in PULS project**

Authors**:** Mian Du, Matthew Pierce, Lidia Pivovarova, Roman Yangarber

Description: PULS is a web-scale monitoring system that finds events (such as outbreaks of infectious disease, or company mergers, etc.) in on-line news streams. The PULS database contains millions of documents, currently in two languages (English and Russian) and multiple domains (business-related news, cross-border security, medical surveillance). PULS utilizes different methods of text analysis: information extraction and natural language processing, supervised and semi-supervised learning, cross-domain and cross-language adaptation. In the demo we will demonstrate -- through a web-based interface -- a variety of tools for searching, clustering and visualization of the discovered information, covering various topics in our ongoing research, including multi-class text classification and cross-language named entity recognition.