IP1

Exploring the Power of Heterogeneous Information Networks in Data Mining

Multiple typed objects in the real world are interconnected, forming complex heterogeneous information networks. Different from some studies on social network analysis where friendship networks or web page networks form homogeneous information networks, heterogeneous information network reflect complex and structured relationships among multiple typed objects. For example, in a university network, objects of multiple types, such as students, professors, courses, departments, and multiple typed relationships, such as teach and advise are intertwined together, providing rich information. We explore methodologies on mining such structured information networks and introduce several interesting new mining methodologies, including integrated ranking and clustering, classification, role discovery, data integration, data validation, and similarity search. We show that structured information networks are informative, and link analysis on such networks becomes powerful at uncovering critical knowledge hidden in large networks.

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

IP2

The Role of Data Mining in Business Optimization

In a trend that reflects the increasing demand for intelligent applications driven by business data, business enterprises are successfully building out a significant number of applications that leverage data mining technologies to optimize business process decisions. This talk highlights this trend; and describes the many different ways in which leading edge data mining, operations research, and information management concepts are being integrated and utilized in business applications.

Chid Apte
IBM Research
apte @us.ibm.com

IP3

Temporal Dynamics and Information Retrieval

Many digital resources, like the Web, are dynamic and ever-changing collections of information. However, most information retrieval tools developed for interacting with Web content, such as browsers and search engines, focus on a single static snapshot of the information. In this talk, I will present analyses of how Web content changes over time, how people re-visit Web pages over time, and how re-visitation patterns are influenced by changes in user intent and content. These results have implications for many aspects of information retrieval and management including crawling policy, ranking and information extraction algorithms, result presentation, and systems evaluation. I will describe a prototype that supports people in understanding how the information they interact with changes over time, and a new retrieval model that incorporates features about the temporal evolution of content to improve core ranking. Finally, I will conclude with an overview of some general challenges that need to be addressed to fully incorporate temporal dynamics in information retrieval and information management systems.

Susan Dumais
Microsoft Research
sdumais@microsoft.com

IP4

Text Mining Using Linear Models of Latent States

Models with hidden states offer an appealing framework for word disambiguation, parsing and other tasks in text mining. The vast repositories of data necessary for text modeling, however, overwhelm iterative, nonlinear algorithms. If we represent a word as an indicator vector, then the text of a single book produces a sequence of roughly 100,000 points in a 10,000 dimensional space. A small library produces on the order of a billion points in 100,000 dimensional space! Developments in the use of random matrix projections allow linear methods to scale to these dimensions,particularly when combined with classical multivariate decompositions that identify interesting, high-energy subspaces. Once projected, the data can be modeled using modern versions of linear methods, such as sequential regression. This use of regression also offers adjustments for complications such as those presented by transfer learning.

Robert A. Stine
University of Pennsylvania
stine@wharton.upenn.edu

CP1

ACE: Anomaly Clustering Ensemble for Multi-perspective Anomaly Detection in Robot Behaviors

This paper addresses an application of anomaly detection from subsequences of time series (STS) to autonomous robots' behaviors. An important aspect of mining sequential data is selecting the temporal parameters, such as the subsequence length and the degree of smoothing1. For example in the task at hand, the patterns of the robot's velocity, which is one of its fundamental features, vary significantly subject to the interval for measuring the displacement. Selecting the time scale and resolution is difficult in unsupervised settings, but often more critical than the choice of the method. In this paper, we propose an ensemble framework for aggregating anomaly detection from different perspectives, i.e., settings of user-defined, temporal parameters. In the proposed framework, each behavior is labeled whether it is an anomaly in multiple settings. The set of labels are used as meta-features of the respective behaviors. Cluster analysis in a meta-feature space partitions anomalous behaviors pertained to a specific range of parameters. The framework also includes a scalable implementation of the instance-based anomaly detection. We evaluate the proposed framework by ROC analysis, in comparison to conventional ensemble methods for anomaly detection.

Shin Ando
Guma University
shin.ando@acm.org

Einoshin Suzuki
Kyushu University
suzuki@inf.kyushu-u.ac.jp

Yoichi Seki, Theerasak Thanongphongphan
Gunma University

seki@cs.gunma-u.ac.jp, nices_jing@hotmail.com

Daisuke Hoshino
Daisuke Hoshino
gunma university

CP1
Fast Algorithms for Finding Extremal Sets

Identifying the extremal (minimal and maximal) sets from a collection of sets is an important subproblem in the areas of data mining and satis?ability checking. For example, extremal set ?nding algorithms are used in the context of mining maximal frequent itemsets, and for simplifying large propositional satis?ability instances derived from real world tasks such as circuit routing and veri?cation. In this paper, we describe two new algorithms for the task and detail their performance on real and synthetic data. Each algorithm leverages an entirely different principle one exploits primarily set cardinality constraints, the other lexicographic constraints. Despite the inherent dif?culty of this problem (the best known worst-case bounds are nearly quadratic), we show that both these algorithms provide excellent performance in practice, and can identify all extremal sets from multi-gigabyte itemset data using only a single processor core. Both algorithms are concise and can be implemented in no more than a few hundred lines of code. Our reference C++ implementations are open source and available for download.1

