CP0

Mining User Interactions for Searching the Web and Social Media

Abstract not available at time of publication.

Eugene Agichtein
Emory University
Mathematics & Computer Science Department
eugene@mathcs.emory.edu

CP0

Discriminant Analysis and Predictive Models in Medicine and Biology

Abstract not available at time of publication.

Eva K. Lee
Georgia Inst of Technology
Sch of Ind & Systems Eng
evakylee@isye.gatech.edu

CP0

Mining the Noise in Biomedical Data

Abstract not available at time of publication.

Brani Vidakovic
Georgia Institute of Technology
Department of Biomedical Engineering
brani@bme.gatech.edu

CP0

Sorting Out Metagenomes Through Identification of Genomic Signatures

Abstract not available at time of publication.

Ying Xu
University of Georgia
Department of Biochemistry and Molecular Biology
xyn@bmb.uga.edu

CP1

Creating a Cluster Hierarchy under Constraints of a Partially Known Hierarchy

Abstract not available at time of Publication.

Korinna Bade
Otto-von-Guericke-University Magdeburg
korinna.bade@ovgu.de

Andreas Nürnberger
Otto-von-Guericke University Magdeburg
andreas.nuernberger@ovgu.de

CP1

Data-Peeler: Constraint-Based Closed Pattern Mining in N-Ary Relations

Set pattern discovery from binary relations has been extensively studied during the last decade. In particular, many complete and efficient algorithms which extract frequent closed sets are now available. Generalizing such a task to n-ary relations ($n \geq 2$) appears as a timely challenge. It may be important for many applications, e.g., when adding the time dimension to the popular $objects \times features$ binary case. The generality of the — no assumption being made on the relation arity or on the size of its attribute domains — makes it computationally challenging. We introduce an algorithm called Data-Peeler. From a n-ary relation, it extracts all closed n-sets satisfying given piecewise (anti)-monotonic constraints. This new class of constraints generalizes both monotonic and anti-monotonic constraints. Considering the special case of ternary relations, Data-Peeler outperforms the state-of-the-art algorithms CubeMiner and Trias by orders of magnitude. These good performances must be granted to a new clever enumeration strategy allowing an efficient closeness checking. An original application on a real-life 4-ary relation is used to assess the relevancy of closed n-set constraint-based mining.

Jeremy Besson, Loic Cerf, Celine Robardet
INSA-Lyon, LIRIS UMR5205, F-69621 Villeurbanne, France
jeremy.besson@insa-lyon.fr, loic.cerf@insa-lyon.fr, celine.robardet@insa-lyon.fr

Jean-Francois Boulicaut
INSA-Lyon
LIRIS CNRS UMR5205
jean-francois.boulicaut@insa-lyon.fr

CP1

***SpaRClus*: Spatial Relationship Pattern-Based Hierarchical Clustering**

In this paper, we, first, show an algorithm, *SpIBag* (**Sp**atial **I**tem **Bag** Mining), which discovers frequent spatial patterns in images. Due to the properties of image data, *SpIBag* considers a bag of items together with a spatial information as a pattern which persists over geometrical transformations, such as scaling, translation, and rotation. Then, based on *SpIBag*, we propose *SpaRClus* (**Spa**tial **R**elationship Pattern-Based Hierarchical **Clus**tering) to cluster image data.

Sangkyum Kim, Xin Jin, Jiawei Han
University of Illinois at Urbana-Champaign
kim71@uiuc.edu, xinjin3@uiuc.edu, hanj@cs.uiuc.edu

CP1

Constrained Co-Clustering of Gene Expression Data

Co-clustering aims at computing a bi-partition that is a collection of co-clusters. We consider constrained co-clustering for extended must-link and cannot-link constraints (i.e., both objects and attributes can be involved) and for interval constraints that enforce properties of co-clusters when considering ordered domains. We propose an iterative co-clustering algorithm which exploits user-defined constraints while minimizing the sum-squared residues. We illustrate the added value of our approach in two applications on gene expression data.

Ruggero G. Pensa
Pisa KDD Laboratory
ISTI-CNR, Pisa, Italy
ruggero.pensa@isti.cnr.it

Jean-Francois Boulicaut
INSA-Lyon
LIRIS CNRS UMR5205

jean-francois.boulicaut@insa-lyon.fr

CP1

Semi-Supervised Clustering Via Matrix Factorizations

The recent years have witnessed a surge of interests of semi-supervised clustering methods, which aim to cluster the data set under the guidance of some supervisory information. Usually those supervisory information takes the form of pairwise constraints that indicate the similarity/dissimilarity between the two points. In this paper, we propose a novel matrix factorization based approach for semi-supervised clustering. In addition, we extend our algorithm to co-cluster the data sets of different types with constraints. Finally the experiments on UCI data sets and real world Bulletin Board Systems (BBS) data sets show the superiority of our proposed method.

Fei Wang
Department of Automation
Tsinghua University
feiwang03@gmail.com

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

Changshui Zhang
Tsinghua University
zcs@mail.thu.edu.cn

CP2

Mining Tree Patterns with Almost Smallest Supertrees

In this work we describe a new algorithm to mine tree structured data. Our method computes an almost smallest supertree, based upon iteratively employing tree alignment. This supertree is a global pattern, that can be used both for descriptive and predictive data mining tasks. Experiments performed on two real datasets, show that our approach leads to a drastic compression of the database. Furthermore, when the resulting pattern is used for classification, the results show a considerable improvement over existing algorithms. Moreover, the incremental nature of the algorithm provides a flexible way of dealing with extension or reduction of the original dataset. Finally, the computation of the almost smallest supertree can be easily parallelized.

Jeroen De Knijf
Universiteit Utrecht
jknijf@cs.uu.nl

CP2

Discovering Relational Item Sets Efficiently

Frequent item set mining is a major data mining research area. Generalising from the standard single table case to a multi-relational setting is simple in principle, but hard in practice. That is, it is simple to define frequent item sets in the multi-relational setting, as well as extending the A-Priori algorithm. It is hard, because the well-known frequent pattern explosion at low min-sup settings is far worse than it is in the standard case. In this paper we introduce an effective algorithm for the discovery of frequent, multi-relational item sets. These relational patterns show which item sets occur together. Answering questions like: 'What type of Books are bought together with what Record types?'. Hence, they provide a symmetric insight in the relation and reveal patterns that are relevant with respect to the relation. It extends our earlier work on using MDL to discover a small set of characteristic item sets. The algorithm, R-Krimp, first discovers the small set of characteristic patterns in the single tables and then combines these to find a small set of characteristic multi-relational item sets. This reduces the original search space dramatically and, hence, brings down the computational complexity by orders of magnitude. In the experiments we show that this approach yields a very good approximation of the naive approach, joining all tables into one huge table, while being far more efficient.

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

CP2

Mining Association Rules of Simple Conjunctive Queries

We present an algorithm for mining association rules in arbitrary relational databases. We define association rules over a simple, but appealing subclass of conjunctive queries, and show that many interesting patterns can be found. We propose an efficient algorithm and a database-oriented implementation in SQL, together with several promising and convincing experimental results.

Wim Le Page, Bart Goethals
University of Antwerp
wim.lepage@ua.ac.be, bart.goethals@ua.ac.be

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

CP2

Maximal Quasi-Bicliques with Balanced Noise Tolerance: Concepts and Co-Clustering Applications

The rigid all-versus-all adjacency required by a maximal biclique for its two vertex sets is extremely vulnerable to missing data. In the past, several types of *quasi-bicliques* have been proposed to tackle this problem, however their noise tolerance is usually unbalanced and can be very skewed. In this paper, we improve the noise tolerance of maximal quasi-bicliques by allowing every vertex to tolerate up to the same number, or the same percentage, of missing edges. This idea leads to a more natural interaction between the two vertex sets—a balanced most-versus-most adjacency. This generalization is also non-trivial, as many large-size maximal quasi-biclique subgraphs do not contain any maximal bicliques. This observation implies that direct expansion from maximal bicliques may not guarantee a complete enumeration of all maximal quasi-bicliques. We present important properties of maximal quasi-bicliques such as a bounded closure property and a fixed point property to design efficient algorithms. Maximal quasi-bicliques are closely related to co-clustering problems such as documents and words co-clustering, images and features co-clustering, stocks and financial ratios co-clustering, etc. Here, we demonstrate the usefulness of our concepts using a new application—a bioinformatics example—where

prediction of true protein interactions is investigated.

