IP1

A Geometric Perspective on Machine Learning and Data Mining

Increasingly, we face machine learning problems in very high dimensional spaces. We proceed with the intuition that although natural data lives in very high dimensions, they have relatively few degrees of freedom. One way to formalize this intuition is to model the data as lying on or near a low dimensional manifold embedded in the high dimensional space. This point of view leads to a new class of algorithms that are ”manifold motivated” and a new set of theoretical questions that surround their analysis. A central construction in these algorithms is a graph or simplicial complex that is data-derived and we will relate the geometry of these to the geometry of the underlying manifold. Applications to embedding, clustering, classification, and semi-supervised learning will be considered.

Partha Niyogi
University of Chicago
Department of Computer Science
niyogi@cs.uchicago.edu

IP2

Automated Learning and Data Visualization

Automated numeric methods of data mining, statistics, and machine learning adapt themselves to systematic patterns in data to carry out predictive tasks, or to describe the patterns in a way that provides fundamental understanding. Data visualization is critical in all phases of the analysis of data, from the moment of arrival when data checking and cleaning are needed, to the final presentation of results. Visualization allows us to learn which patterns occur out of an immensely broad collection of possible patterns; it is difficult to select and carry out, a priori, automated learning methods to cover nearly as broad a collection of possibilities. It is widely accepted that an effective knowledge of patterns is necessary for fundamental understanding. But the knowledge can be of immense benefit for predictive tasks as well because it gives us valuable information about which automated numeric methods will likely produce best performance. Selecting best automated methods by trying a number of them in a training-test framework runs the risk of simply finding the best among a collection of poor performers. So visualization supports the automated methods. But the reverse is true, too. It is difficult to make progress just displaying raw data without the benefit of automated methods that provide fits to patterns, which are then displayed, and provide displays of remaining variation in the data after adjusting for the fits. Automation and visualization are symbiotic. Today, an immense challenge to data visualization, as it is to all technical areas of data analysis, is the rapid expansion in the size and complexity of datasets. This should not deter our commitment to an understanding of patterns in data, but does require new frameworks for how we approach data visualization. One such framework is visualization databases; for a single complex dataset, it consists of a large number of displays, many of which consist of many pages. The displays become a new database that is queried and studied on an as-needed basis. Production, management, and viewing a visualization database need many new ideas. For example, methods are needed for view selection to populate the database when the number of views can be millions or more. Examples are statistical sampling methods that find a representative collection of views, and automation algorithms that find interesting views by searching for certain patterns.

William S. Cleveland
Purdue University
Department of Statistics
wsc@purdue.edu

IP3

Semantics on the Web: How Do We Get There?

It is becoming increasingly clear that the next generation of web search and advertising will rely on a deeper understanding of user intent and task modeling, and a correspondingly richer interpretation of content on the web. How we get there, in particular, how we understand web content in richer terms than bags of words and links, is a wide open and fascinating question. I will discuss some of the options here, and look closely at the role that information extraction can play.

Raghu Ramakrishnan
Yahoo! Research
ramakris@yahoo-inc.com

IP4

Applied Nonparametric Bayes

Computer Science has historically been strong on data structures and weak on inference from data, whereas Statistics has historically been weak on data structures and strong on inference from data. One way to draw on the strengths of both disciplines is to develop “inferential methods for data structures’; i.e., methods that are based on probability distributions on recursively-defined objects such as trees, graphs, grammars and function calls. This is accommodated in the world of “nonparametric Bayes,’ where prior and posterior distributions are allowed to be general stochastic processes. In this talk I discuss a variety of applied problems that are naturally tackled from this point of view. I will discuss nonparametric Bayesian solutions to problems in natural language parsing, computational vision, information retrieval, statistical genetics and protein structural modeling.

Michael I. Jordan
University of California, Berkeley
jordan@cs.berkeley.edu

CP1

GAD: General Activity Detection for Fast Clustering on Large Data

Abstract not available at time of publication.

Xin Jin, Sangkyum Kim, Jiawei Han, Liangliang Cao, Zhijun Yin
University of Illinois at Urbana-Champaign
xinjin3@illinois.edu, kim71@illinois.edu, hanj@cs.uiuc.edu, cao4@ifp.uiuc.edu, zyin3@illinois.edu

CP1

Hybrid Clustering of Text Mining and Bibliometrics Applied to Journal Sets

A new hybrid clustering framework of integrating text mining and bibliometics is proposed . We propose a novel adaptive kernel K-means clustering algorithm to combine textual content an citation information for clustering.

Based on several validation indices, the experimental results, on a clustering problem of 1869 journals published in 2002-2006, demonstrate that our hybrid clustering strategy is able to provide clustering result as well as the best individual data source.

Xinhai Liu
K.U. Leuven, ESAT-SCD
Wuhan University of Science and Technology, College of Infor
xinhai.liu@esat.kuleuven.be

Shi Yu, Yves Moreau, Bart De Moor
K.U. Leuven, Dept. of Electrical Engineering ESAT-SCD
shi.yu@esat.kuleuven.be, yves.moreau@esat.kuleuven.be, bart.demoor@esat.kuleuven.be

Wolfgang Glanzel
K.U. Leuven, Steunpunt O
&O Indicatoren
wolfgang.glanzel@econ.kuleuven.ac.be

Frizo Janssens
K.U. Leuven, Dept. of Electrical Engineering ESAT-SCD
Attentio Company in Brussels
frizo.janssens@esat.kuleuven.be

CP1
Constraint-Based Subspace Clustering

Since the performance of traditional clustering algorithms decreases in high-dimensional data, subspace clustering techniques, which compute clusters in subsets of dimensions, have been developed. Nevertheless, due to the huge number of subspaces to consider, they often lack efficiency. In this paper we integrate background knowledge and, in particular, instance-level constraints to subspace clustering techniques and show experimentally that this increases not only the efficiency of the techniques but also the accuracy of the resultant clustering.

Elisa Fromont
Universite de Lyon, Laboratoire Hubert-Curien
UMR CNRS 5516
elisa.fromont@univ-st-etienne.fr

Adriana Prado
University of Antwerp
adriana.prado@ua.ac.be

Celine Robardet
INSA-Lyon, LIRIS UMR5205, F-69621 Villeurbanne, France
celine.robardet@insa-lyon.fr

CP1
Core: Nonparametric Clustering of Large Numeric Datasets

We propose CORE, a new nonparametric clustering technique that explicitly computes local maxima of the density and represents them with cores. CORE proposes an adaptive grid and gradients to define and compute cores of clusters. The incrementally constructed adaptive grid and gradients make the identification of cores robust, scalable, and independent of density fluctuations. The experimental studies show that CORE without any model parameters produces better quality clustering than related techniques and is efficient for large datasets.

Andrej Taliun
Free University of Bolzen-Bolzano
taliun@inf.unibz.it

Michael Böhlen
Free University of Bozen-Bolzano
boehlen@inf.unibz.it

Arturas Mazeika
Max-Planck-Institut für Informatik
amazeika@mpi-inf.mpg.de

CP1
Integrated Kl (K-Means - Laplacian) Clustering: A New Clustering Approach by Combining Attribute Data and Pairwise Relations

Most datasets in real applications come in from multiple sources. As a result, we often have attributes information about data objects and various pairwise relations (similarity) between data objects. Traditional clustering algorithms use either data attributes only or pairwise similarity only. We propose to combine K-means clustering on data attributes and normalized cut spectral clustering on pairwise relations. We show that these two methods can be coherently integrated together to make use of different data sources to obtain good clustering results. We also show that our integrated KL (K-means - Laplacian) clustering method can be naturally extended to semi-supervised clustering, data embedding and metric learning. Finally the experimental results on benchmark data sets are presented to show the effectiveness of our method.

Fei Wang
School of CIS, FIU
feiwang03@gmail.com

Chris Ding
University of Texas at Arlington
chqding@uta.edu

Tao Li
Florida International University
taoli@cs.fiu.edu

CP2
Proximity-Based Anomaly Detection Using Sparse Structure Learning

We proposed a new anomaly detection framework for correlation anomalies in highly noisy multivariate data. We show that fitting a *sparse* graphical model to the data is extremely useful to capture meaningful correlation changes. We then define the correlation anomaly scores by evaluating the distances between the fitted conditional distributions. Using real-world data, we demonstrate that our matrix-based sparse structure learning approach successfully detects correlation anomalies under collinearites and heavy noise.

Tsuyoshi Ide
IBM Research
Tokyo Research Lab.
goodidea@jp.ibm.com

Aurelie Lozano, Naoki Abe, Yan Liu

IBM Research
T. J. Watson Research Center
aclozano@us.ibm.com, nabe@us.ibm.com,
liuya@us.ibm.com

CP2
FuncICA for Time Series Pattern Discovery

FuncICA is a new independent component analysis method for pattern discovery in functional data like time series. FuncICA is an analog to functional PCA; instead of extracting components to minimize L_2 loss, we maximize independence of optimally-smoothed components over functional observations. Results for synthetic, gene expression, and electroencephalographic event-related potential data show FuncICA recovers scientific phenomena and improves classification accuracy. We conclude with a novel framework for fMRI data analysis using FuncICA.

Nishant A. Mehta
College of Computing
Georgia Institute of Technology
niche@cc.gatech.edu

Alexander Gray
Georgia Institute of Technology
agray@cc.gatech.edu

CP2
Event Discovery in Time Series

The discovery of events in time series can have important implications, such as identifying microlensing events in astronomical surveys. In this work, we develop probability models for calculating the significance of an arbitrary-sized sliding window and use these probabilities to find areas of significance. We apply our method to over 100,000 astronomical time series from the MACHO survey. In addition to successfully identifying known events, we were able to identify events that do not pass traditional event discovery procedures.

Dan R. Preston
Initiative in Innovative Computing, Harvard University
Department of Computer Science, Tufts University
dan.preston@tufts.edu

Pavlos Protopapas
Initiative in Innovative Computing, Harvard University
Harvard-Smithsonian Center for Astrophysics
pprotopapas@cfa.harvard.edu

Carla Brodley
Department of Computer Science, Tufts University
brodley@cs.tufts.edu

CP2
Optimal Distance Bounds on Time-Series Data

We present new mechanisms for very fast search operations over the compressed time-series data, with specific focus on weblog data. An important contribution of this work is the derivation of optimally tight bounds on the Euclidean distance estimation between compressed sequences. Since our methodology is applicable to sequential data in general, the proposed technique is of independent interest.Additionally, our distance estimation strategy is not tied to a specific compression methodology, but can be applied on top of any orthonormal based compression technique (Fourier, Wavelet, PCA, etc). The experimental results indicate that the new optimal bounds lead to a significant improvement in the pruning power of search compared to previous state-of-the-art, in many cases eliminating more than 80% of the candidate search sequences.

