ALGODAN
Algorithmic Data Analysis

POSTER SESSION May 16th 2012

Combinatorial Pattern Matching

1)
Title: **Algorithms for genome assembly**
Authors: Leena Salmela, Veli Mäkinen, Niko Välimäki, Johannes Ylinen, Esko Ukkonen
Presenter: Leena Salmela
Description: Current DNA sequencing technologies can produce huge amounts of short reads. The de novo genome assembly problem is to infer the genome of an organism based on these reads. We have studied several problems related to the assembly problem including correcting sequencing errors in the reads, computing overlaps between the reads, and organizing longer contiguous sequences into gapped sequences called scaffolds based on paired reads.

2)
Title: **Identifying regulatory modules in genome**
Authors: Jarkko Toivonen, Arttu Jolma, Jussi Taipale, Esko Ukkonen, Pasi Rastas, Teemu Kivioja, Mikko Sillanpää
Presenter: Jarkko Toivonen
Description: We present a model for recognising transcription factor binding sites and describing the cooperation between factors.

3)
Title: **Mining the UKIDSS GPS: star formation and embedded clusters**
Authors: Otto Solin, Esko Ukkonen, Lauri Haikala
Presenter: Otto Solin
Description: The aim of this research is to locate previously unknown stellar clusters from the near-infrared UKIDSS Galactic Plane 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 167 new cluster candidates.

4)

Title: **Accelerating Burrows-Wheeler Compression with Grammar Precompression**

Authors: Dominik Kempa, Juha Kärkkäinen, Pekka Mikkola

Presenter: Pekka Mikkola

Description: Text compression algorithms based on the Burrows-Wheeler transform (BWT) typically achieve a good compression ratio but are slow compared to Lempel-Ziv type compression algorithms. The main culprit is the time needed to compute the BWT during compression and its inverse during decompression. We propose to speed up BWT-based compression by performing grammar-based compression before the transform. The idea is to reduce the amount of data that BWT and its inverse have to process. We have developed a very fast grammar compressor using pair replacement for the purpose. Experiments show a substantial speed up in practice without a signicant effect on compression ratio.

5)

Title: **Slashing the Time for BWT Inversion**

Authors: Juha Kärkkäinen, Dominik Kempa, Simon J. Puglisi

Presenter: Dominik Kempa

Description: The Burrows-Wheeler transform (BWT) is a powerful tool for data compression used for example in the popular compression program bzip2. We describe new algorithms for inverting the BWT, which is usually the bottleneck in the decompression phase due to a high number of CPU cache misses. One of the algorithms is consistently 2.3--4 times as fast as the state-of-the-art. Another algorithm achieves an asymptotic reduction in cache misses in theory and is the fastest algorithm in practice for highly repetitive data.

6)

Title: **Indexing Finite Language Representation of Population Genotypes**

Authors: Jouni Sirén, Niko Välimäki, Veli Mäkinen

Presenter: Jouni Sirén

Description: Compressed full-text indexes based on the Burrows-Wheeler transform are widely used in bioinformatics. Their most succesful application so far has been mapping short reads to a reference sequence. We generalize these indexes to handle finite automata, e.g. those representing the known genetic variation within a population.

7)

Title: **Distributed String Mining Algorithm for High-Throughput Sequencing Data**

Author: Niko Välimäki

Presenter: Niko Välimäki

Description: The goal of frequency constrained string mining is to extract substrings that discriminate two (or more) datasets. The existing algorithms are practical only up to a few gigabytes of input. We introduce a distributed algorithm and apply it to a large-scale metagenomics study.

8)

Title: **Geometric Data Summarization Simplified and Improved**

Authors: Dan Feldman, Juha-Antti Isojärvi, Valentin Polishchuk

Presenter: Juha-Antti Isojärvi

Description: Coresets are a powerful tool in approximating extent measures of high-dimensional data. We designed a new coreset; our construction is simple and better conforms to the data.

9)

Title: **Compression-Based Clustering of Chromagram Data: New Method and Representations**

Author: Teppo E. Ahonen

Presenter: Teppo E. Ahonen

Description: We study clustering recorded music using novel quantized pitch class profile representations and data compression based similarity measuring. Based on work to be presented at CMMR 2012.

10)

Title: **Distinguishing between major and minor chords in automatic chord transcription**

Author: Antti Laaksonen

Presenter: Antti Laaksonen

Description: Automatic chord transcription is a problem of extracting the harmonic content from a music signal and representing it through chord symbols. We focus on distinguishing between major and minor chords in automatic chord transcription. We are especially interested in the role of the musical context in this process. We conduct an experiment where human listeners are asked to classify chords which a computer transcriber has failed to recognize when evaluated using a collection of Beatles songs. Based on this experiment and our analysis, we conclude that the musical context is often needed in distinguishing between major and minor chords. Furthermore, sometimes the quality of a chord cannot be unambiguously determined even if the full musical context is available.