Jinyan Li
School of Computer Engineering
Nanyang Technological University
jyli@ntu.edu.sg

Kelvin Sim
Institute for Infocomm Research
shsim@i2r.a-star.edu.sg

Guimei Liu, Limsoon Wong
NUS
liugm@comp.nus.edu.sg, wongls@comp.nus.edu.sg

CP2

Cispan: Comprehensive Incremental Mining Algorithms of Closed Sequential Patterns for Multi-Versional Software Mining

Recently, frequent sequential pattern mining algorithms have been widely used in software engineering field to mine various source code or specification patterns. In practice, software evolves from one version to another in its life span. The effort of mining frequent sequential patterns across multiple versions of a software can be substantially reduced by efficient incremental mining. This problem is challenging in this domain since the databases are usually updated in all kinds of manners including insertion, various modifications as well as removal of sequences. Also, different mining tools may have various mining constraints, such as low minimum support. None of the existing work can be applied effectively due to various limitations of such work. For example, our recent work, IncSpan, failed solving the problem because it could neither handle low minimum support nor removal of sequences from database. In this paper, we propose a novel, comprehensive incremental mining algorithm for frequent sequential pattern, CISpan (Comprehensive Incremental Sequential Pattern mining). CISpan supports both closed and complete incremental frequent sequence mining, with all kinds of updates to the database. Compared to IncSpan, CISpan tolerates a wide range for minimum support threshold (as low as 2). Our performance study shows that in addition to handling more test cases on which IncSpan fails, CISpan outperforms IncSpan in all test cases which IncSpan could handle, including various sequence length, number of sequences, modification ratio, etc., with an average of 3.4 times speedup. We also tested CISpan's performance on databases transformed from 20 consecutive versions of Linux Kernel source code. On average, CISpan outperforms the non-incremental CloSpan by 42 times.

Ding Yuan
University of Illinois at Urbana-Champaign
dyuan3@cs.uiuc.edu

Kyuhyung Lee, Hong Cheng, Gopal Krishna
University of Illinois at Urbana-Champain
kyuhlee@cs.uiuc.edu, hcheng3@cs.uiuc.edu,
gkrishn2@cs.uiuc.edu

Zhenmin Li
CleanMake Inc.
zhenmin.li@cleanmake.com

Xiao Ma, Yuanyuan Zhou
University of Illinois at Urbana-Champain
xiaoma2@cs.uiuc.edu, yyzhou@cs.uiuc.edu

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

CP3

An Efficient Local Algorithm for Distributed Multivariate Regression in Peer-to-Peer Networks

This paper offers a **local** distributed algorithm for multivariate regression in large peer-to-peer environments. The algorithm is designed for distributed inferencing, data compaction, data modeling and classification tasks in many emerging peer-to-peer applications for bioinformatics, astronomy, social networking, sensor networks and web mining. Computing a global regression model from data available at the different peer-nodes using a traditional centralized algorithm for regression can be very costly and impractical because of the large number of data sources, the asynchronous nature of the peer-to-peer networks, and dynamic nature of the data/network. This paper proposes a two-step approach to deal with this problem. First, it offers an efficient local distributed algorithm that monitors the quality of the current regression model. If the model is outdated, it uses this algorithm as a feedback mechanism for rebuilding the model. The local nature of the monitoring algorithm guarantees low monitoring cost. Experimental results presented in this paper strongly support the theoretical claims.

Kanishka Bhaduri
Department of Computer Science,
University of Maryland Baltimore County
kanishk1@cs.umbc.edu

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

CP3

Semi-Supervised Learning Based on Semiparametric Regularization

Semi-supervised learning plays an important role in the recent literature on machine learning and data mining and the developed semi-supervised learning techniques have led to many data mining applications in recent years. This paper addresses the semi-supervised learning problem by developing a semiparametric regularization based approach, which attempts to discover the marginal distribution of the data to learn the parametric function through exploiting the geometric distribution of the data. This learned parametric function can then be incorporated into the supervised learning on the available labeled data as the prior knowledge. Specifically, our contributions are: (1) We present a semi-supervised learning approach which incorporates the unlabeled data into the supervised learning by a parametric function learned from the whole data including the labeled and unlabeled data. The parametric function reflects the geometric structure of the marginal distribution of the data. Furthermore, the proposed approach which naturally extends to the out-of-sample data is an inductive learning method in nature. (2) This approach allows a family of algorithms to be developed based on various choices of the original RKHS and the loss function. (3) We provide experimental comparisons showing that the proposed approach leads the state-of-the-art performance on a variety of classification tasks. In particular, we demonstrate that this approach can be used successfully in both

transductive and semi-supervised settings.

Zhen Guo, Zhongfei Zhang
Computer Science Department
State University of New York at Binghamton
zguo@cs.binghamton.edu, zhongfei@cs.binghamton.edu

Eric Xing
School of Computer Science
CMU
epxing@cs.cmu.edu

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

CP3

Roughly Balanced Bagging for Imbalanced Data

Imbalanced class problems appear in many real applications of classification learning. We propose a novel sampling method to improve bagging for data sets with skewed class distributions. In our new sampling method "Roughly Balanced Bagging" (RB Bagging), the number of samples in the largest and smallest classes are different, but they are effectively balanced when averaged over all subsets, which supports the approach of bagging in a more appropriate way. Our method is different from the existing bagging methods for imbalanced data which draw exactly the same numbers of majority and minority examples for the sampled subset data. In addition, our method makes full use of all of the minority examples by under-sampling, which is efficiently done by using negative binomial distributions. RB Bagging outperforms the existing "balanced" methods and other common methods, as shown by the experiments using benchmark and real-world data sets.

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

CP3

Aerosol Optical Depth Prediction from Satellite Observations by Multiple Instance Regression

Aerosol optical thickness prediction from satellite observations is considered to be one of the most important climate related research topics. Motivated by this application, a number of settings for multiple instance regression were studied. In our study, an EM based algorithm was proposed and was demonstrated to provide more accurate aerosol predictions as compared to the alternative methods. More accurate results were also obtained on synthetic datasets of various complexities.

Zhuang Wang, Vladan Radosavljevic, Bo Han, Zoran Obradovic, Slobodan Vucetic
Temple University
zhuang@temple.edu, vladan@ist.temple.edu, bohan@ist.temple.edu, zoran@ist.temple.edu, vucetic@ist.temple.edu

CP3

A Stagewise Least Square Loss Function for Classification

This paper presents a stagewise least square (SLS) loss function for classification. It uses a least square form to approximate a bounded monotonic nonconvex loss in a stagewise manner. Several benefits are obtained from using the SLS loss function, such as: (i) higher generalization accuracy and better scalability than classical least square loss; (ii) improved performance and robustness than convex loss; (iii) computational advantages compared with nonconvex loss; (iv) ability to boost the margin without boosting the classifier complexity. In addition, it naturally results in a kernel machine which is as sparse as SVM, yet much faster and simpler to train.