Michail Vlachos
IBM Research
michalis0@gmail.com

Serdar Kozat
Koc University
Istanbul, Turkey
serdarkozat@gmail.com

Philip S Yu
University of Chicago
psyu@cs.uic.edu

CP2
Autocannibalistic and Anyspace Indexing Algorithms with Applications to Sensor Data Mining

Efficient indexing is at the heart of many data mining algorithms. A simple and extremely effective algorithm for indexing under any metric space was introduced in 1991 by Orchard. Orchard's algorithm has not received much attention in the data mining and database community because of a fatal flaw; it requires quadratic space. In this work we show that we can produce a reduced version of Orchard's algorithm that requires much less space, but produces nearly identical speedup. We achieve this by casting the algorithm in an anyspace framework, allowing deployed applications to take as much of an index as their main memory/sensor can afford.

Lexiang Ye, Xiaoyue Wang, Eamonn Keogh
University of California, Riverside
lexiangy@cs.ucr.edu, xwang@cs.ucr.edu,
eamonn@cs.ucr.edu

Agenor Mafra-Neto
ISCA Technologies
president@iscatech.com

CP3
Prior-Free Rare Category Detection

Rare category detection is an open challenge in machine learning. In this paper, we propose a new method for rare category detection named SEDER, which requires no prior information about the data set. It implicitly performs semiparametric density estimation using specially designed exponentially families, and then picks the examples for labeling where the neighborhood density changes the most. Experimental results on both synthetic and real data sets demonstrate the superiority of SEDER.

Jingrui He
Machine Learning Department
Carnegie Mellon University
jingruih@cs.cmu.edu

Jaime Carbonell
Language Technologies Institute
Carnegie Mellon University
jgc@cs.cmu.edu

CP3

Learning Random-Walk Kernels for Protein Remote Homology Identification and Motif Discovery

It is very difficult to choose the optimal number of random steps in random-walk kernels. In this paper, we will discuss how to better identify protein remote homology than any other algorithm using a learned random-walk kernel based on a positive linear combination of random-walk kernels with different random steps, which leads to a convex combination of kernels. The resulting kernel has much better prediction performance than the state-of-the-art profile kernel for protein remote homology identification. Moreover, our approach based on learned random-walk kernels can effectively identify meaningful protein sequence motifs that are responsible for discriminating the memberships of protein sequences' remote homology in SCOP.

Renqiang Min
Dept Computer Science
University of Toronto
minrq@cs.toronto.edu

Rui Kuang
Dept Computer Science
University of Minnesota
kuang@cs.umn.edu

Anthony Bonner
Dept Computer Science
University of Toronto
bonner@cs.toronto.edu

Zhaolei Zhang
Banting and Best Dept of Medical Research
University of Toronto
zhaolei.zhang@utoronto.ca

CP3

Application of Bayesian Partition Models in Warranty Data Analysis

Warranty data analysis helps automotive engineers in their task of resolving manufacturing or design related quality issues. In this contribution we outline how Bayesian partition models can be integrated with interactive decision trees to support root cause investigations. Our approach considers taxonomies and identifies the most likely, semantically meaningful partitions that are close to the concept that actually caused a quality issue. Real-world case studies illustrate how the approach is applied in practice.

Markus Mueller, Christoph Schlieder
University of Bamberg
Markus.Mueller@uni-bamberg.de,
christoph.schlieder@uni-bamberg.de

Axel Blumenstock
Daimler AG
axel.blumenstock@daimler.com

CP3

A Family of Large Margin Linear Classifiers and Its Application in Dynamic Environments

We combine regularization mechanisms with online large margin learning algorithms to learn robust classifiers in nonstationary environments. We prove bounds on their error and show that removing features with small weights has little influence on the accuracy, suggesting that these methods exhibit feature selection ability. We show that such regularized learning algorithms automatically decrease the influence of the old training instances and focus on the more recent ones.

Jianqiang Shen, Thomas Dietterich
School of EECS, Oregon State University
shenj@eecs.oregonstate.edu, tgd@eecs.oregonstate.edu

CP3

Outlier Detection with Globally Optimal Exemplar-Based Gmm

Outlier detection has recently become an important problem in many data mining applications. In this paper, a novel unsupervised algorithm for outlier detection is proposed. First we apply a provably globally optimal Expectation Maximization (EM) algorithm to t a Gaussian Mixture Model (GMM) to a given data set. In our approach, a Gaussian is centered at each data point, and hence, the estimated mixture proportions can be interpreted as probabilities of being a cluster center for all data points. The outlier factor at each data point is then dened as a weighted sum of the mixture proportions with weights representing the similarities to other data points. The proposed outlier factor is thus based on global properties of the data set. This is in contrast to most existing approaches to outlier detection, which are strictly local. Our experiments performed on several simulated and real life data sets demonstrate superior performance of the proposed approach. Moreover, we also demonstrate the ability to detect unusual shapes.

Xingwei Yang
Department of Computer and Information Science
Temple University
xingwei@temple.edu

Longin Jan Latecki
Temple University
Department of Computer and Information Science
latecki@temple.edu

Dragoljub Pokrajac
Delaware State University
dragoljub.pokrajac@comcast.net

CP4

Discovering Substantial Distinctions Among Incremental Bi-Clusters

A fundamental task of data analysis is comprehending what distinguishes clusters found within the data. We present the problem of mining distinguishing sets which seeks to find sets of objects or attributes that induce that most change among the incremental bi-clusters of a binary dataset. Unlike emerging patterns and contrast sets which only focus on statistical differences between support of itemsets, our approach considers distinctions in both the attribute space and the object space. Viewing the lattice of bi-clusters formed within a data set as a weighted directed graph, we mine the most significant distinguishing sets by growing a maximal cost spanning tree of the lattice. In this paper we present a weighting function for measuring distinction among bi-clusters in the lattice and the novel MIDS algorithm. MIDS simultaneously enumerates bi-clusters, constructs the bi-cluster lattice, and computes

the distinguishing sets. The efficient computational performance of MIDS is exhibited in a performance test on real world and benchmark data sets. The utility of distinguishing sets is also demonstrated with experiments on synthetic and real data.

Faris Alqadah, Raj Bhatnagar
Universtiy of Cincinnati
alqadaf@email.uc.edu, raj.bhatnagar@uc.edu

CP4
A Framework for Exploring Categorical Data

In this paper, we present a framework for categorical data analysis which allows such data sets to be explored using a rich set of techniques that are only applicable to continuous data sets. We introduce the concept of separability statistics in the context of exploratory categorical data analysis. We show how these statistics can be used as a way to map categorical data to continuous space given a labeled reference data set. This mapping enables visualization of categorical data using techniques that are applicable to continuous data. We show that in the transformed continuous space, the performance of the standard k-nn based outlier detection technique is comparable to the performance of the k-nn based outlier detection technique using the best of the similarity measures designed for categorical data. The proposed framework can also be used to devise similarity measures best suited for a particular type of data set.

Varun Chandola, Shyam Boriah
Department of Computer Science
University of Minnesota
chandola@cs.umn.edu, sboriah@cs.umn.edu

Vipin Kumar
University of Minnesota
kumar@cs.umn.edu

CP4
DensEst: Density Estimation for Data Mining in High Dimensional Spaces

Subspace clustering and frequent itemset mining algorithms do not scale to large high dimensional databases as the search space gets enormous. Efficiency improvements can be achieved by estimates of object counts in selective subspace regions. In this work, we propose DensEst, an efficient density estimator. By incorporating correlations between dimensions DensEst achieves highly accurate estimations. We integrated DensEst into subspace clustering and frequent itemset mining algorithms and show both, their improved efficiency and accuracy.

Emmanuel Müller
RWTH Aachen University
mueller@cs.rwth-aachen.de

Ira Assent
Aalborg University
ira@cs.aau.dk

Ralph Krieger, Stephan Günnemann, Thomas Seidl
RWTH Aachen University
krieger@cs.rwth-aachen.de,
guennemann@informatik.rwth-aachen.de,
seidl@informatik.rwth-aachen.de

CP4
Bayesian Cluster Ensembles

Cluster ensembles provide a framework for combining multiple base clusterings of a dataset to generate a stable and robust consensus clustering. There are important variants of the basic cluster ensemble problem, notably including cluster ensembles with missing values, as well as row-distributed or column-distributed cluster ensembles. Existing cluster ensemble algorithms are applicable only to a small subset of these variants. In this paper, we propose Bayesian Cluster Ensembles (BCE), which is a mixed-membership model for learning cluster ensembles, and is applicable to all the primary variants of the problem. We propose two methods, respectively based on variational approximation and Gibbs sampling, for learning a Bayesian cluster ensemble. We compare BCE extensively with several other cluster ensemble algorithms, and demonstrate that BCE is not only versatile in terms of its applicability, but also outperforms the other algorithms in terms of stability and accuracy.

Hanhuai Shan
Department of Computer Science and Engineering
University of Minnesota, Twin Cities
shan@cs.umn.edu

Hongjun Wang
School of Computer Science
Sichuan University, Chengdu, China
wanghongjun@cs.scu.edu.cn

Arindam Banerjee
University of Minnesota
banerjee@cs.umn.edu

CP4
Agglomerative Mean-Shift Clustering Via Query Set Compression

Mean-Shift (MS) is a powerful non-parametric clustering method. Although good accuracy can be achieved, its computational cost is particularly expensive even on moderate data sets. In this paper, for the purpose of algorithm speedup, we develop an agglomerative MS clustering method called Agglo-MS, along with its mode-seeking ability and convergence property analysis. Our method is built upon an iterative query set compression mechanism which is motivated by the quadratic bounding optimization nature of MS. The whole framework can be efficiently implemented in linear running time complexity. Furthermore, we show that the pairwise constraint information can be naturally integrated into our framework to derive a semi-supervised non-parametric clustering method. Extensive experiments on toy and real-world data sets validate the speedup advantage and numerical accuracy of our method, as well as the superiority of its semi-supervised version.

Xiao-Tong Yuan, Bao-Gang Hu, Ran He
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
xtyuan@nlpr.ia.ac.cn, hubaogang@gmail.com,
rhe@nlpr.ia.ac.cn

CP5
Scalable Distributed Change Detection from Astronomy Data Streams Using Local, Asynchronous

Eigen Monitoring Algorithms

This paper considers the problem of change detection using distributed eigen monitoring algorithms for astronomy data pipelines. Change point detection in such datasets may provide useful insights to unique astronomical phenomenon. However, this is a challenging problem for such high-throughput distributed data streams. In this paper we propose a highly scalable and distributed asynchronous algorithm for monitoring the eigenstates of such data streams. Experiments performed on SDSS catalogue data show the effectiveness of the algorithm.