11)

Title: **Analysis of Etymological Data via MDL**

Authors: Hannes Wettig, Roman Yangarber

Presenter: Javad Nouri

Description: Etymological databases contain sets of genetically related words (strings of symbols) within a language family, i.e., words that derive from the same ancestor form through globally regular correspondences. We present several MDL-based models for aligning the etymological data (pair-wise and in higher dimensions) and discovering the rules of correspondence among the languages. We evaluate the models in terms of compression power, imputation of unseen data, finding linguistically meaningful rules of correspondence, and several methods for reconstruction of phylogenetic trees and networks based on the resulting alignments.

Data Mining: Theory and Applications

12)

Title: **Analysis of Linguistic Variation**
Authors: Jefrey Lijffijt et al.
Presenter: Jefrey Lijffijt
Description: In the past decades, many medium to large text corpora have been compiled and annotated. This enables the study of more diverse and detailed aspects of language, e.g., differences between writing style of various age groups, or in various media. Alongside these developments, new computational and statistical challenges arise.

13)

Title: **Ensemble Computation with OR- and SUM-circuits**
Authors: Matti Järvisalo, Petteri Kaski, Mikko Koivisto, Janne Korhonen
Presenter: Janne Korhonen
Description: Given a Boolean function as input, a fundamental problem is to find a Boolean circuit with the least number of elementary gates (AND, OR, NOT) that computes the function. The problem generalises naturally to the setting of multiple Boolean functions: find the smallest Boolean circuit that computes all the functions simultaneously. We study an NP-complete variant of this problem titled Ensemble Computation under two monotone circuit classes: OR-circuits and SUM-circuits. In particular, we are interested in understanding the separation between these classes. The main motivation for this work is the relationship between the problem of rewriting in subquadratic time a given OR-circuit to a SUM-circuit and the existence of non-trivial algorithms for NP-hard problems, e.g. CNF-SAT. We also present computational results on the sizes of small OR- and SUM-circuits.

14)

Title: **Partial order MCMC for structure discovery in Bayesian networks**
Authors: Teppo Niinimäki, Pekka Parviainen, Mikko Koivisto
Presenter: Teppo Niinimäki
Description: We present a new Markov chain Monte Carlo method for estimating posterior probabilities of structural features in Bayesian networks. The method samples partial orders on the nodes; for each sample, the conditional probabilities of interest are computed exactly. Compared to previous methods our algorithm obtains a significant reduction in the size of sample space with negligible increase in computation time.

15)

Title: **Ancestor relations in the presence of unobserved variables**
Authors: Pekka Parviainen, Mikko Koivisto
Presenter: Pekka Parviainen
Description: We present an exact dynamic programming algorithm for computing posterior probabilities ancestor relations, that is, directed paths in Bayesian networks. Our experimental results show that ancestor relations can be learned with good power even when a majority of involved variables are unobserved.

16)
Title: **Backward Model Selection in Finite Mixture Models**
Authors: Prem Raj Adhikari, Jaakko Hollmén
Presenter: Prem Raj Adhikari
Description: The poster presents a search-based backward model selection method for finite mixture models using progressive merging of mixture components. The poster also presents a data driven, fast approximation of Kullback-Leibler (KL) divergence as a criterion to merge the mixture components.

17)
Title: **Environmental proxy selection problems in temperature reconstruction**
Authors: Mikko Korpela, Jaakko Hollmén
Presenter: Mikko Korpela
Description: Direct temperature measurements are only available from the past few hundred years. Therefore, proxy measurements must be used. We study the use of different environmental proxy variables for temperature reconstruction. Differences in both the time coverage of the proxies and the temperature signal present in them pose a challenge to the recovery of reliable temperature records.

18)
Title: **Damage detection methods for Structural Health Monitoring with Wireless Sensor Networks**
Authors: Janne Toivola, Jaakko Hollmén
Presenter: Janne Toivola
Description: Detecting changes in the condition of large structures, like bridges, offers a challenging data analysis task, as there are no practical sensors that would directly indicate damages and the potential future damages may have unpredictable effect on the measurements. Thus, we need to extract indirect features, insensitive to environmental variability, and novelty detection methods to detect possible unforeseen changes. This work considers the following data processing chain: feature extraction from low-power wireless accelerometer sensors, centralized and distributed feature space dimensionality reduction methods, and the final damage detection in the novelty detection framework.