Roberto Bayardo, Biswanath Panda
Google
bayardo@alum.mit.edu, bpanda@google.com

CP1
The Network Completion Problem: Inferring Missing Nodes and Edges in Networks

We address the Network Completion Problem: Given a network with missing nodes and edges, can we complete the missing part of the network? By combining the Kronecker graphs model with the expectation-maximization, we design a scalable algorithm that estimates the model parameters as well as missing nodes and edges. Experiments on synthetic and real networks show that our approach effectively recovers the missing edges even when half of the nodes are missing.

Myunghwan Kim
Stanford University
mykim@stanford.edu

CP1
Interpreting and Unifying Outlier Scores

Outlier scores provided by different outlier models differ widely in their meaning, range, and contrast between different outlier models and, hence, are not easily comparable or interpretable. We propose a unification of outlier scores provided by various outlier models and a translation of the arbitrary "outlier factors' to values in the range $[0, 1]$ interpretable as values describing the probability of a data object of being an outlier. As an application, we show that this unification facilitates enhanced ensembles for outlier detection.

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

Peer Kroeger, Erich Schubert
LMU Munich
kroeger@dbs.ifi.lmu.de, schube@dbs.ifi.lmu.de

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

CP1
Structural Diversity for Privacy in Publishing Social Networks

Concerning the privacy in publishing social networks, we show that k-degree anonymity is not sufficient to provide protection against the vertex degree attacks, since vertices are usually associated with not only vertex identities but also communities. If the vertices of certain degree appear together in a dense sub-graph (community), an attacker can infer the neighborhood of a victim with his vertex degree without identifying the vertex of the victim. Therefore, we propose k-structural diversity anonymity.

Chih-Hua Tai
Dept of EE
National Taiwan University
hana@arbor.ee.ntu.edu.tw

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

DE-NIAN Yang
Institute of Information Science
Academia Sinica
dnyang@iis.sinica.edu.tw

MING-SYAN Chen
Dept of EE
National Taiwan University
mschen@cc.ee.ntu.edu.tw

CP2
A Gaussian Process Based Online Change Detection Algorithm for Monitoring Periodic Time Series

Online time series change detection is a critical component of many monitoring systems, such as space and air-borne remote sensing instruments, cardiac monitors, and network traffic profilers, which continuously analyze observations recorded by sensors. Data collected by such sensors typically has a periodic (seasonal) component. Most existing time series change detection methods are not directly applicable to handle such data, either because they are not designed to handle periodic time series or because they cannot operate in an online mode. We propose an online change detection algorithm which can handle periodic time series. The algorithm uses a *Gaussian process* based non-parametric time series prediction model and monitors the difference between the predictions and actual observations within a statistically principled control chart framework to identify changes. A key challenge in using Gaussian process in an online mode is the need to solve a large system of equations involving the associated covariance matrix which grows with every time step. The proposed algorithm ex-

ploits the special structure of the covariance matrix and can analyze a time series of length T in $O(T^2)$ time while maintaining a $O(T)$ memory footprint, compared to $O(T^4)$ time and $O(T^2)$ memory requirement of standard matrix manipulation methods. We experimentally demonstrate the superiority of the proposed algorithm over several existing time series change detection algorithms on a set of synthetic and real time series. Finally, we illustrate the effectiveness of the proposed algorithm for identifying land use land cover changes using *Normalized Difference Vegetation Index* (NDVI) data collected for an agricultural region in Iowa state, USA. Our algorithm is able to detect different types of changes in a NDVI validation data set (with $\approx 80\%$ accuracy) which occur due to crop type changes as well as disruptive changes (e.g., natural disasters).

Varun Chandola
Oak Ridge National Laboratory
chandolav@ornl.gov

CP2

Discovering Dynamic Dipoles in Climate Data

Pressure dipoles are important long distance climate phenomena (teleconnection) characterized by pressure anomalies of opposite polarity appearing at two different locations at the same time. Such dipoles have proven important for understanding and explaining the variability in climate in many regions of the world, e.g., the El Nino climate phenomenon is known to be responsible for precipitation and temperature anomalies worldwide. This paper presents a novel approach for dipole discovery that outperforms existing state of the art algorithms. Our approach is based on a climate anomaly network that is constructed using the correlation of time series of climate variables at all the locations on the Earth. One novel aspect of our approach to the analysis of such networks is a careful treatment of negative correlations, whose proper consideration is critical for finding dipoles. Another key insight provided by our work is the importance of modeling the time dependent patterns of the dipoles in order to better capture the impact of important climate phenomena on land. The results presented in this paper show that these innovations allow our approach to produce better results than previous approaches in terms of matching existing climate indices with high correlation and capturing the impact of climate indices on land.