Shuang-Hong Yang, Bao-Gang Hu
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
eeshyang@gmail.com, hubg@nlpr.ia.ac.cn

CP4

Dynamic Non-Parametric Mixture Models and The Recurrent Chinese Restaurant Process : with Applications to Evolutionary Clustering

Clustering is an important data mining task for exploration and visualization of different data types like news stories, scientific publications, weblogs, etc. Due to the evolving nature of these data, evolutionary clustering, also known as dynamic clustering, has recently emerged to cope with the challenges of mining temporally smooth clusters over time. A good evolutionary clustering algorithm should be able to fit the data well at each time epoch, and at the same time results in a smooth cluster evolution that provides the data analyst with a coherent and easily interpretable model. In this paper we introduce the temporal Dirichlet process mixture model (TDPM) as a framework for evolutionary clustering. TDPM is a generalization of the DPM framework for clustering that automatically grows the number of clusters with the data. In our framework, the data is divided into epochs; all data points inside the same epoch are assumed to be fully exchangeable, whereas the temporal order is maintained across epochs. Moreover, The number of clusters in each epoch is unbounded: the clusters can retain, die out or emerge over time, and the actual parameterization of each cluster can also evolve over time in a Markovian fashion. We give a detailed and intuitive construction of this framework using the recurrent Chinese restaurant process (RCRP) metaphor, as well as a Gibbs sampling algorithm to carry out posterior inference in order to determine the optimal cluster evolution. We demonstrate our model over simulated data by using it to build an infinite dynamic mixture of Gaussian factors, and over real dataset by using it to build a simple non-parametric dynamic clustering-topic model and apply it to analyze the NIPS12 document collection.

Amr Ahmed
School of Computer Science
Carnegie Mellon University
amahmed@cs.cmu.edu

Eric Xing
School of Computer Science
CMU
epxing@cs.cmu.edu

CP4

Deterministic Latent Variable Models and Their Pitfalls

We derive a number of well known deterministic latent variable models such as PCA, ICA, EPCA, NMF and PLSA

as variational EM approximations with point posteriors. We show that the often practiced heuristic of "folding-in" can lead to overly optimistic estimates of the test-set log-likelihood and we verify this result experimentally. We trace this problem back to an infinitely negative entropy term that is ignored in the variational approximation.

Chaitanya Chemudugunta
Bren School of Information and Computer Science
chandra@ics.uci.edu

Max Welling, Nathan Sutter
University of California, Irvine
welling@ics.uci.edu, nsutter@uci.edu

CP4
Feature Selection with the LogRatio Kernel

In this article we present a novel kernel function, logRatio, which was designed to address two common problems in biological applications: data preprocessing and attribute interaction modelling. An extension of the SVMRFE feature selection algorithm was built around this new kernel function and compared with the original on a number of biological data and text classification problems. Experiments showed that SVMRFE based on the logRatio kernel detects relevant information and handles attribute redundancy more effectively than SVMRFE coupled with other kernels.

Julien Prados
CUI, Centre Universitaire d'Informatique
University of Geneva, Switzerland
julien.prados@cui.unige.ch

Alexandros Kalousis
Computer Science Department,
University of Geneva, Switzerland
kalousis@cui.unige.ch

Melanie Hilario
CUI, Centre Universitaire d'Informatique
University of Geneva, Switzerland
hilario@cui.unige.ch

CP4
Massive-Scale Kernel Discriminant Analysis: Mining for Quasars

We describe a fast algorithm for kernel discriminant analysis and its application to quasar identification in the Sloan Digital Sky Survey (40m points with four color dimensions). We empirically demonstrate asymptotic speed-up over the previous best approach and find approximately one million quasars, ten-fold more than the previous largest catalog. The algorithm achieves its speed-up using a new pattern of processing data, properties of the Epanechnikov kernel, and work-sharing between simultaneous computations for different bandwidths.

Ryan N. Riegel, Alexander Gray
Georgia Institute of Technology
rriegel@cc.gatech.edu, agray@cc.gatech.edu

Gordon Richards
Johns Hopkins University
gtr@physics.drexel.edu

CP4
A Relief Based Feature Extraction Algorithm

RELIEF is considered one of the most successful algorithms for assessing the quality of features due to its simplicity and effectiveness. It has been recently proved that RELIEF is an online algorithm that solves a convex optimization problem with a margin-based objective function. Starting from this mathematical interpretation, we propose a novel feature extraction algorithm, referred to as LFE, as a natural generalization of RELIEF. LFE collects discriminant information through local learning, and is solved as an eigenvalue decomposition problem with a closed-form solution. A fast implementation is also derived. Experiments on synthetic and real-world data are presented. The results demonstrate that LFE performs significantly better than other feature extraction algorithms in terms of both computational efficiency and accuracy.

Yijun Sun, Dapeng Wu
University of Florida
sunyijun@biotech.ufl.edu, wu@ece.ufl.edu

CP5
Similarity Measures for Categorical Data: A Comparative Evaluation

Measuring similarity or distance between two entities is a key step for several data mining and knowledge discovery tasks. The notion of similarity for continuous data is relatively well-understood, but for categorical data, the similarity computation is not straightforward. Several data-driven similarity measures have been proposed in the literature to compute the similarity between two categorical data instances but their relative performance has not been evaluated. In this paper we study the performance of a variety of similarity measures in the context of a specific data mining task: outlier detection. Results on a variety of data sets show that while no one measure dominates others for all types of problems, some measures are able to have consistently high performance.

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

CP5
Practical Private Computation and Zero-Knowledge Tools for Privacy-Preserving Distributed Data Mining

We introduce a practical scheme for privacy-preserving data mining. Our scheme is based on secret sharing over small field and enjoys very high efficiency. Verification of user data is via an extremely efficient zero-knowledge proof that uses a linear number of inexpensive small field operations, and only a logarithmic number of large-field cryptographic operations, achieving orders of magnitude reduction in running time over standard techniques (from hours to seconds) for large scale problems.

Yitao Duan, John Canny
University of California, Berkeley
duan@cs.berkeley.edu, jfc@cs.berkeley.edu

CP5
Latent Variable Mining with Its Applications to Anomalous Behavior Detection

We propose a new approach to anomaly detection by looking at the latent variable space to make the first step toward *latent anomaly detection*. We attempt to track changes of behavior patterns, which are information of more abstract level than observed elements such as UNIX command lines. The key ideas of the methods are: 1) construction of the model variation vectors, 2) the change-point detection for the time series of model variation vectors.

Shunsuke Hirose, Kenji Yamanishi
NEC
Common Platform Software Research Laboratories
s-hirose@ap.jp.nec.com, k-yamanishi@cw.jp.nec.com

CP5
A Spamicity Approach to Web Spam Detection

In this paper, we study the problem of unsupervised web spam detection. We introduce the notion of spamicity to measure how likely a page is spam. Spamicity is a more flexible and user-controllable measure than the traditional supervised classification methods. We propose efficient online link spam and term spam detection methods using spamicity. Our methods do not need training and are cost effective. A real data set is used to evaluate the effectiveness and the efficiency of our methods.

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

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

Zhaohui Tang
Microsoft AdLab
zhaotang@microsoft.com

CP5
Gaussian Process Learning for Cyber-Attack Early Warning

Network security information sharing systems share reports of cyberattacks such that a participant can be forewarned of the attacks observed by others. Because the number of reports is huge, identifying the attackers that are most relevant to each individual network becomes a challenging problem. We present a Gaussian process learning framework to address this problem. Our experiments using DShield data show that attackers found relevant are indeed more likely to attack in the future.

Jian Zhang, Phillip Porras
SRI International
jian.zhang@sri.com, porras@csl.sri.com

Johannes Ullrich
SANS Technology Institute
jullrich@sans.org

CP6
Integration of Multiple Networks for Robust Label Propagation

We address a label propagation problem on multiple networks and present a new algorithm that automatically integrates structure information brought in by multiple networks. The proposed method is robust in that irrelevant networks are automatically deemphasized, which is an advantage over existing approaches. We also show that the proposed algorithm can be interpreted as an EM algorithm with a Student-t prior. Finally, we demonstrate the usefulness of our method in protein function prediction.