Kamalika Das
University of Maryland, Baltimore County
kdas1@cs.umbc.edu

Kanishka Bhaduri
Mission Critical Tech
NASA Ames Research Center
kanishka.bhaduri-1@nasa.gov

Sugandha Arora, Wesley Griffin
Dept of CSEE
University of Maryland, Baltimore County
a56@umbc.edu, griffin5@umbc.edu

Kirk Borne
Computational and Data Sciences Dept
George Mason University
kborne@gmu,edu

Chris Giannella
Computer Science Dept
New Mexico State University
cgiannel@acm.org

Hillol Kargupta
Department of Computer Science
University of Maryland Baltimore County
hillol@cs.umbc.edu

CP5
Adaptive Concept Drift Detection

An established method to detect concept drift in data streams is to perform statistical hypothesis testing on the multivariate data in the stream. Statistical decision theory offers rank-based statistics for this task. However, these statistics depend on a fixed set of characteristics of the underlying distribution. Thus, they work well whenever the change in the underlying distribution affects these properties measured by the statistic, but they perform not very well, if the drift influences the characteristics caught by the test statistic only to a small degree. To address this problem, we present three novel drift detection tests, whose test statistics are dynamically adapted to match the actual data at hand. The first one is based on a rank statistic on density estimates for a binary representation of the data, the second compares average margins of a linear classifier induced by the 1-norm support vector machine (SVM), and the last one is based on the average zero-one or sigmoid error rate of an SVM classifier. Experiments show that the margin- and error-based tests outperform the multivariate Wald-Wolfowitz test for concept drift detection. We also show that the tests work even if the drift is gradual in nature and that the new methods are faster than the Wald-Wolfowitz test.

Ulrich Rückert
International Computer Science Institute
rueckert@icsi.berkeley.edu

Anton Dries
Katholieke Universiteit Leuven
anton.dries@cs.kuleuven.be

CP5
Positive Unlabeled Learning for Data Stream Classification

This paper studies how to devise PU learning techniques for the data stream environment. Unlike existing data stream classification methods that assume both positive and negative training data are available for learning, we propose a novel PU learning technique LELC (PU Learning by Extracting Likely positive and negative micro-Clusters) for document classification. LELC only requires a small set of positive examples and a set of unlabeled examples which is easily obtainable in the data stream environment to build accurate classifiers.

Xiaoli Li
Institute for Infocomm Research
xlli@i2r.a-star.edu.sg

Philip Yu, Bing Liu
University of Illinois at Chicago
psyu@cs.uic.edu, liub@cs.uic.edu

See-Kiong Ng
Institute for Infocomm Research
skng@i2r.a-star.edu.sg

CP5
Time-Decayed Correlated Aggregates over Data Streams

Data stream analysis frequently relies on identifying correlations and posing conditional queries on the data after it has been seen. *Correlated aggregates* form an important example of such queries. Since recent events are typically more important, *time decay* is used to downweight old values. In this talk, we present space-efficient algorithms as well as space lower bounds for the time-decayed correlated sum, a problem at the heart of many related aggregations.

Graham Cormode
AT&T Labs-Research
graham@research.att.com

Srikanta Tirthapura, Bojian Xu
Department of Electrical and Computer Engineering
Iowa State University
snt@iastate.edu, bojianxu@iastate.edu

CP5
Multi-Modal Hierarchical Dirichlet Process Model for Predicting Image Annotation and Image-Object Label Correspondence

We address the problem of predicting image captions and labels for individual objects in the image using a multimodal hierarchical Dirichlet Process model (MoM-HDP). The model groups related words and image features using hidden mixture components and using a stochastic process for generating the mixture components, thus allowing to circumvent the need for a priori choice of the number of mixture components or the computational expense of model selection. The model parameters are estimated efficiently using variational inference. It is then evaluated for image annotation task and object recognition task on 2 large scale real world image datasets.

Oksana Yakhnenko, Vasant Honavar
Computer Science Department
Iowa State University
oksayakh@cs.iastate.edu, honavar@cs.iastate.edu

CP6

Hierarchical Linear Discriminant Analysis for Beamforming

We demonstrate the applicability of the recently proposed hierarchical linear discriminant analysis (h-LDA) to beamforming. h-LDA tackles the unimodal limitation of LDA by variance decomposition in subcluster level. We present an efficient h-LDA algorithm using Cholesky decomposition and generalized singular value decomposition for oversampled data, and analyze its data model. Our experiments for beamforming simulation show that h-LDA outperforms LDA, kernel discriminant analysis, the regularized least squares and the kernelized support vector regression.

Jaegul Choo
College of Computing
Georgia Institute of Technology
joyfull@cc.gatech.edu

Barry L. Drake
SEAL/AMDD
Georgia Tech Research Institute
barry.drake@gtri.gatech.edu

Haesun Park
Georgia Institute of Technology
hpark@cc.gatech.edu

CP6

Toward Optimal Ordering of Prediction Tasks

We study the problem of ordering a series of interdependent prediction tasks that must be accomplished sequentially through user interaction. We propose an approximate formulation in terms of pairwise task order preferences, reducing it to the well-known Linear Ordering Problem. Our experiments on two practical applications show encouraging improvements in predictive performance, as compared to approaches that do not take task dependencies into account.

Abhimanyu Lad
Language Technologies Institute
Carnegie Mellon University
alad@cs.cmu.edu

Yiming Yang
Carnegie Mellon University
yiming@cs.cmu.edu

Rayid Ghani
Accenture Technology Labs
rayid.ghani@accenture.com

Bryan Kisiel
Language Technologies Institute
Carnegie Mellon University
bkisiel@cs.cmu.edu

CP6

Amori: A Metric-Based One Rule Inducer

The objectives of data mining applications vary extensively. We have implemented a supervised concept learner called A Metric-based One Rule Inducer (AMORI), for which it is possible to select the learning/objective metric based on the problem at hand. We have compared the performance of this algorithm on 19 UCI data sets by embedding three different learning metrics. Experiments show that a performance gain is achieved when using identical metrics for learning and evaluation.

Niklas Lavesson, Paul Davidsson
Blekinge Institute of Technology
niklas.lavesson@bth.se, paul.davidsson@bth.se

CP6

The Metric Dilemma: Competence-Conscious Associative Classification

The classification performance of an associative classifier is strongly dependent on the statistic measure or metric that is used to quantify the strength of the association between features and classes (i.e., confidence, correlation etc.). Previous studies have shown that classifiers produced by different metrics may provide conflicting predictions, and that the best metric to use is data-dependent and rarely known while designing the classifier. This uncertainty concerning the optimal match between metrics and problems is a dilemma, and prevents associative classifiers to achieve their maximal performance. This dilemma is the focus of this paper.

Adriano Veloso
UFMG
adrianov@dcc.ufmg.br

Mohammed Zaki
Rensselaer Polytechnic Institute
zaki@cs.rpi.edu

Wagner Meira JR., Marcos GONALVES
UFMG
meira@dcc.ufmg.br, mgoncalv@dcc.ufmg.br

CP6

Twin Vector Machines for Online Learning on a Budget

This paper proposes Twin Vector Machine (TVM), a constant space and sublinear time Support Vector Machine (SVM) algorithm for online learning. TVM achieves its favorable scaling by maintaining only a fixed number of examples, called the twin vectors, and their associated information in memory during training. In addition, TVM guarantees that Kuhn-Tucker conditions are satisfied on all twin vectors at any time. To maximize the accuracy of TVM, twin vectors are adjusted during the training phase in order to approximate the data distribution near the decision boundary. Given a new training example, TVM is updated in three steps. First, the new example is added as a new twin vector if it is near the decision boundary.

If this happens, two twin vectors are selected and merged into a single twin vector to maintain the budget. Finally, TVM is updated by incremental and decremental learning to account for the change in twin vectors. Several methods for twin vector merging were proposed and experimentally evaluated. TVMs were thoroughly tested on 12 large data sets. In most cases, the accuracy of low-budget TVMs was comparable to the state of the art resource-unconstrained SVMs. Additionally, the TVM accuracy was substantially larger than that of SVM trained on a random sample of the same size. Even larger difference in accuracy was observed when comparing to Forgetron, a popular kernel perceptron algorithm on a budget. The results illustrate that highly accurate online SVMs could be trained from large data streams using devices with severely limited memory budgets.

Slobodan Vucetic, Zhuang Wang
Temple University
vucetic@ist.temple.edu, zhuang@temple.edu

CP7

Identifying Unsafe Routes for Network-Based Trajectory Privacy

We propose a privacy model that offers trajectory privacy to the requesters of LBSs. Our model assumes movement on a road network as well as attackers who have knowledge of the users' movement statistics. The privacy model has been implemented as a framework that automatically identifies routes where user privacy is at risk. Then, it anonymizes user requests based on the location of the requester (w.r.t. his/her unsafe routes) from the time of request until the service provision.

Aris Gkoulalas-Divanis, Vassilios Verykios
Department of Computer & Communication Engineering
University of Thessaly
arisgd@inf.uth.gr, verykios@inf.uth.gr

Mohamed Mokbel
Department of Computer Science and Engineering
University of Minnesota
mokbel@cs.umn.edu

CP7

Privacy Preservation in Social Networks with Sensitive Edge Weights

In addition to the current social network anonymity de-identification techniques, in this paper we consider perturbing the weights of some edges to preserve data privacy when the network is published, while retaining the shortest path and the approximate cost of the path between some pairs of nodes in the original network. We develop two privacy-preserving strategies for this application, the Gaussian randomization multiplication and the greedy perturbation.

Jun Zhang, Lian Liu
University of Kentucky
Department of Computer Science
jzhang@cs.uky.edu, lliuc@csr.uky.edu

Jie Wang
Minnesota State University Mankato
Computer Science Department
jie.wang@mnsu.edu

Jinze Liu
University of Kentucky
Department of Computer Science
liuj@cs.uky.edu

CP7

A Bayesian Approach Toward Finding Communities and Their Evolutions in Dynamic Social Networks

In this paper, we propose a dynamic stochastic block model for finding communities and their evolutions in a dynamic social network. In this study, we employ a Bayesian treatment for parameter estimation that computes the posterior distributions for all the unknown parameters. Extensive experimental studies based on both synthetic data and real-life data demonstrate that our model achieves higher accuracy and reveals more insights in the data than several state-of-the-art algorithms.

Tianbao Yang
Michigan State University
yangtia1@msu.edu

Yun Chi, Shenghuo Zhu, Yihong Gong
NEC Laboratories America
ychi@sv.nec-labs.com, zsh@sv.nec-labs.com,
ygong@sv.nec-labs.com

Rong Jin
Michigan State University
rongjin@msu,.edu

CP7

Graph Generation with Prescribed Feature Constraints

In this paper, we study the problem of how to generate synthetic graphs matching various properties of a real social network with two applications, privacy preserving social network publishing and significance testing of network analysis results. We investigate potential disclosures of sensitive links due to the preserved features. Our algorithms on graph generation are based on the Metropolis-Hastings sampling.

Xiaowei Ying, Xintao Wu
University of North Carolina at Charlotte
xying@uncc.edu, xwu@uncc.edu

CP7

Detecting Communities in Social Networks Using Max-Min Modularity

Many datasets can be described in the form of graphs or networks where nodes in the graph represent entities and edges represent relationships between pairs of entities. A common property of these networks is their community structure, considered as clusters of densely connected groups of vertices, with only sparser connections between groups. The identification of such communities relies on some notion of clustering or density measure. which defines the communities that can be found. However, previous community detection methods usually apply the same structural measure on all kinds of networks, despite their distinct dissimilar features. In this paper, we present a new community mining measure, Max-Min Modularity, which considers both connected pairs and criteria defined by domain experts in finding communities, and then specify a hierarchical clustering algorithm to detect communities in networks. When applied to real world networks for which the community structures are already known, our method shows improvement over previous algorithms. In addition, when applied to randomly generated networks for which we only have approximate information about communities, it gives promising results which shows the algorithm's robustness against noise.