Jaya Kawale, Michael Steinbach, Vipin Kumar
University of Minnesota
kawale@cs.umn.edu, steinbac@cs.umn.edu,
kumar@cs.umn.edu

CP2

Data Integration Via Constrained Clustering: An Application to Enzyme Clustering

When multiple data sources are available for clustering, an a priori data integration process is usually required, which may be costly and not lead to good clusterings because important information is likely to be discarded. We propose constrained clustering as a strategy for integrating data sources without losing any information, and apply it to the problem of enzyme function prediction. The use of the additional information improves clustering quality in an enzyme clustering application scenario.

Elisa B. Lima
Universidade Federal de Minas Gerais
eblima@dcc.ufmg.br

CP2

Integrating Distance Metrics Learned from Multiple Experts and Its Application in Patient Similarity Assessment

Patient similarity assessment is an important task in the context of patient cohort identification for comparative effectiveness studies and clinical decision support applications. The goal is to derive clinically meaningful distance metric to measure the similarity between patients represented by their key clinical indicators. It is desirable to learn the distance metric based on experts' knowledge of clinical similarity among subjects. However, often different physicians have different understandings of patient similarity based on the specifics of the cases. The distance metric learned for each individual physician often leads to a limited view of the true underlying distance metric. The key challenge will be how to integrate the individual distance metrics obtained for a group of physicians into a globally consistent unified metric. In this paper, we propose the Composite Distance Integration (Comdi) approach. In this approach we first construct discriminative neighborhoods from each individual metrics, then we combine them into a single optimal distance metric. We formulate Comdi as a quadratic optimization problem and propose an efficient alternating strategy to find the optimal solution. Besides learning a globally consistent metric, Comdi provides an elegant way to share knowledge acrossmultiple experts (physicians) without sharing the underlying data, which enables the privacy preserving collaboration. Our experiments on several benchmark data sets show approximately 10% improvement in classification accuracy over baseline. These results show that Comdi is an effective and general metric learning approach. An application of our approach to real patient data has also been presented in the results.

Fei Wang
Department of Statistical Science, Cornell University
feiwang03@gmail.com

Jimeng Sun
IBM Research
jimeng@us.ibm.com

Shahram Ebadollahi
IBM
ebad@us.ibm.com

CP2

Learning Feature Dependencies for Noise Correction in Biomedical Prediction

Noisy feature values in biomedical data can lead to incorrect prediction. We introduce a Bayesian Network-based Noise Correction framework, named BN-NC. BN-NC learns the dependencies between features, using them to deduce and compensate for any noise in the input during prediction. BN-NC also generates probabilistic rules to explain the class prediction and feature noise correction. Experimental results on HIV-1 drug resistance and leukemia subtype prediction demonstrate the efficacy of BN-NC.

Ghim-Eng Yap
Institute for Infocomm Research, A*STAR
geyap@i2r.a-star.edu.sg

Ah-Hwee Tan
Nanyang Technological University
asahtan@ntu.edu.sg

Hwee-Hwa Pang
Singapore Management University
hhpang@smu.edu.sg

CP3

Polonium: Tera-Scale Graph Mining and Inference for Malware Detection

We present *Polonium*, a novel Symantec malware detection technology, based on *Belief Propagation*, that infers every file's reputation, flagging the low-reputation ones as malware. We evaluated Polonium with a billion-node graph constructed from 60TB of data. Polonium attained a high *true positive rate* of 87% in detecting malware and lifted existing methods' by 10 *absolute* percentage points. Polonium has served 120 million people and answered *one trillion* queries.

Duen Horng Chau
Carnegie Mellon University
dchau@cs.cmu.edu

Carey Nachenberg, Jeffrey Wilhelm, Adam Wright
Symantec
cnachenberg@symantec.com,
jeffrey_wilhelm@symantec.com,
adam_wright@symantec.com

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

CP3

Clustered Low Rank Approximation of Graphs in Information Science Applications

In this paper we present a fast and accurate procedure called *clustered low rank matrix approximation* for massive graphs. The procedure involves a fast clustering of the graph and then approximates each cluster separately using existing methods, e.g. the singular value decomposition, or stochastic algorithms. The cluster-wise approximations are then extended to approximate the entire graph. This approach has several benefits: (1) important community structure of the graph is preserved due to the clustering; (2) highly accurate low rank approximations are achieved; (3) the procedure is efficient both in terms of computational speed and memory usage; (4) better performance in problems from various applications compared to standard low rank approximation. Further, we generalize stochastic algorithms to the clustered low rank approximation framework and present theoretical bounds for the approximation error. Finally, a set of experiments, using large scale and real-world graphs, show that our methods outperform standard low rank matrix approximation algorithms.