Tsuyoshi Kato
Graduate School of Frontier Sciences
University of Tokyo
mailto:kato-tsuyoshi@cb.k.u-tokyo.ac.jp

Hisashi Kashima
Tokyo Research Laboratory
IBM Research
kashi_pong@yahoo.co.jp

Masashi Sugiyama
Tokyo Institute of Technology
sugi@cs.titech.ac.jp

CP6
Statistical Density Prediction in Traffic Networks

Recently, modern tracking methods started to allow capturing the position of massive numbers of moving objects. Given this information, it is possible to analyze and predict the traffic density in a network which offers valuable information for traffic control, congestion prediction and prevention. In this paper, we propose a novel statistical approach to predict the density on any edge of such a network at some time in the future. Our method is based on short-time observations of the traffic history. Therefore, knowing the destination of each travelling individual is not required. Instead, we assume that the individuals will act rationally and choose the shortest path from their starting points to their destinations. Based on this assumption, we introduce a statistical approach to describe the likelihood of any given individual in the network to be located at a certain position at a certain time. Since determining this likelihood is quite expensive when done in a straightforward way, we propose an efficient method to speed up the prediction which is based on a suffix-tree. In our experiments, we show the capability of our approach to make useful predictions about the traffic density and illustrate the efficiency of our new algorithm when calculating these predictions.

Matthias Renz, Matthias Schubert, Hans-Peter Kriegel, Andreas Zuefle
Ludwig-Maximilians University Munich
renz@dbs.ifi.lmu.de, schubert@dbs.ifi.lmu.de, kriegel@dbs.ifi.lmu.de, zuefle@dbs.ifi.lmu.de

CP6
Proximity Tracking on Time-Evolving Bipartite Graphs

Given an author-conference network that evolves over time, which are the conferences that a given author is most closely related with, and how do they change over time? Large time-evolving bipartite graphs appear in many set-

tings, such as social networks, co-citations, market-basket analysis, and collaborative filtering. Our goal is to monitor (i) the centrality of an individual node (e.g., who are the most important authors?); and (ii) the proximity of two nodes or sets of nodes (e.g., who are the most important authors with respect to a particular conference?) Moreover, we want to do this efficiently and incrementally, and to provide "any-time" answers. We propose pTrack and cTrack, which are based on random walk with restart, and use powerful matrix tools. Experiments on real data show that our methods are effective and efficient: the mining results agree with intuition; and we achieve up to 15 176 times speed-up, without any quality loss.

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

Spiros Papadimitriou
IBM T.J. Watson Lab
spapadim@us.ibm.com

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

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

CP6

Spatial Scan Statistics for Graph Clustering

We present a measure associated with detection and inference of statistically anomalous clusters of a graph based on the likelihood test of observed and expected edges in a subgraph. This measure is adapted from spatial scan statistics for point sets and provides quantitative assessment for clusters. We discuss some important properties of this statistic and its relation to modularity and Bregman divergences. We apply a simple clustering algorithm to find clusters with large values of this measure in a variety of real-world data sets, and we illustrate its ability to identify statistically significant clusters of selected granularity.

Bei Wang
Department of Computer Science
Duke University
beiwang@cs.duke.edu

Jeff Phillips
Duke University
jeffp@cs.duke.edu

Robert Schreiber, Dennis Wilkinson
HP Labs
rob.schreiber@hp.com, dennis.wilkinson@hp.com

Nina Mishra
Visiting Search Labs, Microsoft Resesarch
University of Virginia
nmishra@cs.virginia.edu

Robert Tarjan
Princeton University
HP Labs
robert.tarjan@hp.com

CP6

Randomizing Social Networks: a Spectrum Preserving Approach

We investigate how various topological properties and spectrum of social networks may be affected due to randomization. We conduct theoretical analysis on the extent to which edge anonymity can be achieved. A spectrum preserving graph randomization method, which can better preserve network properties while protecting edge anonymity, is then presented and empirically evaluated.

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

CP7

The Relevant-Set Correlation Model for Data Clustering

This paper introduces a model for clustering, the *Relevant-Set Correlation* (RSC) model, that requires no direct knowledge of the nature or representation of the data. The quality of cluster candidates, the degree of association between pairs of cluster candidates, and the degree of association between clusters and data items are all assessed according to the statistical significance of a form of correlation among pairs of relevant sets and/or candidate cluster sets.

Michael E. Houle
National Institute of Informatics (Japan)
meh@nii.ac.jp

CP7

Robust Clustering in Arbitrarily Oriented Subspaces

In this paper, we propose an efficient and effective method to find arbitrarily oriented subspace clusters by mapping the data space to a parameter space defining the set of possible arbitrarily oriented subspaces. The objective of a clustering algorithm based on this principle is to find those among all the possible subspaces, that accommodate many database objects. In contrast to existing approaches, our method can find subspace clusters of different dimensionality even if they are sparse or are intersected by other clusters within a noisy environment. A broad experimental evaluation demonstrates the robustness, efficiency and effectivity of our method.

Peer Kröger, Elke Achtert, Christian Böhm
Ludwig-Maximilians-Universität München
kroegerp@dbs.ifi.lmu.de, achtert@dbs.ifi.lmu.de,
boehm@dbs.ifi.lmu.de

Jörn David
Technische Universität München
david@in.tum.de

Arthur Zimek
Ludwig-Maximilians-Universität München
zimek@dbs.ifi.lmu.de

CP7

Weighted Consensus Clustering

Consensus clustering has emerged as an important ex-

tension of the classical clustering problem. We propose *weighted consensus clustering*, where each input clustering is weighted and the weights are determined in such a way that the final consensus clustering provides a better quality solution, in which clusters are better separated comparing to standard consensus clustering. Theoretically, we show that a reformulation of the well-known L_1 regularization LASSO problem is equivalent to the weight optimization of our weighted consensus clustering, and thus our approach provides sparse solutions which may resolve the difficult situation when the input clusterings diverge significantly. We also show that the weighted consensus clustering resolves the redundancy problem when many input clusterings correlate highly. Detailed algorithms are given. Experiments are carried out to demonstrate the effectiveness of the weighted consensus clustering.

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

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

CP7

Cluster Ensemble Selection

We define and study the ensemble selection problem for unsupervised learning. Given a large library of clustering solutions, our goal is to select a subset to form a smaller but better performing cluster ensemble. Focusing on quality and diversity, the two factors that has been shown to influence ensemble performance, we jointly consider both factors in our selection strategies. Empirical evaluation indicates that our methods can achieve statistically significant performance improvement over full ensembles.

Xiaoli Z. Fern, Wei Lin
Oregon State University
xfern@eecs.oregonstate.edu, linwe@eecs.oregonstate.edu

CP7

Efficient Maximum Margin Clustering Via Cutting Plane Algorithm

This paper presents a cutting plane algorithm for maximum margin clustering. The proposed algorithm constructs a nested sequence of successively tighter relaxations of the original MMC problem, and each optimization problem in this sequence could be efficiently solved using the constrained concave-convex procedure (CCCP). Experimental evaluations on several real world datasets show that CPMMC performs better than existing MMC methods, both in efficiency and accuracy.

Bin Zhao, Fei Wang, Changshui Zhang
Tsinghua Univ.
zhaobinhere@hotmail.com, feiwang03@mails.tsinghua.edu.cn, zcs@mail.tsinghua.edu.cn

CP8

Simultaneous Unsupervised Learning of Disparate Clusterings

In this paper, we address the difficult problem of uncovering disparate clusterings from the data in a *totally unsupervised manner*. We propose two new approaches for this problem. In the first approach we give a new and tractable characterization of decorrelation between clusterings, and present an objective function to capture it. In the second approach, we model the data as a sum of mixtures and associate each mixture with a clustering which leads us to the problem of learning a convolution of mixture distributions. We evaluate our methods on two real-world data sets - a music data set from the text mining domain, and a portrait data set from the computer vision domain. Our methods achieve a substantially higher accuracy than existing factorial learning as well as traditional clustering algorithms.