Jiyang Chen, Osmar Zaiane, Randy Goebel
University of Alberta
jiyang@cs.ualberta.ca, zaiane@cs.ualberta.ca, goebel@cs.ualberta.ca

CP8
Efficient Discovery of Interesting Patterns Based on Strong Closedness

Regarding all patterns above a certain frequency threshold as interesting is one way of defining interestingness in frequent pattern mining. We argue that in many applications, a different notion of interestingness is required in order to be able to capture "long', and thus particularly informative, patterns that are correspondingly of low frequency. To identify such patterns, we propose a new measure of interestingness that is based on their degree of closedness.

Mario Boley
Fraunhofer IAIS
mario.boley@iais.fraunhofer.de

Tamas Horvath
University of Bonn
Fraunhofer IAIS
tamas.horvath@iais.fraunhofer.de

Stefan Wrobel
Fraunhofer IAIS
University of Bonn
stefan.wrobel@iais.fraunhofer.de

CP8
Efficient Computation of Partial-Support for Mining Interesting Itemsets

Mining interesting itemsets is a popular topic in the data mining community. The objective of this problem is to mine all interesting itemsets, with respect to a given interestingness measure. While considerable efforts have being spent on justifying the various interestingness measures, the algorithms that mine them are not quite well-studied, except in the case *support*, which has resulted in the famous *frequent* itemset mining (FIM) problem. In this paper, we show that a certain *class* of interesting itemsets can be represented by functions of their *partial support.* This class includes some definitions of fault-tolerant itemsets, *estimated support* of itemsets in noisy data, and *bond* of itemsets. As the name implies, partial support of an itemset is the number of transactions containing *some part* of the given itemset. This paper addresses the problem of efficiently calculating partial supports, which leads to efficient algorithms for mining interesting itemsets in that class. We show that there exists a recurrence relation between partial supports. Hence, we can calculate the partial supports of itemset by simply extending any FIM algorithm (even the implementation). This allows us to benefit from innovations and optimizations in FIM algorithms. Theoretical analysis shows that our approaches retain the running time complexity of the base FIM algorithms for only a small cost in space. Extensive experiments on several real-world datasets also demonstrate that algorithms based on our approach are significantly faster than previously proposed techniques for corresponding definitions.

Ardian K. Poernomo, Vivekanand Gopalkrishnan
Nanyang Technological University, Singapore
ardi0002@ntu.edu.sg, asvivek@ntu.edu.sg

CP8
Top-K Correlative Graph Mining

We study the problem of mining top-k correlative subgraphs from a graph database, which share similar occurrence distributions with a given query graph. We propose an efficient algorithm, TopCor, which effectively directs the search to highly correlative candidates by three key techniques: an effective correlation checking mechanism, a powerful pruning criteria, and a set of rules for candidate exploration. Experiments show that TopCor is significantly faster than CGSearch, the state-of-the-art threshold-based correlative graph mining algorithm.

Yiping Ke
The Chinese University of Hong Kong
ypke@se.cuhk.edu.hk

James Cheng
Nanyang Technological University
j.cheng@acm.org

Jeffrey Yu
The Chinese University of Hong Kong
yu@se.cuhk.edu.hk

CP8
Grammar Mining

We introduce the problem of grammar mining, where patterns are context-free grammars, as a generalization of a large number of common pattern mining tasks, such as tree, sequence and itemset mining. The proposed system offers data miners the possibility to specify and explore pattern domains declaratively, in a way which is very similar to the declarative specification of regular expressions in popular scripting languages.

Siegfried Nijssen, Luc De Raedt
Katholieke Universiteit Leuven
siegfried.nijssen@cs.kuleuven.be,
luc.deraedt@cs.kuleuven.be

CP8
High Performance Parallel/Distributed Biclustering Using Barycenter Heuristic

Biclustering refers to simultaneous clustering of objects and their features. It has been shown that Bipartite Spectral Partitioning can be reformulated as a graph drawing problem where objective is to minimize Hall's energy of the bipartite graph representation of the input data. We provide an embarrassingly parallel algorithm for biclustering, based on parallel energy minimization using barycenter heuristic. Experimental evaluation shows large superlinear speedups, scalability and high level of accuracy.

Arifa Nisar

Northwestern University
Department of Electrical Engineering and Computer Science
a-nisar@u.northwestern.edu

Waseem Ahmad
A9.COM
wahmad@acm.org

Wei-Keng Liao, Alok Choudhary
Department of Electrical Engineering and Computer Science
Northwestern University
wkliao@ece.northwestern.edu,
choudhar@ece.northwestern.edu

CP9

Polynomial-Delay and Polynomial-Space Algorithms for Mining Closed Sequences, Graphs, and Pictures in Accessible Set Systems

This paper studies efficient closed pattern mining from semi-structured data. By modeling semi-structured data with a framework of set systems, we present an efficient depth-first algorithm that finds all closed patterns in accessible set systems without duplicates in polynomial-delay and polynomial-space w.r.t. the total input size. We also apply this result to efficient closed pattern mining for classes of semi-structured patterns including rigid sequence motifs, itemset sequences, relational graphs, 2-D convex hulls, and 2-D picture patterns.

Hiroki Arimura
Graduate School of IST, Hokkaido University
arim@ist.hokudai.ac.jp

Takeaki Uno
National Institute of Informatics
uno@nii.jp

CP9

Link Propagation: A Fast Semi-Supervised Learning Algorithm for Link Prediction

We propose Link Propagation as a new semi-supervised learning method for link prediction problems, where the task is to predict unknown parts of the network structure by using auxiliary information such as node similarities. Since the proposed method can fill in missing parts of tensors, it is applicable to multi-relational domains, allowing us to handle multiple types of links simultaneously. We also give a novel efficient algorithm for Link Propagation based on an accelerated conjugate gradient method.

Hisashi Kashima
IBM Research, Tokyo Research Laboratory
hkashima@jp.ibm.com

CP9

MultiVis: Content-Based Social Network Exploration Through Multi-Way Visual Analysis

With the explosion of social media, scalability becomes a key challenge. There are two main aspects of the problems that arise: 1) data volume: how to manage and analyze huge datasets to efficiently extract patterns, 2) data understanding: how to facilitate understanding of the patterns by users? To address both aspects of the scalability challenge, we present a hybrid approach that leverages two complementary disciplines, data mining and information visualization. In particular, we propose 1) an analytic data model for content-based networks using tensors; 2) an efficient high-order clustering framework for analyzing the data; 3) a scalable context-sensitive graph visualization to present the clusters. We evaluate the proposed methods using both synthetic and real datasets. In terms of computational efficiency, the proposed methods are an order of magnitude faster compared to the baseline. In terms of effectiveness, we present several case studies of real corporate social networks.

Jimeng Sun
IBM T.J. Watson Research Center
jimeng@cs.cmu.edu

Spiros Papadimitriou, Ching-Yung Lin, Nan Cao, Shixia Liu, Weihong Qian
IBM Reserach
spapadim@cs.cmu.edu, chingyung@us.ibm.com, nancao@cn.ibm.com, liusx@cn.ibm.com, qianwh@cn.ibm.com

CP9

Near-Optimal Supervised Feature Selection Among Frequent Subgraphs

Graph classification is an increasingly important step in numerous application domains. A popular classification approach using frequent subgraphs suffers from the enormous problem that the number of extracted features may grow exponentially with the size of the graphs. In order to ensure an efficient graph representation of high discriminative power, we propose a submodular approach to feature selection on frequent subgraphs which can be integrated into gSpan, the state-of-the-art tool for frequent subgraph mining.

Marisa Thoma
Institute for Informatics
Ludwig-Maximilians-Universität München
thoma@dbs.ifi.lmu.de

Hong Cheng
Department of Systems Engineering and Engineering Management
Chinese University of Hong Kong
hcheng3@ad.uiuc.edu

Arthur Gretton
Max-Planck Institute for Biological Cybernetics
Tübingen
arthur.gretton@tuebingen.mpg.de

Jiawei Han
University of Illinois at Urbana-Champaign
hanj@cs.uiuc.edu

Hans-Peter Kriegel
Ludwig-Maximilians University Munich
kriegel@dbs.ifi.lmu.de

Alex Smola
Yahoo! Research, Santa Clara, California
alex@smola.org

Le Song
School of Computer Science

Carnegie Mellon University
lesong@it.usyd.edu.au

Philip Yu
University of Illinois at Chicago
psyu@cs.uic.edu

Xifeng Yan
Department of Computer Science
University of California at Santa Barbara
xyan@cs.ucsb.edu

Karsten Borgwardt
University of Cambridge
Max-Planck-Institutes for Biological
Cybernetics,Tübingen
karsten.borgwardt@tuebingen.mpg.de

CP9

Understanding Importance of Collaborations in Co-Authorship Networks: A Supportiveness Analysis Approach

In co-authorship networks, the fact two authors co-author one paper can be regarded as one author supports the other's scientific work. Such characteristics can be measured by supportiveness, a novel and interesting measure on co-authorship relation. In our work, several efficient algorithms are developed to compute the top-n most supportive authors and most supportive groups. The empirical study conducted on DBLP data set indicates the supportiveness measures are interesting, and methods are effective and efficient.

Yi Han
National University of Defense Technology
yihan@nudt.edu.cn

Bin Zhou
Simon Fraser University
bzhou@cs.sfu.ca

Jian Pei
School of Computing Science
Simon Fraser University
jpei@cs.sfu.ca

Yan Jia
National University of Defense Technology
jiayanjy@vip.sina.com

CP10

Local Relevance Weighted Maximum Margin Criterion for Text Classification

We propose a feature extraction method for text classification, named Local Relevance Weighted Maximum Margin Criterion. It aims to learn a subspace in which the documents in the same class are as near as possible while the documents in the different classes are as far as possible in the local region of each document. Furthermore, the relevance is taken into account as a weight to determine the extent to which the documents will be projected.

Quanquan Gu, Jie Zhou
Department of Automation, Tsinghua University
gqq03@mails.thu.edu.cn, jzhou@tsinghua.edu.cn

CP10

Parallel Large Scale Feature Selection for Logistic Regression

In this paper we examine the problem of efficient feature evaluation for logistic regression on very large data sets. We present a new forward feature selection heuristic that ranks features by their estimated effect on the resulting model's performance. An approximate optimization, based on backfitting, provides a fast and accurate estimate of each new feature's coefficient in the logistic regression model. Further, the algorithm is highly scalable by parallelizing simultaneously over both features and records, allowing us to quickly evaluate billions of potential features even for very large data sets.

Sameer Singh
University of Massachusetts, Amherst
Department of Computer Science
sameer@cs.umass.edu

Jeremy Kubica, Scott Larsen
Google Inc.
Pittsburgh PA 15213
jkubica@google.com, esl@google.com

Daria Sorokina
Carnegie Mellon University
Pittsburgh PA 15213
daria@cs.cmu.edu

CP10

Multi-Topic Based Query-Oriented Summarization

In this paper, we study a new setup of the problem of multi-topic based query-oriented summarization. We propose using a probabilistic approach to solve this problem. More specifically, we propose two strategies to incorporate the query information into a probabilistic model. Experimental results on two different genres of data show that our proposed approach can effectively extract a multi-topic summary from a document collection and the summarization performance is better than baseline methods.