Berkant Savas, Inderjit S. Dhillon
University of Texas at Austin
berkant@cs.utexas.edu, inderjit@cs.utexas.edu

CP3

Centralities in Large Networks: Algorithms and Observations

Node centrality measures are important in a large number of graph applications. Various definitions for centrality have been proposed. However, measuring centrality in billion-scale graphs poses several challenges since the traditional definitions were not designed with scalability in mind. In this paper, we propose centrality measures suitable for very large graphs, as well as scalable methods to effectively compute them. We present extensive experimental results on both synthetic and real datasets.

U Kang
Carnegie Mellon University
Computer Science Department
ukang@cs.dot.cmu.dot.edu

Spiros Papadimitriou
Google
spapadim@gmail.com

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

Hanghang Tong
IBM T.J. Watson
htong@us.ibm.com

CP3

Kernel-Based Similarity Search in Massive Graph Databases with Wavelet Trees

Similarity search in databases of labeled graphs is a fundamental task in managing graph data such as XML, chemical compounds and social networks. Typically, a graph is decomposed to a set of substructures (e.g., paths, trees and subgraphs) and a similarity measure is defined via the number of common substructures. Using the representation, graphs can be stored in a document database by regarding graphs as documents and substructures as words. A graph similarity query then translates to a semi-conjunctive query that retrieves graphs sharing at least k substructures in common with the query graph. We argue that this kind of query cannot be solved efficiently by conventional inverted indexes, and develop a novel recursive search algorithm on wavelet trees (Grossi et al., SODA03). Unlike gIndex, it does not require frequent subgraph mining for indexing. In experiments, our method was successfully applied to 25 million chemical compounds.

Yasuo Tabei
JST ERATO Minato Project
yasuo.tabei@gmail.com

Koji Tsuda
Computational Biology Research Center, National
Institute of
Advanced Industrial Science and Technology, Tokyo,
Japan
koji.tsuda@aist.go.jp

CP3

Non-Negative Residual Matrix Factorization with Application to Graph Anomaly Detection

Given an IP source-destination traffic network, how do we spot mis-behavioral IP sources (e.g., port-scanner)? How do we find strange users in a user-movie rating graph?

Moreover, how can we present the results intuitively so that it is relatively easier for data analysts to interpret? We propose NrMF, a non-negative residual matrix factorization framework, to address such challenges. We present an optimization formulation as well as an effective algorithm to solve it. Our method can naturally capture abnormal behaviors on graphs. In addition, the proposed algorithm is linear wrt the size of the graph therefore it is suitable for large graphs. The experimental results on several data sets validate its effectiveness as well as efficiency.

Hanghang Tong
IBM T.J. Watson
htong@us.ibm.com

CP4
Mkboost: A Framework of Multiple Kernel Boosting

We investigate a boosting framework of exploring multiple kernel learning for classification. In particular, we present a novel framework of Multiple Kernel Boosting (MKBoost), which applies boosting techniques for learning kernel-based classifiers with multiple kernels. Based on the proposed framework, we develop several variants of MKBoost algorithms and examine their performance in comparisons to several state-of-the-art MKL algorithms. Experimental results show that our method is more effective and efficient than the existing MKL techniques.

Steven C.H. Hoi, Hao Xia
Nanyang Technological University
chhoi@ntu.edu.sg, xiah0002@e.ntu.edu.sg

CP4
Efficient Kernel Approximation for Large-Scale Support Vector Machine Classification

Training support vector machines (SVMs) with nonlinear kernel functions on large-scale data are usually very time-consuming. In contrast, there exist faster solvers to train the linear SVM. We propose a technique which sufficiently approximates the infinite-dimensional implicit feature mapping of the Gaussian kernel function by a low-dimensional feature mapping. By explicitly mapping data to the low-dimensional features, efficient linear SVM solvers can be applied to train the Gaussian kernel SVM, which leverages the efficiency of linear SVM solvers to train a nonlinear SVM. Experimental results show that the proposed technique is very efficient and achieves comparable classification accuracy to a normal nonlinear SVM solver.

Keng-Pei Lin
National Taiwan University
kplin@arbor.ee.ntu.edu.tw

MING-SYAN Chen
Dept of EE
National Taiwan University
mschen@cc.ee.ntu.edu.tw

CP4
A Quadratic Mean Based Supervised Learning Model for Managing Data Skewness

In this paper, We address the problem of class skewness for supervised learning models which are based on optimizing a regularized empirical risk function. These include both classification and regression models for discrete and continuous dependent variables. Classical empirical risk minimization is akin to minimizing the arithmetic mean of prediction errors, in which approach the induction process is biased towards the majority class for skewed data. To overcome this drawback, we propose a quadratic mean based learning framework (QMLearn) that is robust and insensitive to class skewness. We will note that minimizing the quadratic mean is a convex optimization problem and hence can be efficiently solved for large and high dimensional data. Comprehensive experiments demonstrate that the QMLearn model significantly outperforms existing statistical learners including logistic regression, support vector machines, linear regression, support vector regression and quantile regression etc.