Prateek Jain, Raghu Meka, Inderjit S. Dhillon
University of Texas at Austin
pjain@cs.utexas.edu, raghu@cs.utexas.edu, inderjit@cs.utexas.edu

CP8

Unsupervised Segmentation of Conversational Transcripts

Contact center calls follow well defined patterns which structure the operational process of call handling. Automatically identifying such patterns in terms of distinct segments from a collection of call transcripts would improve productivity of agents and track compliance to guidelines. In this paper, we propose an algorithm to segment conversational transcripts in an unsupervised way, improve the segmentation using limited supervision and show that our algorithms performs well with respect to various evaluation measures.

Krishna Kummamuru, Deepak P
IBM India Research Lab, Bangalore
kkummamu@in.ibm.com, deepak.s.p@in.ibm.com

Shourya Roy, L Venkata Subramaniam
IBM India Research Lab, New Delhi
rshourya@in.ibm.com, lvsubram@in.ibm.com

CP8

A General Model for Multiple View Unsupervised Learning

Multiple view data, which have multiple representations from different feature spaces or graph spaces, arise in various data mining applications such as information retrieval, bioinformatics and social network analysis. Since different representations could have very different statistical properties, how to learn a consensus pattern from multiple representations is a challenging problem. In this paper, we propose a general model for multiple view unsupervised learning. The proposed model introduces the concept of mapping function to make the different patterns from different pattern spaces comparable and hence an optimal pattern can be learned from the multiple patterns of multiple representations. Under this model, we formulate two specific models for two important cases of unsupervised learning, clustering and spectral dimensionality reduction; we derive an iterating algorithm for multiple view clustering, and a simple algorithm providing a global optimum to multiple spectral dimensionality reduction. We also extend the proposed model and algorithms to evolutionary clustering and unsupervised learning with side information. Empirical evaluations on both synthetic and real data sets demonstrate the effectiveness of the proposed model and

algorithms.

Bo Long
SUNY at Binghamton
blong1@binghamton.edu

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

Zhongfei Zhang
Computer Science Department
State University of New York at Binghamton
zhongfei@cs.binghamton.edu

CP8
Large-Scale Many-Class Learning

We investigate algorithms for learning sparse feature-to-class indicies, to address the challenges of efficient learning in the presence of myriad classes. When compared to other approaches, including one-versus-rest and top-down methods using support vector machines, we find that indexing is highly advantageous in terms of space and time efficiency, at both training and classification times, while yielding similar and often better accuracies.

Omid Madani
SRI International, AI Center
madani@ai.sri.com

Michael Connor
UIUC, Computer Science
connor2@uiuc.edu

CP8
A General Framework for Estimating Similarity of Datasets and Decision Trees: Exploring Semantic Similarity of Trees

Decision trees are among the most popular pattern types in data mining due to their intuitive representation. However, little attention has been given on the definition of semantic similarity measures between decision trees. In this work, we present a general framework for similarity estimation that includes as special cases the estimation of semantic similarity between decision trees, as well as various forms of similarity estimation on classification datasets.

Irene C. Ntoutsi
Department of Informatics,University of Piraeus
ntoutsi@unipi.gr

Alexandros Kalousis
Computer Science Department,
University of Geneva, Switzerland
kalousis@cui.unige.ch

Yannis Theodoridis
Department of Informatics
University of Piraeus, Greece
ytheod@unipi.gr

PP0
Outlier Detection with Uncertain Data

In recent years, many new techniques have been developed for mining and managing uncertain data. This is because of the new ways of collecting data which has resulted in enormous amounts of inconsistent or missing data. Such data is often remodeled in the form of uncertain data. In this paper, we will examine the problem of outlier detection with uncertain data sets. The outlier detection problem is particularly challenging for the uncertain case, because the outlier-like behavior of a data point may be a result of the uncertainty added to the data point. Furthermore, the uncertainty added to the other data points may skew the overall data distribution in such a way that true outliers may be masked. Therefore, it is critical to be able to remove the effects of the uncertainty added both at the aggregate level as well as at the level of individual data points. In this paper, we will examine a density based approach to outlier detection, and show how to use it to remove the uncertainty from the underlying data. We present experimental results illustrating the effectiveness of the method.

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

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

PP0
On Indexing High Dimensional Data with Uncertainty

In this paper, we will examine the problem of distance function computation and indexing uncertain data in high dimensionality for nearest neighbor and range queries. Because of the inherent noise in uncertain data, traditional distance function measures such as the L_q-metric and their probabilistic variants are not qualitatively effective. This problem is further magnified by the sparsity issue in high dimensionality. In this paper, we examine methods of computing distance functions for high dimensional data which are qualitatively effective and friendly to the use of indexes. In this paper, we show how to construct an effective index structure in order to handle uncertain similarity and range queries in high dimensionality. Typical range queries in high dimensional space use only a subset of the ranges in order to resolve the queries. Furthermore, it is often desirable to run similarity queries with only a subset of the large number of dimensions. Such queries are difficult to resolve with traditional index structures which use the entire set of dimensions. We propose query-processing techniques which use effective search methods on the index in order to compute the final results. We discuss the experimental results on a number of real and synthetic data sets in terms of effectiveness and efficiency. We show that the proposed distance measures are not only more effective than traditional L_q-norms, but can also be computed more efficiently over our proposed index structure.

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

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

PP0
Semi-Supervised Learning of a Markovian Metric

In classification and clustering tasks metrics like Euclidean

do not necessarily reflect the true structure (clusters or manifolds) in the data, it becomes imperative that an appropriate metric be learned from training or labeled data. In this paper we present a Markov random walk based semi-supervised method for metric learning. A nearest neighbor graph representation of the data is created and a semidefinite program that learns the random walk on the corresponding graph is proposed.

Avleen Bijral
Dept. of Computer Science
University of Colorado Boulder
avleen.bijral@colorado.edu

Manuel Lladser
Dept. of Applied Mathematics
University of Colorado Boulder
manuel.lladser@colorado.edu

Gregory Grudic
Dept. of Computer Science
University of Colorado Boulder
gregory.grudic@colorado.edu

PP0

Semi-Supervised Multi-Label Learning by Solving a Sylvester Equation

We present a novel Semi-supervised algorithm for Multi-label learning by solving a *Sylvester Equation* (*SMSE*). Two graphs are first constructed on *instance* level and *category* level respectively. A regularization framework combining two regularization terms for the two graphs is suggested. We show that the labels of unlabeled data finally can be obtained by solving a *Sylvester Equation*.

Gang Chen, Yangqiu Song, Fei Wang
Department of Automation
Tsinghua University
g-c05@mails.thu.edu.cn, songyq99@mails.thu.edu.cn,
feiwang03@mails.thu.edu.cn

Changshui Zhang
Tsinghua University
zcs@mail.thu.edu.cn

PP0

The PageTrust Algorithm: How to Rank Web Pages When Negative Links Are Allowed?

The paper introduces a novel algorithm derived from the PageRank algorithm of Brin and Page. The PageRank algorithm interprets an hyperlink from page a to page b as being a positive vote from a to b. Starting from this interpretation, it attributes a rank to each page. However, it does not offer the possibility to take into account negative votes. The PageTrust algorithm includes negative links and converges to a trust value for each page. PageTrust appears as a natural extension of PageRank and it preserves good properties of robustness against possible spammers. Moreover several parameters allow us to strengthen or weaken the role played by negative links.