Jie Tang
Tsinghua University
jietang@tsinghua.edu.cn

Limin Yao
University of Massachusetts Amherst
lmyao@cs.umass.edu

Dewei Chen
Tsinghua University
chendw@keg.cs.tsinghua.edu.cn

CP10

Straightforward Feature Selection for Scalable Latent Semantic Indexing

Latent Semantic Indexing (LSI) has been validated to be effective on many small scale text collections. However, little evidence has shown its effectiveness on unsampled large scale text corpus due to its high computational complexity. In this paper, we propose a straightforward feature selection strategy, which is named as Feature Selection for Latent Semantic Indexing (FSLSI), as a preprocessing step such that LSI can be efficiently approximated on large scale

text corpus.

Jun Yan
Microsoft Research Asia
junyan@microsoft.com

Shuicheng Yan
National University of Singapore
eleyans@nus.edu.sg

Ning Liu, Zheng Chen
Microsoft Research Asia
ningl@microsoft.com, zhengc@microsoft.com

CP10
Topic Cube: Topic Modeling for OLAP on Multidimensional Text Databases

As the amount of textual information grows explosively in various kinds of business systems, it becomes more and more desirable to analyze both structured data records and unstructured text data simultaneously. While OLAP techniques have been proven very useful for analyzing and mining structured data, they face challenges in handling text data. On the other hand, probabilistic topic models are among the most effective approaches to latent topic analysis and mining on text data. In this lecture, we will describe a new data model called *topic cube* which combines OLAP with probabilistic topic modeling and enables OLAP on the dimension of text data in a multidimensional text database.

Duo Zhang, Chengxiang Zhai
University of Illinois at Urbana Champaign
dzhang22@uiuc.edu, czhai@cs.uiuc.edu

Jiawei Han
University of Illinois at Urbana-Champaign
hanj@cs.uiuc.edu

CP11
Travel-Time Prediction Using Gaussian Process Regression: A Trajectory-Based Approach

We tackle the task of travel-time prediction for an arbitrary origin-destination pair on a map. Unlike most of the existing studies, our method allows us to probabilistically predict the travel time along an unknown path if the similarity between paths is defined as a kernel function. Our first innovation is to use a string kernel to represent the similarity between paths. Our second new idea is to apply Gaussian process regression for probabilistic travel-time prediction.

Tsuyoshi Ide
IBM Research
Tokyo Research Lab.
goodidea@jp.ibm.com

Sei Kato
IBM Research, Tokyo Research Lab.
seikato@jp.ibm.com

CP11
Discretized Spatio-Temporal Scan Window

The focus of this paper is the discovery of anomalous spatio-temporal windows. We propose a Discretized Spatio- Temporal Scan Window approach to address the question of how we can treat Space and Time together without compromising on the properties of each and their impact on each other. In doing so we discover anomalous Spatio- Temporal windows, identify at what point in time the window changes, identify the spatial patterns of change over time and identify a spatial extent in time which is completely deviant with respect to the rest of the anomalous spatio- temporal windows. None of the current approaches address all these issues in combination. Subsequently we perform experiments on several real world datasets to validate our approach while comparing with the established approach of discovering a cylindrical spatio-temporal Scan window.

Vandana Janeja
Information Systems Department
University Of Maryland Baltimore County
vjaneja@umbc.edu

Seyed Mohammadi
UMBC, Johns Hopkins University
smohamm1@jhuadig.admin.jhu.edu

Aryya Gangopadhyay
University of Maryland, Baltimore County
gangopad@umbc.edu

CP11
Efficient Multiplicative Updates for Support Vector Machines

The dual formulation of the support vector machine (SVM) objective function is an instance of a nonnegative quadratic programming problem. We reformulate the SVM objective function as a matrix factorization problem which establishes a connection with the regularized nonnegative matrix factorization (NMF) problem. This allows us to derive a novel multiplicative algorithm for solving hard and soft margin SVM. The algorithm follows as a natural extension of the updates for NMF and semi-NMF. No additional parameter setting, such as choosing learning rate, is required. Exploiting the connection between SVM and NMF formulation, we show how NMF algorithms can be applied to the SVM problem. Multiplicative updates that we derive for SVM problem also represent novel updates for semi-NMF. Further this unified view yields algorithmic insights in both directions: we demonstrate that the Kernel Adatron algorithm for solving SVMs can be adapted to NMF problems. Experiments demonstrate rapid convergence to good classifiers. We analyze the rates of asymptotic convergence of the updates and establish tight bounds. We test them on several datasets using various kernels and report equivalent classification performance to that of a standard SVM.

Vamsi Potluru
Dept of Computer Science, University of New Mexico
ismav@cs.unm.edu

Sergey Plis
Department of Computer Science
University of New Mexico
s.m.plis@gmail.com

Morten Morup
Technical University of Denmark
morten.morup@gmail.com

Vince Calhoun
Electrical and Computer Engineering

University of New Mexico
vcalhoun@unm.edu

Terran Lane
Department of Computer Science
University of New Mexico
terran.lane@gmail.com

CP11

Finding Links and Initiators: a Graph-Reconstruction Problem

Consider a 0–1 observation matrix M, where rows correspond to entities and columns correspond to signals; a value of 1 (or 0) in cell (i, j) of M indicates that signal j has been observed (or not observed) in entity i. Given such a matrix we study the problem of inferring the underlying directed links between entities (rows) and finding which entries in the matrix are initiators. We formally define this problem and propose an MCMC framework for estimating the links and the initiators given the matrix of observations M. We also show how this framework can be extended to incorporate a temporal aspect; instead of considering a single observation matrix M we consider a sequence of observation matrices $M_1, \ldots, M_t$ over time. We show the connection between our problem and several problems studied in the field of social-network analysis. We apply our method to paleontological and ecological data and show that our algorithms work well in practice and give reasonable results.

Heikki Mannila
HIIT, Helsinki University of Technology
University of Helsinki
mannila@cs.helsinki.fi

Evimaria Terzi
IBM Almaden Research Center
eterzi@us.ibm.com

CP11

Efficient Active Learning with Boosting

We construct a novel objective function to unify semi-supervised learning and active learning boosting. Minimization of this objective is achieved through alternating optimization w.r.t the classifier ensemble and the queried data set iteratively. More important, we derive an efficient active learning algorithm under this framework, based on a novel query mechanism called query by incremental committee. It does not only save considerable computational cost, but also outperforms conventional active learning methods based on boosting.

Zheng Wang
Department of Automation, Tsinghua University
wangzhengthu@gmail.com

Yangqiu Song
Tsinghua University
songyq99@mails.tsinghua.edu.cn

Changshui Zhang
Tsinghua Univ.
zcs@mail.tsinghua.edu.cn

PP0

On Segment-Based Stream Modeling and Its Applications

The primary constraint in the effective mining of data streams is the large volume of data which must be processed in real time. In many cases, it is desirable to store a summary of the data stream segments in order to perform data mining tasks. Since density estimation provides a comprehensive overview of the probabilistic data distribution of a stream segment, it is a natural choice for this purpose. A direct use of density distributions can however turn out to be an inefficient storage and processing mechanism in practice. In this paper, we introduce the concept of cluster histograms, which provides an efficient way to estimate and summarize the most important data distribution profiles over different stream segments. These profiles can be constructed in a supervised or unsupervised way depending upon the nature of the underlying application. The profiles can also be used for change detection, anomaly detection, segmental nearest neighbor search, or supervised stream segment classification. The flexibility of the tasks which can be performed from the cluster histogram framework follows from its generality in storing the historical density profile of the data stream. As a result, this method provides a holistic framework for density based mining of data streams. We discuss and test the application of the cluster histogram framework to a variety of interesting data mining applications such as speaker recognition and intrusion detection.

Charu C. Aggarwal
IBM T. J. Watson Research Center
charu@us.ibm.com

PP0

Structure and Dynamics of Research Collaboration in Computer Science

We use the DBLP bibliographic database of Computer Science publications in top tier conferences to construct collaboration networks and examine the properties of these networks. We perform community structure analysis, examine various forms of centralization, and use PCA on the various areas of computer science research to compare and contrast their collaboration patterns. Our analysis examines the entire network, separate networks based on research area, and looks at how they have changed over time.

Christian A. Bird
University of California, Davis
cabird@ucdavis.edu

Earl Barr, Andre Nash, Premkumar Devanbu, Vladimir Filkov, Zhendong Su
UC Davis
etbarr@ucdavis.edu, alnash@ucdavis.edu, ptdevanbu@ucdavis.edu, vfilkov@ucdavis.edu, su@ucdavis.edu

PP0

On the Comparison of Relative Clustering Validity Criteria

The present paper presents an alternative methodology for comparing clustering validity criteria and uses it to make an extensive comparison of the performances of 4 well-known validity criteria and 20 variants of them over a collection of 142,560 partitions of 324 different data sets of a given class of interest.

Lucas Vendramin
Department of Computer Sciences
University of Sao Paulo at Sao Carlos
vendra@grad.icmc.usp.br

Ricardo J. Campello
Department of Computer Sciences
University of So Paulo at So Carlos
campello@icmc.usp.br

Eduardo Hruschka
Department of Computer Sciences
University of Sao Paulo at Sao Carlos
erh@icmc.usp.br

PP0

Context Aware Trace Clustering: Towards Improving Process Mining Results

Process Mining refers to the extraction of process models from event logs. Real-life processes tend to be less structured and more flexible. Traditional process mining algorithms have problems dealing with such unstructured processes and generate spaghetti-like process models that are hard to comprehend. An approach to overcome this is to cluster process instances such that each of the resulting clusters correspond to a coherent set of process instances that can be adequately represented by a process model. In this paper, we propose a context aware approach to trace clustering based on generic edit distance. It is well known that the generic edit distance framework is highly sensitive to the costs of edit operations. We define an automated approach to derive the costs of edit operations. The method proposed in this paper outperforms contemporary approaches to trace clustering in process mining. We evaluate the goodness of the formed clusters using established fitness and comprehensibility metrics defined in the context of process mining. The proposed approach is able to generate clusters such that the process models mined from the clustered traces show a high degree of fitness and comprehensibility when compared to contemporary approaches.

Jagadeesh Chandra Bose R.P
Department of Mathematics and Computer Science
University of Technology Eindhoven (TU/e), The Netherlands
j.c.b.rantham.prabhakara@tue.nl

Wil Van Der Aalst
Department of Mathematics and Computer Science
University of Technology, Eindhoven, The Netherlands
w.m.p.v.d.aalst@tm.tue.nl

PP0

A Semi-Supervised Framework for Feature Mapping and Multiclass Classification

We propose a semi-supervised framework incorporating feature mapping with multiclass classification. By learning multiple classification tasks simultaneously, this framework can learn the latent feature space effectively for both labeled and unlabeled data. The knowledge in the transformed space can be transferred not only between the labeled and unlabeled data, but also across multiple classes, so as to improve the classification performance given a small amount of labeled data. We show that this problem is equivalent to a sequential convex optimization problem by applying constraint concave-convex procedure (CCCP). Efficient algorithm with theoretical guarantee is proposed and computational issue is investigated. Extensive experiments have been conducted to demonstrate the effectiveness of our proposed framework.