Wei Liu, Sanjay Chawla
School of IT, the University of Sydney
weiliu.au@gmail.com, chawla@it.usyd.edu.au

CP4
A Sequential Dual Method for Structural Svms

We propose a fast sequential dual method(SDM) for structural SVMs. The method makes repeated passes over the training set and optimizes the dual variables associated with one example at a time.We present an extensive empirical evaluation of the proposed method on several sequence learning problems. Our experiments demonstrate that our method is an order of magnitude faster than state of the art methods and is thus a useful alternative for large-scale structured output learning.

Balamurugan P
Indian Institute of Science, Bangalore
balamurugan@csa.iisc.ernet.in

Shirish Shevade
Indian Institute Of Science
Bangalore
shirish@csa.iisc.ernet.in

Sundararajan S
Yahoo! Labs, Bangalore, India
ssrajan@yahoo-inc.com

Sathiya Keerthi S
Yahoo! Labs, Santa Clara, USA
selvarak@yahoo-inc.com

CP4
Who Should Label What? Instance Allocation in Multiple Expert Active Learning

The *active learning* (AL) framework is an increasingly popular strategy for reducing the amount of human labeling effort required to induce a predictive model. Most work in AL has assumed that a single, infallible oracle provides labels requested by the learner at a fixed cost. However, real-world applications suitable for AL often include multiple domain experts who provide labels of varying cost and quality. We explore this *multiple expert active learning* (MEAL) scenario and develop a novel algorithm for instance allocation that exploits the meta-cognitive abilities of novice (cheap) experts in order to make the best use of the experienced (expensive) annotators. We demonstrate that this strategy outperforms strong baseline approaches to MEAL on both a sentiment analysis dataset and two datasets from our motivating application of biomedical citation screening. Furthermore, we provide evidence that novice labelers are often aware of which instances they are likely to mislabel.

Byron C. Wallace
Department of Computer Science, Tufts University
ICHRPS, Tufts Medical Center
byron.wallace@gmail.com

Kevin Small, Carla Brodley
Department of Computer Science, Tufts University
kevin.small@gmail.com, brodley@cs.tufts.edu

Thomas Trikalinos
Institute for Clinical Research and Health Policy Studies
Tufts Medical Center
ttrikalinos@tuftsmedicalcenter.org

CP5
Segmented Nestedness in Binary Data

A binary matrix is *nested* if its columns can be ordered by the subset relation. In a *k-nested* matrix the data consists of k nested blocks. Such nested patterns occur in species occurrence data in ecology. The recognition of k-nestedness is polynomial in noiseless case; it becomes NP-hard if noise is present. Heuristic methods and an MDL-based technique choose k and discover a k-nested pattern in geographical mammal occurrence data.

Esa Junttila
University of Helsinki
esa.junttila@cs.helsinki.fi

Petteri Kaski
Aalto University and University of Helsinki
petteri.kaski@aalto.fi

CP5
Weighted Rank-One Binary Matrix Factorization

Rank-one binary matrix approximation has been actively studied recently for mining discrete patterns from binary data. However,this approach suffers from some limitations. We propose weighted rank-one binary matrix approximation. It enables the tradeoff between the accuracy and succinctness in approximate data and allows users to impose their personal preferences on the importance of different error types.

Haibing Lu, Jaideep Vaidya, Vijay Atluri, Heechang Shin
Rutgers University
luhaibingster@gmail.com, jsvaidya@rbs.rutgers.edu,
atluri@rutgers.edu, hshin@cimic.rutgers.edu

Lili Jiang
Lanzhou University
jianglili06@lzu.cn

CP5
Extracting Interpretable Features for Early Classification on Time Series

Early classification on time series data has been found highly useful in a few important applications, such as medical and health informatics, industry production management, safety and security management. While some classifiers have been proposed to achieve good earliness in classification, the interpretability of early classification remains largely an open problem. Without interpretable features, application domain experts such as medical doctors may be reluctant to adopt early classification. In this paper, we tackle the problem of extracting interpretable features on time series for early classification. Specifically, we advocate local shapelets as features, which are segments of time series remaining in the same space of the input data and thus are highly interpretable. We extract local shapelets distinctly manifesting a target class locally and early so that they are effective for early classification. Our experimental results on a few benchmark real data sets clearly shows that the local shapelets extracted by our methods are highly interpretable and can achieve effective early classification.