Machine Learning

19)

Title: **Density and entropy estimation with NML histogram**

Authors: Panu Luosto, Ciprian Giurcaneanu, Petri Kontkanen

Presenter: Panu Luosto

Description: Theoretical advances of the last decade have led to novel methodologies for density and entropy estimation by irregular histograms and penalized maximum likelihood. We compare empirically four histogram methods. They include the normalized maximum likelihood (NML) histogram by Kontkanen and Myllymäki, and its novel variant that is based on NML as well. As an extension to irregular histograms, we also test the new MDL based clustgram.

20)

Title: **Metabolite Identification and molecular fingerprint prediction via machine learning**

Authors: Markus Heinonen, Huibin Shen, Juho Rousu

Presenter: Markus Heinonen

Description: Identification of metabolites from tandem mass spectrometry measurements is a prerequisite step for metabolic modeling and network analysis. Currently this task requires matching of measured mass spectra against annotated databases of reference spectra, and extensive manual work. We propose a machine learning framework, which identifies the metabolite structures based on the mass spectral signals. We decompose the problem into binary subproblems each predicting an individual property of the structure. The complete structure is then inferred from the properties. We show promising results in identifying metabolites using several mass spectra datasets.

21)

Title: **Efficient Path Kernels for Reaction Function Prediction**

Authors: Markus Heinonen, Niko Välimäki, Veli Mäkinen, Juho Rousu

Presenter: Markus Heinonen

Description: We propose the first efficient path-based graph kernel for classification of reaction graphs. The path kernel utilizes efficient compressed path index data structure to store the potentially millions of paths. In our experiments we outperform state-of-the-art graph kernels in prediction of the EC code of organic reactions.

22)

Title: **Protein Interaction Network Prediction in Yeast based on Sequence Features**

Author: Jana Kludas

Presenter: Jana Kludas

Description: The over all goal of this work is to investigate and ultimately improve biological network reconstruction tools that are based on heterogeneous biological data. Preliminary experiments focus on Baker's yeast (Saccharomyces cerevisiae), but it is planned to experiment with other yeast and fungi in future work. A simple local modelling approach is implemented that trains SVM classifiers for a set of seed proteins in a network based on known protein interactions from the STRING data base and BLAST alignment scores, Global Trace Graph (GTG) and IPRscan features. Feature selection is applied to the input space to improve the classification results on the high dimensional and small scale data set.

23)

Title: **Random Graph Ensemble in Multi-Task Classification**

Authors: Hongyu Su, Juho Rousu

Presenter: Hongyu Su

Description: We present an ensemble of multi-task classifiers for multilabel classification. As the base classifiers of ensemble, we use Maximum Margin Conditional Random Field (MMCRF) Model. Source diversity of base classifiers arises from the different random output structures, a different approach from boosting or bagging. Experimental result shows that ensembles of random networks outperforms other approaches.

Neuroinformatics

24)

Title: **Bregman divergence as general framework to estimate unnormalized statistical models**

Authors: Michael Gutmann, Jun-ichiro Hirayama

Presenter: Michael Gutmann

Description: We show that the Bregman divergence provides a rich framework to estimate unnormalized statistical models for continuous or discrete random variables, that is, models which do not integrate or sum to one, respectively. We prove that recent estimation methods such as noise-contrastive estimation, ratio matching, and score matching belong to the proposed framework, and explain their interconnection based on supervised learning. Further, we discuss the role of boosting in unsupervised learning.

25)

Title: **Decoding Sensory Modalities from MEG Signals on the Basis of Spectrospatial Information**

Authors: Jukka-Pekka Kauppi, Lauri Parkkonen, Riitta Hari, Aapo Hyvärinen

Presenter: Jukka-Pekka Kauppi

Description: Multivariate pattern classification methods are currently widely applied to investigate how sensory, motor and cognitive information is represented in brain imaging signals. The goal is often to decode distinct brain states related to processing of different stimuli. Here, we used magnetoencephalography (MEG) that has millisecond-range temporal resolution and hypothesized that spectrospatial information present in MEG is useful in separating different stimulus categories from each other. To this aim, we designed four classifiers and evaluated their performance in a four-category decoding task.

26)

Title: **Learning Topographic Representations for Linearly Correlated Components**

Authors: Hiroaki Sasaki, Michael U. Gutmann, Hayaru Shouno, Aapo Hyvärinen

Presenter: Hiroaki Sasaki

Description: The poster describes a new extension of ICAstatistical model for the estimation of correlated topographic representations. Applications to real data such as natural images are also shown.