Bo Chen, Wai Lam
Chinese University of Hong Kong
bchen@se.cuhk.edu.hk, wlam@se.cuhk.edu.hk

Ivor Tsang
Nanyang Technological University
ivortsang@ntu.edu.sg

Tak-Lam Wong
Chinese University of Hong Kong
wongtl@cse.cuhk.edu.hk

PP0

Divide and Conquer Strategies for Effective Information Retrieval

Latent Semantic Indexing, a well-known technique for information retrieval, requires the computation of a partial SVD of the term-document matrix. This computation becomes infeasible for large document collections, since it is very demanding both in time and memory. We discuss two divide and conquer strategies, with the goal of alleviating these difficulties. An additional benefit is that the computation can be easily adapted to a parallel computing environment. Experimental results confirm that the proposed strategies are effective.

Jie Chen
Department of Computer Science and Engineering
University of Minnesota
jchen@cs.umn.edu

Yousef Saad
Department of Computer Science
University of Minnesota
saad@cs.umn.edu

PP0

A Bayesian Approach to Graph Regression with Relevant Subgraph Selection

This paper introduces a Bayesian approach to graph regression problems requiring relevant subgraph selection which provides a posterior distribution on the target variable as opposed to a single estimate. The intractability issue arisen from the representation of the graphs as binary vectors of indicators of subgraphs is solved using a column generation approach, where the most violated constraints are found by weighted subgraph mining. The model is evaluated on several molecular graph datasets.

Silvia Chiappa
Max-Planck Institute for Biological Cybernetics
silvia.chiappa@tuebingen.mpg.de

Hiroto Saigo
Max-Planck Institute for Informatics
Campus E1 4, 66123 Saarbruecken, Germany
hiroto.saigo@mpi-inf.mpg.de

Koji Tsuda

Max Planck Institute for Biological Cybernetics
koji.tsuda@tuebingen.mpg.de

PP0
A New Constraint for Mining Sets in Sequences

Discovering interesting patterns in event sequences is a popular task in the field of data mining. Most existing methods try to do this based on some measure of cohesion to determine an occurrence of a pattern, and a frequency threshold to determine if the pattern occurs often enough. We introduce a new constraint based on a new interestingness measure combining the cohesion and the frequency of a pattern.

Boris Cule, Bart Goethals
University of Antwerp
boris.cule@ua.ac.be, bart.goethals@ua.ac.be

Celine Robardet
INSA-Lyon, LIRIS UMR5205, F-69621 Villeurbanne, France
celine.robardet@insa-lyon.fr

PP0
Non-Parametric Information-Theoretic Measures of One-Dimensional Distribution Functions from Continuous Time Series

We study non-parametric measures for the problem of comparing distributions, which arise in anomaly detection for continuous time series. Some of these measures are for PDFs and others are for CDFs. We show how to adapt PDF measures to compare CDFs —we compare 23 CDF measures. We provide a unified functional form for all measure. We determine the measure significance by simulations only. Finally, we evaluate them for the anomaly detection in continuous time series.

Paolo D'Alberto, Ali Dasdan
Yahoo! Inc.
pdalbert@yahoo-inc.com, dasdan@yahoo-inc.com

PP0
Noise Robust Classification Based On Spread Spectrum

In this paper we develop a robust classification mechanism based on a connectionist model in order to classify objects from arbitrary feature spaces. Our main contribution is to adapt the spread spectrum method from signal transmission technology to the noise-robust classification of feature vectors using a recurrent neural network. We applied our technique to four publicly available classification benchmarks, providing higher classification accuracies (2% to 16% improvement) than support vector machines and meta-classification techniques.

Joern David
Technical University Munich
david@in.tum.de

PP0
Finding Representative Association Rules from Large Rule Collections

One of the most well-studied problems in data mining is computing association rules from large transactional databases. Often, the rule collections extracted from existing data-mining methods can be far too large to be carefully examined and understood by the data analysts. In this paper, we address exactly this issue of overwhelmingly large rule collections by introducing and studying the following problem: Given a large collection $\mathcal{R}$ of association rules we want to pick a subset of them $S \subseteq \mathcal{R}$ that best represents the original collection $\mathcal{R}$ as well as the dataset from which $\mathcal{R}$ was extracted. We first quantify the notion of the goodness of a ruleset using two very simple and intuitive definitions. Based on these definitions we then formally define and study the corresponding optimization problems of picking the best ruleset $S \subseteq \mathcal{R}$. We propose algorithms for solving these problems and present experiments to show that our algorithms work well for real datasets and lead to large reduction in the size of the original rule collection.

Warren L. Davis
IBM Almaden Research Center
wldavis@us.ibm.com

Peter Schwarz
IBM
schwarz@almaden.ibm.com

Evimaria Terzi
IBM Almaden Research Center
eterzi@us.ibm.com

PP0
Discovery of Geospatial Discriminating Patterns from Remote Sensing Datasets

Large amounts of remotely sensed data calls for data mining techniques to fully utilize their rich information content. In this paper, a new value-iteration method is introduced to optimally split the spatial domain of the selected variable into two classes. This division is used to calculate the set of patterns that are emerging with respect to the two classes. A new method for a concise summarization is introduced to construct super patterns of controlling factors.

Wei Ding
UMass Boston
wei.ding@umb.edu

Tomasz Stepinski
Lunar and Planetary Institute
tom@lpi.usra.edu

Josue Salazar
University of Houston-Clear Lake
salazarj4857@uhcl.edu

PP0
Mining for Surprise Events Within Text Streams

Text streams are a fundamental source of information that can be used to detect and characterize strategic intent of individuals and organizations as well as detecting abrupt or surprising events. We describe our algorithm development and analysis methodology for mining the evolving content in text streams. Our approach focuses on the temporal characteristics in a text stream to identify relevant features, and on the analysis and algorithmic methodology to communicate these characteristics to a user.

Dave Engel, Paul Whitney, Nick Cramer

Pacific Northwest National Laboratory
dave.engel@pnl.gov, paul.whitney@pnl.gov,
nick.cramer@pnl.gov

PP0
Topic Evolution in a Stream of Documents

Document collections evolve over time, new topics emerge and old ones decline. At the same time, the terminology evolves as well. We propose Topic Monitor for monitoring and understanding of topic and vocabulary evolution over an infinite document stream. We use PLSA for topic modeling and propose new folding-in techniques for topic adaptation under an evolving vocabulary. We extract a series of models, on which we detect topic threads as descriptions of topic evolution.

André Gohr
Leibniz Institute of Plant Biochemistry,
IPB, Germany
agohr@ipb-halle.de

Alexander Hinneburg
Martin-Luther-University Halle-Wittenberg
hinneburg@informatik.uni-halle.de

René Schult, Myra Spiliopoulou
Otto-von-Guericke-University Magdeburg
Germany
schult@iti.cs.uni-magdeburg.de,
myra@iti.cs.uni-magdeburg.de

PP0
Randomization Techniques for Graphs

Within the framework of statistical hypothesis testing, we focus on randomization techniques for unweighted undirected graphs. Given an input graph, our randomization method will sample data from the class of graphs that share certain structural properties with the input graph. We present three alternative algorithms based on local edge swapping and Metropolis sampling. We test our framework in graph clustering and frequent subgraph mining.

Sami Hanhijärvi
Helsinki Institute for Information Technology HIIT
Helsinki University of Technology
sami.hanhijarvi@tkk.fi

Gemma Garriga, Kai Puolamäki
Helsinki Institute for Information Technology HIIT
gemma.garriga@tkk.fi, kai.puolamaki@tkk.fi

PP0
Musk: Uniform Sampling of k Maximal Patterns

We propose Musk, an algorithm to obtain representative frequent patterns by sampling uniformly from the pool of all maximal frequent patterns; uniformity is achieved by a variant of Markov Chain Monte Carlo (MCMC) algorithm. Musk simulates a random walk on the frequent pattern partial order graph with a prescribed transition probability matrix, whose values are computed locally during the simulation. In the stationary distribution of the random walk, all maximal frequent pattern nodes in the partial order graph are sampled uniformly. Experiments on various large datasets validate that Musk is effective in obtaining representative frequent patterns when complete enumeration of all the frequent patterns are infeasible by traditional algorithms.

Mohammad A. Hasan
Department of Computer Science
Rensselaer Polytechnic Institute
alhasan@cs.rpi.edu

Mohammed Zaki
Rensselaer Polytechnic Institute
zaki@cs.rpi.edu

PP0
Low-Entropy Set Selection

Most pattern discovery algorithms easily generate very large numbers of patterns, making the results impossible to understand and hard to use. In this paper we present a succinct way of representing data on the basis of itemsets that identify strong interactions. This new approach, LESS, provides a powerful and general MDL-based technique to data description. We consider the data symmetrically and describe all interactions between attributes, not just co-occurrences, in only a handful of sets.

Hannes Heikinheimo
Helsinki University of Technology TKK
Department of Information and Computer Science, HIIT
hannes.heikinheimo@tkk.fi

Jilles Vreeken
Department of Computer Science
Universiteit Utrecht
jillesv@cs.uu.nl

Arno Siebes
Dept. of Information and Computing Sciences
Universiteit Utrecht
arno@cs.uu.nl

Heikki Mannila
HIIT, Helsinki University of Technology
University of Helsinki
mannila@cs.helsinki.fi

PP0
A Re-Evaluation of the Over-Searching Phenomenon in Inductive Rule Learning

We evaluate the spectrum of different search strategies to see whether separate-and-conquer rule learners are able to gain performance by using more powerful search strategies like beam or exhaustive search. Unlike previous results that demonstrated that rule learners suffer from over-searching, our work pays particular attention to the connection between the search heuristic and the search strategy, and we show that for some heuristics, complex search algorithms will consistently improve results without suffering from over-searching.

Frederik Janssen
TU Darmstadt, Knowledge Engineering Group
janssen@ke.informatik.tu-darmstadt.de

Johannes Fürnkranz
TU Darmstadt
Knowledge Engineering Group
juffi@ke.informatik.tu-darmstadt.de

PP0

Change-Point Detection in Time-Series Data by Direct Density-Ratio Estimation

Change-point detection is the problem of discovering time points at which properties of time-series data change. This covers a broad range of real-world problems and has been actively discussed in the community of statistics and data mining. In this paper, we present a novel non-parametric approach to detecting the change of probability distributions of sequence data. Our key idea is to estimate the ratio of probability densities, not the probability densities themselves. This formulation allows us to avoid non-parametric density estimation, which is known to be a difficult problem. We provide a change-point detection algorithm based on direct density-ratio estimation that can be computed very efficiently in an online manner. The usefulness of the proposed method is demonstrated through experiments using artificial and real datasets.

Yoshinobu Kawahara, Masashi Sugiyama
Tokyo Institute of Technology
kawahara@sg.cs.titech.ac.jp, sugi@cs.titech.ac.jp

PP0

PICC Counting: Who Needs Joins when you Can Propagate Efficiently?