Cristobald De Kerchove, Paul Van Dooren
Department of mathematical engineering
Universite catholique de Louvain
c.dekerchove@uclouvain.be, vdooren@inma.ucl.ac.be

PP0

A Bayesian Technique for Estimating the Credibility of Question Answerers

We present a technique for ranking question answerers according to their credibility, characterized here by the probability that a given question answerer (user) will be awarded a *best answer* on a question given the answerer's question-answering history. This probability θ, is considered to be a hidden variable that can only be estimated statistically from the number b of best answers awarded, associated with the number n of questions answered.

Byron E. Dom, Deepa Paranjpe
Yahoo! Inc.
bdom@yahoo-inc.com, deepap@yahoo-inc.com

PP0

A Pattern Mining Approach Toward Discovering Generalized Sequence Signatures

Typically, sequence signatures, such as motifs and domains, are assumed to be localized in one region of a sequence or are derived as combinations of the former. We generalize the concept of sequence signatures and introduce an algorithm for efficiently determining signatures based on subsequences that may be located anywhere on a sequence. We evaluate our signatures in relation to those in the InterPro database and highlight the differences between them.

Dietmar Dorr
Department of Computer Science
North Dakotat State University
dietmar.dorr@ndsu.edu

Anne M. Denton
Department of Computer Science
North Dakota State University
anne.denton@ndsu.edu

PP0

Type Independent Correction of Sample Selection Bias Via Structural Discovery and Re-Balancing

Sample selection bias is a common problem in many applications, we propose to discover the natural structure of the target distribution, by which different types of sample selection bias can be observed and reduced by generating a new sample set from the structure. One main advantage of the approach is that it can correct all types of sample selection bias, while most of the previously proposed approaches are designed for some specific types of bias.

Xiaoxiao Shi, Jiangtao Ren
Computer Department of Sun Yat-sen University
xiao.x.shi@gmail.com, issrjt@mail.sysu.edu.cn

Wei Fan
IBM T.J.Watson Research,
weifan@us.ibm.com

Philip S. Yu
Department of Computer Science
University of Illinois at Chicago
psyu@cs.uic.edu

PP0

Theoretical Analysis of Subsequence Time-Series

Clustering from a Frequency-Analysis Viewpoint

We present; 1) a theoretical analysis of Subsequence Time Series (STS) clustering from a frequency-analysis viewpoint, and we identify a mathematical background on which STS clustering generates sine wave patterns. This also gives a novel theoretical analysis methodology for pattern discovery from time-series data, and 2) Phase Alignment STS clustering algorithm which employs a phase alignment preprocessing to avoid sine-wave patterns.

Ryohei Fujimaki
NEC
r-fujimaki@bx.jp.nec.com

Shunsuke Hirose, Takayuki Nakata
NEC
Common Platform Software Research Laboratories
s-hirose@ap.jp.nec.com, t-nakata@bk.jp.nec.com

PP0

Mining Abnormal Patterns from Heterogeneous Time-Series with Irrelevant Features for Fault Event Detection

We present an abnormal pattern mining algorithm, which characterizes a fault as an anomaly score vector, to resolve the issue of detecting fault events in the following realistic situation: A) the features to which multivariate time series correspond are heterogeneous; B) relative to a large number of normal examples, only a small number of examples of fault events are available in advance; and C) many features irrelevant to fault events are included.

Ryohei Fujimaki
NEC
r-fujimaki@bx.jp.nec.com

Takayuki Nakata
NEC
Common Platform Software Research Laboratories
t-nakata@bk.jp.nec.com

Hidenori Tsukahara, Akinori Sato
NEC
ITS Business Development Center
h-tsukahara@ab.jp.nec.com, a-satou@cw.jp.nec.com

Kenji Yamanishi
NEC
Common Platform Software Research Laboratories
k-yamanishi@cw.jp.nec.com

PP0

Learning Markov Network Structure Using Few Independence Tests

We present the Dynamic Grow-Shrink Inference-based Markov network learning algorithm (DGSIMN), a state-of-the-art algorithm for learning the structure of the domain Markov network from independence tests on data. DGSIMN dynamically selects the locally optimal test that is expected to increase the state of knowledge about the structure the most. Experiments show significant savings on the weighted number of tests on both sampled and real-world data while achieving similar or better accuracy in most cases.

Parichey Gandhi, Facundo Bromberg, Dimitris Margaritis
Iowa State University
parichey@cs.iastate.edu, bromberg@cs.iastate.edu, dmarg@cs.iastate.edu

PP0

Preemptive Measures Against Malicious Party in Privacy-Preserving Data Mining

In the paper, we study the problem of security violations when a malicious party provides false data. We identify four privacy vulnerabilities of secure scalar product protocols. We propose a general more model of two-party interaction and demonstrate its applicability to securely compute $(x_1 + y_1)(x_2 + y_2)$ and $(x + y)\log_2(x + y)$. We show how the proposed model can be used to securely compute four commonly used kernel functions. We also propose two necessary conditions and two basic measures.

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

Wee Keong Ng
Nanyang Technological University, Singapore
awkng@ntu.edu.sg

PP0

ROC-Tree: A Novel Decision Tree Induction Algorithm Based on Receiver Operating Characteristics to Classify Gene Expression Data

We introduce a new decision tree induction technique that is better suited for gene expression data. Our method is based on the area under the Receiver Operating Characteristics (ROC) curve, to help determine decision tree characteristics, such as node selection and stopping criteria. We experimentally compare our algorithm, called *ROC-tree*, against other well-known decision tree techniques. The experimental results clearly demonstrate that *ROC-tree* can deliver better classification accuracy in a range of gene expression data.

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

PP0

Exploration and Reduction of the Feature Space by Hierarchical Clustering

In this paper we propose the use of Ward hierarchical clustering for feature selection with a distance measure based on Goodman-Kruskal tau. It produces the feature subsets dendrogram, a valuable tool to study features relevance relationships. It is used to select the features by a wrapper method. We apply hierarchical clustering to many UCI data-sets and perform comparisons with other methods. Our method allows classifiers to generally outperform their corresponding ones without feature selection.

Dino Ienco
University of Torino
ienco@di.unito.it

Rosa Meo

University of Torino
Italy
meo@di.unito.it

PP0

The Asymmetric Approximate Anytime Join: A New Primitive with Applications to Data Mining

Abstract not available at time of publication.

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

PP0

Generic Methods for Multi-Criteria Evaluation

The concept of generic multi-criteria (MC) learning algorithm evaluation is investigated by comparing existing methods. These methods can be described as frameworks for integrating evaluation metrics and are generic in the sense that the metrics used are not dictated by the methods; the choice of metrics is instead problem dependent. We present a case study, in which we demonstrate how a new method, the candidate evaluation function (CEF), can be used to trade-off multiple criteria.

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

PP0

Efficiently Mining Closed Subsequences with Gap Constraints

Inspired by some state-of-the-art closed or constrained sequential pattern mining algorithms, the paper proposes an efficient approach to finding the complete set of closed sequential patterns with gap constraints. The approach combines the newly devised constrained pattern closure checking scheme and pruning techniques with the pattern growth based subsequence enumeration framework. Our extensive performance study shows that our approach is very efficient in mining frequent closed subsequences with gap constraints.

Chun Li
Department of Computer Science and Technology
Tsinghua University
socrates.lee@gmail.com

Jianyong Wang
Department of Computer Science and Technology
Tsinghua University
jianyong@tsinghua.edu.cn

PP0

Exact and Approximate Reverse Nearest Neighbor Search for Multimedia Data

Reverse nearest neighbor queries are useful in identifying objects that are of significant influence or importance. Existing methods either rely on pre-computation of nearest neighbor distances, do not scale well with high dimensionality, or do not produce exact solutions. In this work we motivate and investigate the problem of reverse nearest neighbor search on high dimensional, multimedia data. We propose exact and approximate algorithms that do not require pre-computation of nearest neighbor distances, and can potentially prune off most of the search space. We demonstrate the utility of reverse nearest neighbor search by showing how it can help improve the classification accuracy.