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

Jian Pei
Simon Fraser University;
jpei@cs.sfu.ca

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

Ke Wang
Simon Fraser University, Canada
wangk@cs.sfu.ca

CP5
Chernoff Dimensionality Reduction-Where Fisher Meets Fkt

Fisher's linear discriminant analysis (LDA) cannot handle heteroscedasticity in data. The Chernoff criterion for dimensionality reduction has been proposed. The technique extends Fisher's LDA and exploits information about heteroscedasticity in the data. While the Chernoff criterion outperforms the Fisher's, a clear understanding of its behavior is lacking. In this paper, we show what can be expected from the Chernoff criterion and its relations to Fisher and Fukunaga-Koontz transforme. We provide experiments validating our analysis.

Jing Peng
Montclair State University
pengj@mail.montclaire.edu

Guna Seetharaman
Information Directorate, AFMC AFRL/RITB
gunasekaran.seetharaman@rl.af.mil

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

Stefan Robila, Aparna Varde
Montclair State University
stefan.robila@montclair.edu, aparna.varde@montclair.edu

CP5
Feature-Based Inductive Transfer Learning Through Minimum Encoding

This paper proposes an Extended Minimum Description

Length Principle (EMDLP) for feature-based inductive transfer learning, in which both the source and the target data sets contain class labels and relevant features are transferred from the source domain to the target one by a code book. Our EMDLP has a solid theoretical framework, is parameter-free and allows us to evaluate the inferiority of the results of transfer learning.

Hao Shao, Einoshin Suzuki
Kyushu University
shaohao@i.kyushu-u.ac.jp, suzuki@inf.kyushu-u.ac.jp

CP6

A Generalized Framework for Mining Arbitrarily Positioned Overlapping Co-Clusters

Co-clustering aims to simultaneously cluster both rows and columns in a given matrix. Here, we present a new algorithm (PONEOCC) that uses a novel ranking-based function to find large and overlapping co-clusters with minimum errors. It allows positively and negatively correlated objects to be in the same co-cluster. The results on synthetic and gene expression datasets showed that PONEOCC is able to find biologically significant co-clusters and it outperforms some of the existing algorithms.

Omar Odibat, Chandan K. Reddy
Department of Computer Science
Wayne State University
odibat@wayne.edu, reddy@cs.wayne.edu

CP6

Sonar: Signal De-Mixing for Robust Correlation Clustering

Our algorithm SONAR is inspired by the idea of an active sonar that reveals hidden objects by sending echo pings: After probing the data with primitive pre-clusters serving as pings, independent component analysis (ICA) allows us to determine the statistically independent response patterns in the data. We combine this idea with the Minimum Description Length (MDL) principle for efficient, outlier-robust and parameter-free detection correlation of clusters.

Claudia Plant
Florida State University
cplant@fsu.edu

CP6

Spatially-Aware Comparison and Consensus for Clusterings

This paper proposes a new distance metric between clusterings that incorporates information about the spatial distribution of points and clusters. Our approach builds on the idea of a Hilbert space-based representation of clusters as a combination of the representations of their constituent points. We use this representation and the underlying metric to design a spatially-aware consensus clustering procedure. This consensus procedure is implemented via a novel reduction to Euclidean clustering, and is both simple and efficient. All of our results apply to both soft and hard clusterings. We accompany these algorithms with a detailed experimental evaluation that demonstrates the efficiency and quality of our techniques.

Parasaran Raman, Jeff Phillips, Suresh
Venkatasubramanian
University of Utah
praman@cs.utah.edu, jeffp@cs.utah.edu,
suresh@cs.utah.edu

CP6

Nonparametric Bayesian Co-clustering Ensembles

A nonparametric Bayesian approach to co-clustering ensembles is presented. Similar to clustering ensembles, co-clustering ensembles combine various base co-clustering results to obtain a more robust consensus co-clustering. To avoid pre-specifying the number of co-clusters, we specify independent Dirichlet process priors for the row and column clusters. Thus, the numbers of row- and column-clusters are unbounded *a priori*; the actual numbers of clusters can be learned *a posteriori* from observations. Next, to model non-independence of row- and column-clusters, we employ a Mondrian Process as a prior distribution over partitions of the data matrix. As a result, the co-clusters are not restricted to a regular grid partition, but form nested partitions with varying resolutions. The empirical evaluation demonstrates the effectiveness of nonparametric Bayesian co-clustering ensembles and their advantages over traditional co-clustering methods.

Pu Wang, Kathryn Laskey, Carlotta Domeniconi
George Mason University
pwang7@gmu.edu, klaskey@gmu.edu,
carlotta@cs.gmu.edu

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

CP6

Abacus: Mining Arbitrary Shaped Clusters from Large Datasets Based on Backbone Identification

We propose a shape-based clustering algorithm, ABACUS, that scales to large datasets. ABACUS is based on the idea of identifying the *intrinsic structure* for each cluster, which we also refer to as the *backbone* of that cluster. ABACUS operates in two stages. In the first stage, we identify the backbone of each cluster via an iterative process. The backbone enables easy identification of the true clusters in a subsequent stage. Experiments on real and synthetic datasets demonstrate the effectiveness of ABACUS.