27)

Title: **Statistical test for consistent estimation of causal effects in linear non-Gaussian models**

Authors: Doris Entner, Patrik O. Hoyer, Peter Spirtes

Presenter: Doris Entner

Description: In many fields of science researchers are faced with the problem of estimating causal effects from non-experimental data. A key issue is to avoid inconsistent estimators due to confounding, a problem commonly solved by 'adjusting for' a subset of the observed variables. When the data generating process is known, there exist simple graphical procedures for determining which subset of covariates should be adjusted for to obtain consistent estimators. However, when the graph is not known no general procedures for this task are available. In this poster we introduce such a method for linear non-Gaussian models, requiring only partial knowledge about the temporal ordering of the variables: We provide a simple statistical test for inferring whether an estimator of a causal effect is consistent when controlling for a subset of measured covariates, and we present heuristics to search for such a set.

28)

Title: **Causal Discovery for Linear Cyclic Models with Latent Variables**

Authors: Antti Hyttinen, Frederick Eberhardt, Patrik O. Hoyer

Presenter: Antti Hyttinen

Description: We show how to optimally use surgical experiments to discover linear causal models with cycles and latent variables.

29)

Title: **Noisy-OR Models with Latent Confounding**

Authors: Antti Hyttinen, Frederick Eberhardt, Patrik O. Hoyer

Presenter: Antti Hyttinen

Description: Generally causal models with latent variables are not identifiable from passive observational data or from experimental data in which only a few variables are subject to interventions at a time. The poster shows that if the local CPDs of the model are restricted to follow the noisy-OR parameterization, we can identify the causal model for example from experiments intervening on a single variable at a time and passive observational data.

Pattern and Link Discovery

30)
Title: **Relevant and Non-redundant Object Retrieval**
Authors: Laura Langohr, Hannu Toivonen
Presenter: Laura Langohr
Description: We address a setting of information retrieval where the user specifies query objects and the problem is to identify other objects that are relevant with respect to the query objects, but non-redundant with respect to each other. For example, in a text mining setting, the user or a program might want to get an overview of different uses or contexts of given terms. In bioinformatics again finding biological processes or pathways that are relevant both to a disease and a set of given genes helps to understand how they are related and may help identify possible shared biological mechanisms. While this area has been a popular topic of research, our contribution is to provide a simple, generic model that covers several related approaches while providing a systematic model for taking account of positive and negative query objects as well as non-redundancy of the output.

31)
Title: **Corpus-Based Generation of Content and Form in Poetry**
Authors: Jukka M. Toivanen, Hannu Toivonen, Alessandro Valitutti, Oskar Gross
Presenter: Oskar Gross
Description: We employ a corpus-based approach to generate content and form in poetry. The main idea is to use two different corpora, on one hand, to provide semantic content for new poems, and on the other hand, to generate a specific grammatical and poetic structure. The approach uses text mining methods, morphological analysis, and morphological synthesis to produce poetry in Finnish. We present some promising results obtained via the combination of these methods and preliminary evaluation results of poetry generated by the system.

32)
Title: **Ambiguous Lexical Resources for Computational Humor Generation**
Author: Alessandro Valitutti
Presenter: Alessandro Valitutti
Description: This work here is aimed to investigate to what extent it is possible to perform a feasible use of ambiguous texts in computational humor generation. The first core of a lexical database was developed in order to collect ambiguous terms in the English lexicon. Then an exploratory use of the resource for computational humor generation was performed. Finally, three existing prototypes of humor generator were simulated in order to generate different form of humorous messages from the same lexical resource.

33)
Title: **The use of weighted metabolic graphs to investigate the large-scale evolution of metabolism**
Authors: Fang Zhou, Ross D. King
Presenter: Fang Zhou
Description: We are interested in better understanding the evolution of metabolic biodiversity in bacteria - Archaea and the Eubacteria. To investigate this question, we introduce the use of weighted graphs to integrate large amounts of genomic data. We propose two ways of measuring the importance of enzymes, and apply the graph compression method to compare the importance of pathways in the different kingdoms.

Software Demonstrations

Title: **PULS: Surveillance of on-line news media**

Authors: Mian Du, Peter von Etter, Silja Huttunen, Roman Yangarber

Presenter: Roman Yangarber

Description: The PULS system mines streams of raw, plain-text news articles for particular kinds of structured, factual information in several languages, aggregates the discovered items into "big-picture" clusters, tries to classify the discovered information according to topic and relevance. We demonstrate the system on several real-world scenarios.