Jessica Lin, David Etter
George Mason University
jessica@ise.gmu.edu, detter@gmu.edu

David DeBarr
Microsoft Corporation
dave.debarr@microsoft.com

PP0

Finding a Haystack in Haystacks – Simultaneous Identification of Concepts in Large Bio-Medical Corpora

Automatically mining information from large text databases is a challenge due to slow pattern/string matching techniques. We introduce a new, fast multi-string pattern matching method called the Block Suffix Shifting (BSS) algorithm, which is based on the well known Aho-Chorasick algorithm. The advantages of our algorithm include: the ability to exploit the natural structure of text, perform significant character shifting, avoid useless backtracking jumps, efficient matching time and avoid the typical "sub-string" false positive errors.

Ying Liu
College of Information Sciences and Technology
The Pennsylvania State University
yliu@ist.psu.edu

Lucian Lita, Stefan Niculescu
Siemens Medical Solutions
lucian.lita@siemens.com, stefan.niculescu@siemens.com

Prasenjit Mitra, Lee Giles
College of Information Sciences and Technology
The Pennsylvania State University
pmitra@ist.psu.edu, giles@ist.psu.edu

PP0

Mining and Ranking Generators of Sequential Patterns

Sequential pattern mining first proposed by Agrawal and Srikant has received intensive research due to its wide range applicability in many real-life domains. Various improvements have been proposed which include mining a closed set of sequential patterns. Sequential patterns supported by the same sequences in the database can be considered as belonging to an equivalence class. Each equivalence class contains patterns partially-ordered by sub-sequence relationship and having the same support. Within an equivalence class, the set of maximal and minimal patterns are referred to as closed patterns and generators respectively. Generators used together with closed patterns can provide additional information which closed patterns alone are not able to provide. Also, as generators are the minimal members, they are preferable over closed patterns for model selection and classification based on the Minimum Description Length (MDL) principle. Several algorithms have been proposed for mining closed sequential patterns,

but none so far for mining sequential generators. This paper fills this research gap by investigating properties of sequential generators and proposing an algorithm to efficiently mine sequential generators. The algorithm works on a three-step process of search space compaction, non-generator pruning and a final filtering step. We also introduce ranking of mined generators and propose mining of a unique generator per equivalence class. Performance study has been conducted on various synthetic and real benchmark datasets. They show that mining generators can be as fast as mining closed patterns even at low support thresholds.

David Lo
Department of Computer Science
National University of Singapore
dlo@comp.nus.edu.sg

Siau-Cheng Khoo
National University of Singapore
khoosc@comp.nus.edu.sg

Jinyan Li
School of Computer Engineering
Nanyang Technological University
jyli@ntu.edu.sg

PP0

A New Method for Rule Finding Via Bootstrapped Confidence Intervals

Association rule discovery in large data sets is vulnerable to producing excessive false positives. Analytical results presented here indicate that Bonferonni-based solutions to this problem may have inherent limitations. The paper proposes a new approach to this problem, based on a novel use of the statistical bootstrap tool. The proposal here differs markedly from previous bootstrap/resampling approaches in basic goal, which is to enable much more active participation by domain experts.

Norman Matloff
University of California, Davis
matloff@cs.ucdavis.edu

PP0

Finding Subgroups Having Several Descriptions: Algorithms for Redescription Mining

In redescription mining, one aims at finding at least two different ways to describe the (approximately) same set of elements. We present two algorithms for this task. Our algorithms are based on heuristic pruning methods that prune the search space based on different interestingness and accuracy measures. We present experimental evaluation showing that the results of our algorithms are both significant and easy to interpret.

Pauli Miettinen
Helsinki Institute for Information Technology
University of Helsinki
Pauli.Miettinen@cs.Helsinki.FI

Arianna Gallo
University of Torino
gallo@di.unito.it

Heikki Mannila
Helsinki Institute for Information Technology
Helsinki University of Technology and University of Helsinki
heikki.mannila@cs.helsinki.fi

PP0

Randomization of Real-Valued Matrices for Assessing the Significance of Data Mining Results

Randomization is an important technique for assessing the significance of data mining results. In this paper, we study the problem of generating randomized real-valued matrices sharing the row and column means and variances with the original matrix. We describe three alternative algorithms based on local transformations and evaluate their performance on real and generated data. The results imply that the methods are usable in practice for significance testing of data mining results on real-valued matrices.

Markus Ojala
HIIT, Helsinki University of Technology
Department of Information and Computer Science
Markus.Ojala@tkk.fi

Niko Vuokko
Helsinki University of Technology
Department of Information and Computer Science
niko.vuokko@tkk.fi

Aleksi Kallio
The Finnish IT Center for Science
aleksi.kallio@helsinki.fi

Niina Haiminen
University of Helsinki
Helsinki Institute for Information Technology
niina.haiminen@cs.helsinki.fi

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

PP0

Clustering from Constraint Graphs

In constrained clustering it is common to model the pairwise constraints as edges on the graph of observations. Using results from graph theory, we analyze such constraint graphs in two contexts, both of immediate value to practitioners. First, we explore the issue of constraint noise under several intuitive noise models. We apply results from random graph theory, which facilitate the analysis of finite-sized graphs and realistic data partitions and noise levels, to obtain a quantification of the effect noisy edges may have on *any* constrained clustering algorithm under a set of commonly-used assumptions. We also demonstrate the dangers in the common practice of connected-component constraint set augmentation, when used in the presence of noise. Second, we describe two practical randomized algorithms that estimate the number of induced clusters using only a small number of constraints. We conclude with an experimental evaluation that shows the effect of noise on common UCI data sets, as well as some aspects of the behavior of our algorithms.

Dan Pelleg, Yossi Richter, Arie Freund
IBM Haifa Lab
daniel+siamconf@pelleg.org, richter@il.ibm.com,

arief@il.ibm.com

PP0

Spatio-Temporal Partitioning for Improving Aerosol Prediction Accuracy

In supervised learning, on data collected over space and time, an appropriate spatio-temporal data partitioning followed by building specialized predictors could often achieve higher overall prediction accuracy than when learning a single predictor on all the data. As an alternative to domain-based partitioning, we proposed a method that automatically discovers a spatio-temporal partitioning through the competition of regression models. The method was evaluated on a challenging problem using satellite observations to predict Aerosol Optical Depth.

Zoran Obradovic, Vladan Radosavljevic, Slobodan Vucetic
Temple University
zoran@ist.temple.edu, vladan@ist.temple.edu, vucetic@ist.temple.edu

PP0

On the Dangers of Cross-Validation. An Experimental Evaluation

Improvements in computational power and recent reductions in the (computational) cost of classification algorithms makes it possible to test a large number of variants of learning models on the data. We empirically show how under such large number of models the risk for overfitting increases and the performance estimated by cross validation is no longer an effective estimate of generalization; hence, this paper provides an empirical reminder of the dangers of cross validation.

Glenn M. Fung, Bharat Rao, Romer Rosales
Siemens Medical Solutions USA
glenn.fung@siemens.com, bharat.rao@siemens.com, romer.rosales@siemens.com

PP0

Efficient Distribution Mining and Classification

We define and solve the problem of "distribution classification", and, in general, "distribution mining". Given n distributions (i.e., clouds) of multi-dimensional points, we want to classify them into k classes, to find patterns, rules and out-lier clouds. For example, consider the 2-d case of sales of items, where, for each item sold, we record the unit price and quantity; then, each customer is represented as a distribution/cloud of 2-d points (one for each item he bought). We want to group similar users together, e.g., for market segmentation, anomaly/fraud detection. We propose D-Mine to achieve this goal. Our main contribution is Theorem 3.1, which shows how to use wavelets to speed up the cloud-similarity computations. Extensive experiments on both synthetic and real multi-dimensional data sets show that our method achieves up to *400 faster* wall-clock time over the naive implementation, with comparable (and occasionally better) classification quality.