Vineet Chaoji
Yahoo! Labs, Bangalore
chaojv@yahoo-inc.com

Geng Li, Hilmi Yildirim
Computer Science Department
Rensselaer Polytechnic Institute, Troy, NY, USA
lig2@cs.rpi.edu, yildih2@cs.rpi.edu

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

CP7

Towards Community Detection in Locally Heterogeneous Networks

In recent years, the size of many social networks such as *Facebook*, *MySpace*, and *LinkedIn* has exploded at a rapid pace, because of its convenience in using the internet in order to connect geographically disparate users. This has lead to considerable interest in many graph-theoretical aspects of social networks such as the underlying communities, the graph diameter, and other structural information which can be used in order to mine useful information from the social network. The graph structure of social networks is influenced by the underlying social behavior, which can vary considerably over different groups of individuals. One of the disadvantages of existing schemes is that they attempt to determine *global communities*, which (implicitly) assume uniform behavior over the network. This is not very well suited to the differences in the underlying density in different regions of the social network. As a result, a global analysis over social community structure can result in either very small communities (in sparse regions), or communities which are too large and incoherent (in dense regions). In order to handle the challenge of local heterogeneity, we will explore a simple property of social networks, which we refer to as the *local succinctness property.* We will use this property in order to extract compressed descriptions of the underlying community representation of the social network with the use of a min-hash approach. We will show that this approach creates balanced communities across a heterogeneous network in an effective way. We apply the approach to a variety of data sets, and illustrate its effectiveness over competing techniques.

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

Yan Xie, Philip Yu
University of Illinois at Chicago
yxie8@uic.edu, psyu@cs.uic.edu

CP7
Influence Maximization in Social Networks When Negative Opinions May Emerge and Propagate

We propose an influence cascade model where negative opinions may emerge and propagate due to product defects, and negativity bias is exhibited when both positive and negative opinions propagate in a social network. We show that influence spread in general networks may be sensitive to the product quality factor in the model. We develop a scalable algorithm for influence maximization for the model and experimentally demonstrate its efficiency and effectiveness.

Wei Chen
Microsoft Research Asia
weic@microsoft.com

CP7
Predicting Item Adoption Using Social Correlation

Users face a dazzling array of choices on the Web when it comes to choosing which product to buy, which video to watch, etc. The trend of social information processing means users increasingly rely not only on their own preferences, but also on friends when making various adoption decisions. In this paper, we investigate the effects of social correlation on users' adoption of items. Given a user-user social graph and an item-user adoption graph, we seek to answer the following questions: 1) whether the items adopted by a user correlate to items adopted by her friends, and 2) how to incorporate social correlation in order to improve prediction of unobserved item adoptions. We propose the *Social Correlation* model based on *Latent Dirichlet Allocation (LDA)* that decomposes the adoption graph into a set of latent factors reflecting user preferences, and a social correlation matrix reflecting the degree of correlation from one user to another. This matrix is learned (rather than pre-assigned), has probabilistic interpretation, and preserves the underlying social network structure. We further devise a *Hybrid* model that combines a user's own latent factors with her friends' for adoption prediction. Experiments on Epinions and LiveJournal data sets show that our proposed models outperform the approach based on latent factors only (LDA).

Freddy Chua
Singapore Management University
freddy.chua.2009@phdis.smu.edu.sg

CP7
On Node Classification in Dynamic Content-Based Networks

In recent years, a large amount of information has become available online in the form of web documents, social networks, blogs, or other kinds of social entities. Such networks are large, heterogeneous, and often contain a huge number of links. This linkage structure encodes rich structural information about the underlying topical behavior of the network. Such networks are often *dynamic* and evolve rapidly over time. Much of the work in the literature has focussed either on the problem of classification with purely text behavior, or on the problem of classification with purely the linkage behavior of the underlying graph. Furthermore, the work in the literature is mostly designed for the problem of static networks. However, a given network may be quite diverse, and the use of either content or structure could be more or less effective in different parts of the network. In this paper, we examine the problem of node classification in *dynamic* information networks with both text content and links. Our techniques use a random walk approach in conjunction with the content of the network in order to facilitate an effective classification process. This results in an effective approach which is more robust to variations in content and linkage structure. Our approach is dynamic, and can be applied to networks which are updated incrementally. Our results suggest that an approach which is based on a combination of content and links is extremely robust and effective. We present experimental results illustrating the effectiveness and efficiency of our approach.