Counting is a common task in many data mining applications. In situations where the attributes of interest span multiple tables in databases, computing instance counts can be expensive. In this paper, we propose PICC, a technique for discovering instance counts. We propose a propagation-based instance counting scheme which avoids joins to obtain a single table. We then present a method for summarizing a database into a concise synopsis and thus estimating the required counts efficiently.

Jong Wook Kim, K. Selcuk Candan
Computer Science and Engineering Dept.
Arizona State University
jong@asu.edu, candan@asu.edu

PP0

Spatially Cost-Sensitive Active Learning

In active learning, one attempts to maximize classifier performance for a given number of labeled training points by allowing the active learning algorithm to choose which points should be labeled. Typically, when the active learner requests labels for the selected points, it assumes that all points require the same amount of effort to label and that the cost of labeling a point is independent of other selected points. In spatially distributed data such as hyperspectral imagery for land-cover classification, the act of labeling a point (i.e., determining the land-type) may involve physically traveling to a location and determining ground truth. In this case, both assumptions about label acquisition costs made by traditional active learning are broken, since costs will depend on physical locations and accessibility of all the visited points. This paper formulates and analyzes the novel problem of performing active learning on spatial data where label acquisition costs are proportional to distance traveled.

Alexander Liu, Goo Jun, Joydeep Ghosh
University of Texas at Austin
ayliu@mail.utexas.edu, gjun@ece.utexas.edu, ghosh@ece.utexas.edu

PP0

Highlighting Diverse Concepts in Documents

We show the underpinnings of a method for summarizing documents: it ingests a document and automatically highlights a small set of sentences that are expected to cover the different aspects of the document. The sentences are picked using simple coverage and orthogonality criteria. We describe a novel combinatorial formulation that captures exactly the document-summarization problem, and we develop simple and efficient algorithms for solving it. We compare our algorithms with many popular document-summarization techniques via a broad set of experiments on real data. The results demonstrate that our algorithms work well in practice and give high-quality summaries.

Kun Liu, Evimaria Terzi, Tyrone Grandison
IBM Almaden Research Center
kun@us.ibm.com, eterzi@us.ibm.com, tyroneg@us.ibm.com

PP0

Fedra: A Fast and Efficient Dimensionality Reduction Algorithm

Motivated by the problems occurring while mining data in high dimensional spaces we propose FEDRA, a fast and efficient dimensionality reduction algorithm that uses a set of landmark points to project data to a lower dimensional Euclidean space. FEDRA is faster and requires less memory than other comparable algorithms, without compromising the projection's quality. We theoretically assess the quality of the resulting projection and provide a bound for the error induced in pairwise distances.

Panagis Magdalinos, Christos Doulkeridis, Michalis‘ Vazirgiannis
Athens University of Economics and Business
pmagdal@aueb.gr, cdoulk@aueb.gr, mvazirg@aueb.gr

PP0

Mining Cohesive Patterns from Graphs with Feature Vectors

In this paper, we introduce the novel problem of mining cohesive patterns from graphs with feature vectors. A cohesive pattern is a dense and connected subgraph that has homogeneous values in a large enough feature subspace. We present the algorithm CoPaM which exploits various pruning strategies. Our theoretical analysis proves the correctness of CoPaM, and our experimental evaluation demonstrates its efficiency and effectiveness in driving applications such as social network analysis and molecular biology.

Flavia S. Moser
Simon Fraser University
fmoser@cs.sfu.ca

PP0

Exact Discovery of Time Series Motifs

Time series motifs are sets of very similar individual time series or subsequences of a long time series. Because of the quadratic search space, only *approximate* motifs have been found in the past. We designed a tractable algorithm (*MK*) to find *exact* motifs for the first time. Empirically, *MK* is way faster than brute-force search and applicable as a subroutine in high level data mining tasks like anytime classification, near-duplicate detection and summarization.

Abdullah Mueen, Eamonn Keogh, Qiang Zhu
University of California, Riverside
mueen@cs.ucr.edu, eamonn@cs.ucr.edu, qzhu@cs.ucr.edu

Sydne Cash
Massachusetts General Hospital
Harvard Medical School
scash@partners.org

Brandon Westover
Massachusetts General Hospital
Brigham and Women's Hospital
mwestover@partners.org

PP0
Providing Privacy Through Plausibly Deniable Search

Query-based web search is an integral part of many peoples daily activities. Most do not realize that their search history can be used to identify them (and their interests). In July 2006, AOL released an anonymized search query log of some 600K randomly selected users. While valuable as a research tool, the anonymization was insufficient: individuals were identified from the contents of the queries alone. Government requests for such logs increases the concern. To address this problem, we propose a client-centered approach of *plausibly deniable search*. Each user query is substituted with a standard, closely-related query intended to fetch the desired results. In addition, a set of k-1 cover queries are issued; these have characteristics similar to the standard query but on unrelated topics. The system ensures that any of these k queries will produce the same set of k queries, giving k possible topics the user could have been searching for. We use a Latent Semantic Indexing (LSI) based approach to generate queries, and evaluate on the DMOZ webpage collection to show effectiveness of the proposed approach.

Mummoorthy Murugesan
Department of Computer Science,Purdue University
mmuruges@cs.purdue.edu

Chris Clifton
Department of Computer Science
Purdue University
clifton@cs.purdue.edu

PP0
The Set Classification Problem and Solution Methods

This paper focuses on developing classification algorithms for problems in which there is a need to predict the class based on multiple observations (examples) of the same phenomenon (class). These problems give rise to a new classification problem, referred to as *set classification*, that requires the prediction of a set of instances given the prior knowledge that all the instances of the set belong to the same unknown class. This problem falls under the general class of problems whose instances have class label dependencies. Four methods for solving the set classification problem are developed and studied. The first is based on a straightforward extension of the traditional classification paradigm whereas the other three are designed to explicitly take into account the known dependencies among the instances of the unlabeled set during learning or classification. A comprehensive experimental evaluation of the various methods and their underlying parameters shows that some of them lead to significant gains in performance.

Xia Ning
University of Minnesota, Twin Cities
xning@cs.umn.edu

George Karypis
University of Minnesota / AHPCRC
karypis@cs.umn.edu

PP0
Text Categorization with All Substring Features

This paper presents a novel document classification method using all substrings as features. Learning by using all substrings has a prohibitive computational cost because the number of candidate substrings can be very large. We show that the idea of equivalent classes of substrings can help determine all effective substrings exhaustively in linear time. In experiments, we show that our method can extract effective substrings efficiently, and achieved more accurate results than the results using previous methods.

Daisuke Okanohara, Jun'ichi Tsujii
University of Tokyo
hillbig@is.s.u-tokyo.ac.jp, tsujii@is.s.u-tokyo.ac.jp

PP0
Exploiting Semantic Constraints for Estimating Supersenses with Crfs

The annotation of words by ontology concepts is extremely helpful for semantic interpretation. We employ conditional random fields to predict the coarse meanings (*supersenses*) of words. As the annotation of training data is costly we modify the CRF algorithm to process a set of possible labels for each training instance (*lumped labels*). Using only unlabelled data for training it turns out that the resulting F-value is only slightly lower than for the fully labelled data.

Frank Reichartz, Gerhard Paass
Fraunhofer IAIS
frank.reichartz@iais.fraunhofer.de,
gerhard.paass@iais.fraunhofer.de

PP0
Analyses for Service Interaction Networks with Applications to Service Delivery

In this work we focus on learning individual and team behavior of different people or agents of a service organization by studying the patterns and outcomes of historical interactions. We develop the notion of service interaction networks which is an abstraction of the historical data and allows one to cast practical problems in a formal setting. Towards this goal we develop new algorithms based on eigen value methods and an iterative approach.

Vinayaka Pandit, SAMEEP MEHTA, GYANA PARIJA
IBM India Research Lab
pvinayak@in.ibm.com, sameepmehta@in.ibm.com,
gyparija@in.ibm.com

S. KAMESHWARAN, VISWANADHAM N.

Indian School of Business (ISB)
kameshwaran_s@isb.edu, n_viswanadham@isb.edu

SUDHANSHU SINGH
University of North Carolina, Chapel Hill
sssingh@email.unc.edu

PP0
Measuring Discrimination in Socially-Sensitive Decision Records

We tackle the problem of determining, given a dataset of historical decision records, a precise measure of the degree of discrimination suffered by a given group of people. This problem is rephrased in a classification rule setting by introducing a collection of quantitative measures of discrimination. Based on these measures, we are able to unveil discriminatory decision patterns hidden in the historical data or in classifiers that learn over training data biased by discriminatory decisions.

Dino Pedreschi, Salvatore Ruggieri, Franco Turini
Dipartimento di Informatica, Universita' di Pisa
pedre@di.unipi.it, ruggieri@di.unipi.it, turini@di.unipi.it

PP0
A Hybrid Data Mining Metaheuristic for the P-Median Problem

The main contribution of this work is a hybrid version of the GRASP metaheuristic, which incorporates a data mining process, to solve the p-median problem. Patterns mined from a set of sub-optimal solutions are used to guide the GRASP procedures in the search for better solutions. Computational experiments, comparing traditional GRASP and different hybrid proposals showed that employing the mined patterns made the hybrid GRASP find better results in less computational time.

Alexandre Plastino, Erick Fonseca, Richard Fuchshuber, Simone Martins
Fluminense Federal University
plastino@ic.uff.br, efonseca@ic.uff.br, rfuchshuber@ic.uff.br, simone@ic.uff.br

Alex Freitas, Martino Luis, Said Salhi
University of Kent
a.a.freitas@kent.ac.uk, ml86@kent.ac.uk, s.salhi@kent.ac.uk

PP0
Aligned Graph Classification with Regularized Logistic Regression

We consider a classification problem in which there is a fixed and known binary relation defined on the features of a set of multivariate random variables, which we call an aligned graph classification problem. We aim to improve classification performance over conventional learning by incorporating feature relation information in the learning process through extending logistic regression to include the normalized Laplacian of the graph. We validate our method using simulated and real data sets.

Brian Quanz, Jun Huan
Information and Telecommunication Technology Center
University of Kansas
bquanz@ittc.ku.edu, jhuan@ku.edu

PP0
Feature Weighted SVMs Using Receiver Operating Characteristics

Support Vector Machines (SVMs) are a leading tool in classification and pattern recognition and the kernel function is one of its most important components. This function is used to map the input space into a high dimensional feature space. However, it can perform rather poorly when there are too many dimensions (e.g. for gene expression data) or when there is a lot of noise. In this paper, we investigate the suitability of using a new feature weighting scheme for SVM kernel functions, based on receiver operating characteristics (ROC). This strategy is clean, simple and surprisingly effective. We experimentally demonstrate that it can significantly and substantially boost classification performance, across a range of datasets.