Yasushi Sakurai
NTT Communication Science Laboratories
yasushi.sakurai@acm.org

Rosalynn Chong
University of British Columbia
rchong@interchange.ubc.ca

Lei Li, Christos Faloutsos
Carnegie Mellon University
leili@cs.cmu.edu, christos@cs.cmu.edu

PP0

Active Learning with Model Selection in Linear Regression

Optimally designing the location of training input points (active learning) and choosing the best model (model selection) are two important components of supervised learning and have been studied extensively. However, these two issues seem to have been investigated separately as two independent problems. If training input points and models are simultaneously optimized, the generalization performance would be further improved. In this paper, we propose a new approach called ensemble active learning for solving the problems of active learning and model selection at the same time. We demonstrate by numerical experiments that the proposed method compares favorably with alternative approaches such as iteratively performing active learning and model selection in a sequential manner.

Masashi Sugiyama, Neil Rubens
Tokyo Institute of Technology
sugi@cs.titech.ac.jp, neil@sg.cs.titech.ac.jp

PP0

A Feature Selection Algorithm Capable of Handling Extremely Large Data Dimensionality

With the advent of high throughput technologies, feature selection has become increasingly important in a wide range of scientific disciplines. We propose a new feature selection algorithm that performs extremely well in the presence of a huge number of irrelevant features. The key idea is to decompose an arbitrarily complex nonlinear models into a set of locally linear ones through local learning, and then estimate feature relevance globally within a large margin framework. The algorithm is capable of processing many thousands of features within a few minutes on a personal computer, yet maintains a close-to-optimum accuracy that is nearly insensitive to a growing number of irrelevant features. Experiments on eight synthetic and real-world datasets are presented that demonstrate the effectiveness of the algorithm.

Yijun Sun
University of Florida
sunyijun@biotech.ufl.edu

Sinisa Todorovic
University of Illinois at Urbana-Champaign
sintod@uiuc.edu

Steve Goodison
University of Florida
steve.goodison@jax.ufl.edu

PP0

Direct Density Ratio Estimation for Large-Scale Covariate Shift Adaptation

Covariate shift is a situation in supervised learning where training and test inputs follow different distributions. A

common approach is to reweight the training samples according to *importance*, which is the ratio of test and training densities. We propose a method to directly estimate the importance without density estimation. An advantage of the proposed method is that the computation time is nearly independent of the number of test input samples.

Yuta Tsuboi, Hisashi Kashima, Shohei Hido
IBM Research, Tokyo Research Laboratory
yutat@jp.ibm.com, hkashima@jp.ibm.com,
hido@jp.ibm.com

Steffen Bickel
Max Planck Institute for Computer Science
bickel@mpi-inf.mpg.de

Masashi Sugiyama
Tokyo Institute of Technology
sugi@cs.titech.ac.jp

PP0

Graph Mining with Variational Dirichlet Process Mixture Models

Graph data such as chemical compounds and XML documents are getting more common in many application domains. We propose a nonparametric Bayesian method for clustering graphs and selecting salient patterns at the same time. Variational inference is adopted here, because sampling is not applicable due to extremely high dimensionality.The feature set minimizing the free energy is efficiently collected with the DFS code tree.

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

Kenichi Kurihara
Tokyo Institute of Technology
kurihara@mi.cs.titech.ac.jp

PP0

Mining Complex, Maximal and Complete Sub-Graphs and Sets of Correlated Variables with Applications to Feature Subset Selection

This work considers the problem of mining 'complex', complete and maximal correlation graphs to describe the correlation structure between variables. It is proved that under a constraint on the minimum level of correlation desired, useful guarantees on the structure of such graphs exist. These results reduce the complexity of the problem and are exploited to develop an efficient data mining algorithm. Additionally, this approach is applied to feature subset selection, creating a new feature selection approach.

Florian Verhein
University of Sydney
fverhein@it.usyd.edu.au

PP0

A Range Query Approach for High Dimensional Euclidean Space Based on EDM Estimation

The need of efficient similarity queries in high dimensional space is rapidly increasing. The primary objective of this study is to propose a novel and efficient algorithm of the range query in high dimensional Euclidean space. The secondary objective is to propose the new principles and algorithms of *PSD* and *EDM estimation* to overcome the above key issue. A novel range query algorithm is developed by combining these techniques.

Takashi Washio
ISIR, Osaka University
washio@ar.sanken.osaka-u.ac.jp

Kentarou Kido, Hiroshi Kuwajima
The Institute of Scientific and Industrial Research
Osaka University
k-kido@ar.sanken.osaka-u.ac.jp,
kuwajima@ar.sanken.osaka-u.ac.jp

PP0

Mining Sequence Classifiers for Early Prediction

In many critical applications of sequence classification such as medical diagnosis and disaster prediction, early prediction is a highly desirable feature of sequence classifiers. In early prediction, a sequence classifier should use a prefix of a sequence as short as possible to make a reasonably accurate prediction. In this paper, we identify the novel problem of mining sequence classifiers for early prediction, and propose two methods to tackle this problem.

Zhengzheng Xing
Schoool of Computer Science, SFU
zxing@cs.sfu.ca

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

Guozhu Dong
Department of Computer Science and Engineering
Wright State University
gdong@cs.wright.edu

Philip S. Yu
Department of Computer Science
University of Illinois at Chicago
psyu@cs.uic.edu

PP0

Semi-Supervised Classification with Universum

The Universum data, defined as a collection of "non-examples" that do not belong to any class of interest, have been shown to encode some prior knowledge for classifiers. In this paper, we address a novel semi-supervised classification problem, called semi-supervised Universum, that can simultaneously utilize the labeled data, unlabeled data and the Universum data to improve the classification performance. The empirical experiments are presented to show the superior performance of the proposed method.

Dan Zhang
Department of Automation
Tsinghua University
danzhang2008@gmail.com

Jingdong Wang
Microsoft Research Asia
i-jingdw@microsoft.com

Fei Wang, Changshui Zhang
Tsinghua Univ.
feiwang03@mails.tsinghua.edu.cn,
zcs@mail.tsinghua.edu.cn

PP0

Exploiting Structured Reference Data for Unsupervised Text Segmentation with Conditional Random Fields

CRFs are a class of discriminative probabilistic models that are gaining acceptance as an effective computing machinery for text segmentation. An important aspect of CRFs is learning model parameters from manually labeled training data. One can avoid the labeling step by using structured reference tables. Inspired by recent work on their use for training HMMs, we developed an unsupervised technique for text segmentation with CRFs using reference tables.

Chang Zhao, JALAL Muhmad, I.V. Ramakrishnan
Stony Brook University
changz@cs.sunysb.edu, jmahmud@cs.sunysb.edu,
ram@cs.sunysb.edu

PP0

Semantic Smoothing for Bayesian Text Classification with Small Training Data

We propose a novel semantic smoothing method to address the data sparsity problem of NB classifier. Our method extracts explicit topic signatures (e.g. multiword phrases and ontology-based concepts) from documents and then statistically maps them into single-word features. When the size of training documents is small, the bayesian classifier with semantic smoothing not only outperforms the classifiers with background smoothing and Laplacian smoothing, but also beats the state-of-the-art active learning classifiers and SVM classifiers.

Xiaohua Zhou, Xiaodan Zhang, Xiaohua Hu
Drexel University
xiaohua.zhou@drexel.edu, xzhang@ischool.drexel.edu,
thu@ischool.drexel.edu