Shaoyi Zhang, M. Maruf Hossain, Md. Rafiul Hassan, James Bailey, Kotagiri Ramamohanarao
Department of Computer Science and Software Engineering
The University of Melbourne, Australia
shaoyi@csse.unimelb.edu.au,
hossain@csse.unimelb.edu.au,
mrhassan@csse.unimelb.edu.au,
jbailey@csse.unimelb.edu.au, rao@csse.unimelb.edu.au

PP0
Identifying Information-Rich Subspace Trends in High-Dimensional Data

Identifying information-rich subsets in high-dimensional spaces and representing them as order revealing patterns is an important research problem in many science and engineering applications. In this paper, we seek an information-revealing representation of the data subsets and formalize the problem of identifying subspace trends focusing on information-rich subsets and develop a new algorithm to extract such subspace trends. We demonstrate our results on both synthetic and real-world datasets and show the advantages of the proposed methodology.

Chandan K. Reddy
Department of Computer Science
Wayne State University
chandankreddy@gmail.com

Snehal Pokharkar
Wayne State University
svpokharkar@gmail.com

PP0
On Maximum Coverage in the Streaming Model and Application To

The set-streaming model is the generalization of graph-streaming model to hyper-graphs. We consider the problem of maximum coverage, in which k sets have to be selected that maximize the total weight of the covered elements in this model and show that our algorithm achieves an approximation factor of 1 /4 . Using this algorithm, we provide efficient online solution to a multi-topic blog-watch application, an extension of blog-alert, for handling simultaneous multiple-topic requests.

Barna Saha
University of Maryland College Park
barna@cs.umd.edu

Lise Getoor
University of Maryland, College Park
getoor@cs.umd.edu

PP0
Multi-Field Correlated Topic Modeling

Popular methods for probabilistic topic modeling like Correlated Topic Models (CTM) and Latent Dirichlet Allocation (LDA) share an important property, i.e. using a common set of topics to model all the data. This can be too restrictive for modeling complex data entries consisting of multiple heterogeneous fields. We propose a new extension of the CTM method to enable modeling with multi-field topics in a global graphical structure.

Konstantin Salomatin, Yiming Yang
Carnegie Mellon University
ksalomat@cs.cmu.edu, yiming@cs.cmu.edu

Abhimanyu Lad
Language Technologies Institute
Carnegie Mellon University
alad@cs.cmu.edu

PP0
FutureRank: Ranking Scientific Articles by Predicting Their Future PageRank

The dynamic nature of citation networks makes the task of ranking scientific articles hard. We argue that what is most useful is the expected *future* references. We define a new measure, *FutureRank*, which is the expected future PageRank score based on citations that will be obtained in the future. In addition to making use of the citation network, FutureRank uses the authorship network and the publication time of the article in order to predict future citations.

Hassan Sayyadi
University of Maryland-College Park
sayyadi@cs.umd.edu

Lise Getoor
University of Maryland, College Park
getoor@cs.umd.edu

PP0
Diversity-Based Weighting Schemes for Clustering Ensembles

We propose general weighting schemes for clustering ensembles. These schemes are independent of the particular method of clustering ensembles and consider the individual clustering solutions in different ways, based on different implementations of the notion of diversity. We show how the proposed schemes can be instantiated into any instance-based, cluster-based and hybrid clustering ensembles methods. Experiments have shown that the performance of clustering ensembles algorithms increases when the proposed weighting schemes are employed.

Andrea Tagarelli
Dept. of Electronics, Computer and System Sciences
University of Calabria, Italy
tagarelli@deis.unical.it

Francesco Gullo
Department of Electronics, Computer and System Sciences
University of Calabria
fgullo@deis.unical.it

Sergio Greco
Dept. of Electronics, Computer and System Sciences
University of Calabria, Italy
greco@deis.unical.it

PP0
Detection and Characterization of Anomalies in Multivariate Time Series

This talk presents a robust algorithm for detecting anomalies in noisy multivariate time series data by employing a kernel matrix alignment method to capture the dependence relationships among variables in the time series. We show that the algorithm is flexible enough to handle different types of time series anomalies including subsequence-based and local anomalies. A case study is also presented to illustrate the ability of the algorithm to detect ecosystem disturbances in Earth science data.

Pang-Ning Tan, Haibin Cheng
Michigan State University
ptan@cse.msu.edu, chenghai@cse.msu.edu

Christopher Potter
NASA Ames Research Center
cpotter@mail.arc.nasa.gov

Steven Klooster
California State University
klooster@gaia.arc.nasa.gov

PP0
Tracking User Mobility to Detect Suspicious Behavior

Popularity of mobile devices is accompanied by widespread security problems, such as MAC address spoofing in wireless networks. We propose a probabilistic approach to temporal anomaly detection using smoothing technique for sparse data. Our technique builds up on the Markov chain, and clustering is presented for reduced storage requirements. Wireless networks suffer from oscillations between locations, which result in weaker statistical models. Our technique identifies such oscillations, resulting in higher accuracy. Experimental results on publicly available wireless network data sets indicate that our technique is more effective than Markov chain to detect anomalies for location, time, or both.

Gaurav Tandon
Nuance Communications
gtandon@fit.edu

Philip Chan
Deptt. of Computer Sciences, Florida Institute of Technology
150 W. University Blvd., Melbourne, FL 32901, USA
pkc@cs.fit.edu

PP0
ShatterPlots: Fast Tools for Mining Large Graphs

Graphs appear in several settings, like social networks, recommendation systems, among others. The main contribution of this paper is ShatterPlots, a simple, scalable (O(E)) and powerful algorithm to extract patterns from real graphs that help us spot synthetic graphs. The highlight patterns are: "30-per-cent", at the Shattering point all real and synthetic graphs have about 30% more nodes than edges; "NodeShatteringRatio", which can almost perfectly separate the real graphs from the synthetic.

Ana Paula Appel
So Paulo University
anaappel@icmc.usp.br

Deepayan Chakrabarti
Yahoo! Research
Sunnyvale, CA
deepay@yahoo-inc.com

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

Ravi Kumar
Yahoo! Research
ravikuma@yahoo-inc.com

Jure Leskovec
Cornell University
Computer Science Department
jure@cs.cornell.edu

Andrew Tomkins
Yahoo! Research
atomkins@yahoo-inc.com

Hanghang Tong
MLD SCS CMU
htong@cs.cmu.edu

PP0
Non-Negative Matrix Factorization, Convexity and Isometry

In this paper we explore avenues for improving the reliability of dimensionality reduction methods such as Non-Negative Matrix Factorization as interpretive exploratory data analysis tools. We first show for the first time that non-trivial NMF solutions always exist and that the optimization problem is actually convex, by using the theory of Completely Positive Factorization. We subsequently explore four novel approaches to finding globally-optimal NMF solutions using various ideas from convex optimization. We then develop a new method, isometric NMF (isoNMF), which preserves non-negativity while also providing an isometric embedding, simultaneously achieving two properties which are helpful for interpretation.

Nikolaos Vasiloglou
Georgia Institute of Technology
nvasil@ieee.org

PP0
Mining Complex Spatio-Temporal Sequence Patterns

Mining sequential movement patterns describing group behaviour in potentially streaming spatio-temporal data sets is a challenging problem. This work mines sequences of rules (k-STARs) that describe complex behaviours including spatio-temporal gaps and paths formed by 'replenishment'. The user may drill down and roll up on a lattice defined over the sequences for exploratory analysis. The algorithm runs linearly in the number of patterns mined and interesting sequences are found in a real world data set.

Florian Verhein
Institut für Informatik
Ludwig-Maximilians-Universität München, Germany
florian@verhein.com

PP0
An Entity Based Model for Coreference Resolution

Recently, many advanced machine learning approaches have been proposed for coreference resolution; however, all of the discriminatively-trained models reason over *mentions*, rather than *entities*. That is, they do not explicitly contain variables indicating the "canonical' values for each attribute of an entity (e.g., name, venue, title, etc.). This *canonicalization* step is typically implemented as a postprocessing routine to coreference resolution prior to adding the extracted entity to a database. In this paper, we propose a discriminatively-trained model that jointly performs coreference resolution and canonicalization, enabling features over hypothesized entities. We validate our approach on two different coreference problems: newswire anaphora resolution and research paper citation matching, demonstrating improvements in both tasks and achieving an error reduction of up to 62% when compared to a method that reasons about mentions only.

Michael Wick
University of Massachusetts
mwick@cs.umass.edu

PP0
Semi-Supervised Learning by Sparse Representation

The L1 graph proposed in this work is motivated by that each datum can be reconstructed by the sparse linear superposition of the training data. The sparse reconstruction coefficients, used to deduce the weights of the directed L1 graph, are derived by solving an L1 optimization problem on sparse representation. Then we propose a semi-supervised learning framework based on L1 graph to utilize both labeled and unlabeled data for inference on a graph.

Shuicheng Yan
National University of Singapore
eleyans@nus.edu.sg

Huan Wang
Chinese University of Hong Kong
hwang5@ie.cuhk.edu.hk

PP0
On Randomness Measures for Social Networks

In this paper, we theoretically analyze graph randomness and present a framework which provides a series of nonrandomness measures at levels of edge, node, and the overall graph. We show that graph non-randomness can be obtained mathematically from the spectra of the adjacency matrix of the network.

Xiaowei Ying, Xintao Wu
University of North Carolina at Charlotte
xying@uncc.edu, xwu@uncc.edu

PP0
Parallel Pairwise Clustering

We propose a simple strategy for pairwise clustering of massive data by randomly splitting their affinity matrix into small manageable affinity matrices that are clustered independently, for example using a parallel platform. We demonstrate that this approach yields high quality clustering for various real world problems, even though at each iteration only small fractions of the original data are examined and at no point is the entire affinity matrix stored in memory or even computed.

Elad Yom-Tov, Noam Slonim
IBM Haifa Research Lab
yomtov@il.ibm.com, noams@il.ibm.com

PP0
Speeding Up Secure Computations via Embedded Caching

High computation overheads of many cryptography-based Privacy Preserving Data Mining algorithms have rendered them less practical. In this paper, we address the efficiency issue of these algorithms by proposing a caching approach/concept. After carefully examining micro-steps of several secure computations blocks, we identify iterative portions and reduce their overall computational cost by caching intermediate results/data. We show empirically that the overall system efficiency would be greatly improved without affecting result quality or compromising data privacy.

Ke Zhai, Wee Keong Ng, Andre Herianto
School of Computer Engineering
Nanyang Technological University
zhai0005@ntu.edu.sg, wkn@acm.org,
andr0027@ntu.edu.sg

Shuguo Han
Nanyang Technological University
hans0004@ntu.edu.sg

PP0
Multiple Kernel Clustering

Maximum margin clustering (MMC) has recently attracted considerable interests in both the data mining and machine learning communities. As in other kernel methods, choosing a suitable kernel function is imperative to the success of maximum margin clustering. In this paper, we propose a multiple kernel clustering (MKC) algorithm that simultaneously finds the maximum margin hyperplane, the best cluster labeling, and the optimal kernel. Experimental results demonstrate the effectiveness and efficiency of the MKC algorithm.

Bin Zhao
Tsinghua Univ.
zhaobinhere@hotmail.com

James Kwok
Dept. Computer Science and Engineering
Hong Kong Univ. of Science and Technology
jamesk@cse.ust.hk

Changshui Zhang
Tsinghua Univ.
zcs@mail.tsinghua.edu.cn