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

Nan Li
University of California at Santa Barbara
nanli@umail.ucsb.edu

CP7
Cost of Collaboration Vs Individual Efforts in Social Networks

We study the dynamics of social networks in terms of population growth and control of user behavior. Most of the current research in social networks focus on static analysis through graph theoretic models to represent the networks or focus on modeling the traffic. Here, we study the cost of collaborative vs individualistic behavior of users in order to grow their network size in a social network. Each user incurs a cost (monetary or emotional) for collaboration. We formulate the behavior of the users as a non-linear optimization problem with a cost. The objective function of the optimization problem is obtained using a stochastic analysis of population growth in social networks, based on the first-passage time of a birth-death process. The stochastic model is validated by comparison with real data obtained from Twitter. Results indicate that a homogeneous social network (in which users have similar characteristics) will be individualistic. However, heterogeneous social networks (users with different characteristics) exhibit a threshold effect, i.e., there is a minimum cost, below which the network is as collabora- tive as desired and a maximum cost above which the network is individualistic as required. *To the best of our knowledge, this is one of the first analysis of dynamics of user behavior and temporal population growth in social networks.*

Anand Santhanakrishnan, Rajarathnam Chandramouli,
Kodavayur Subbalakshmi
Stevens Institute of Technology
anands72@gmail.com, mouli@stevens.edu,
ksubbala@stevens.edu

CP8

Block-Lda: Jointly Modeling Entity-Annotated Text and Entity-Entity Links

Identifying latent groups of entities from observed interactions between pairs of entities is a frequently encountered problem in areas like analysis of protein interactions and social networks. We present a model that combines aspects of mixed membership stochastic block models and topic models to improve entity-entity link modeling by jointly modeling links and text about the entities that are linked.

Ramnath Balasubramanyan, William Cohen
Carnegie Mellon University
rbalasub@cs.cmu.edu, wcohen@cs.cmu.edu

CP8

Distributed Monitoring of the R2 Statistic for Linear Regression

The problem of monitoring a multivariate linear regression model is relevant in studying the relationship between a set of input variables and dependent target variables. This problem becomes challenging for large scale data in a distributed computing environment when the local data changes frequently. In this paper we develop a distributed algorithm for monitoring the quality of a regression model. We show that our proposed method is highly scalable and eventually correct.

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

Kamalika Das
SGT Inc, NASA Ames Research Center
Moffett Field, CA 94035
kamalika.das@nasa.gov

Chris Giannella
The MITRE Corp, 7525 Colshire Drive
McLean, VA 22102-7539
cgiannella@mitre.org

CP8

A Tractable Pseudo-Likelihood Function for Bayes Nets Applied to Relational Data

Bayes nets (BNs) for relational databases are a major research topic in machine learning and artificial intelligence. When the database exhibits cyclic probabilistic dependencies, measuring the fit of a BN model to relational data with a likelihood function is a challenge. A common approach to difficulties in defining a likelihood function is to employ a pseudo-likelihood. This paper proposes a new pseudo likelihood P^* for Parametrized Bayes Nets (PBNs) [Poole 2003] and other relational versions of Bayes nets. The pseudo log-likelihood $L^* = ln(P^*)$ is similar to the single-table BN log-likelihood, where row counts in the data table are replaced by frequencies in the database. We introduce a new type of semantics based on the concept of random instantiations (groundings) from classic AI research [Halpern 1990, Bacchus 1990]: The measure L^* is the expected log-likelihood of a random instantiation of the 1st-order variables in the PBN. For parameter learning, the L^*-maximizing estimates are the empirical conditional frequencies in the databases. For structure learning, we show that the state of the art learn-and-join method of Khosravi *et al.* [AAAI 2010] implicitly maximizes the L^* measure. The measure provides a theoretical foundation for this algorithm, while the algorithm's empirical success provides experimental validation for its usefulness.

Oliver N. Schulte
School of Computing Science
Simon Fraser University
oschulte@cs.sfu.ca

CP8

Are Your Items in Order?

Items in many datasets can be arranged to an order. Such orders can provide new knowledge about the data and may ease further data exploration. Our goal is to define a statistically well-founded measure for the quality of an order. We achieve this by fitting an order-sensitive model to the dataset and using the BIC score as the quality measure. For computing the measure we introduce a fast dynamic program.

Nikolaj Tatti
Adrem, University of Antwerp,
Antwerpen, Belgium
nikolaj.tatti@gmail.com

CP8

Probabilistic Models over Ordered Partitions with Applications in Document Ranking and Collaborative Filtering

Ranking is an important task for handling a large amount of content. Ideally, training data for supervised ranking would include a complete rank of documents (or other objects such as images or videos) for a particular query. However, this is only possible for small sets of documents. In practice, one often resorts to document rating, in that a subset of documents is assigned with a small number indicating the degree of relevance. This poses a general problem of modelling and learning rank data with ties. In this paper, we propose a probabilistic generative model, that models the process as permutations over partitions. This results in super-exponential combinatorial state space with unknown numbers of partitions and unknown ordering

among them. We approach the problem from the discrete choice theory, where subsets are chosen in a stagewise manner, reducing the state space per each stage significantly. Further, we show that with suitable parameterisation, we can still learn the models in linear time. We evaluate the proposed models on two application areas: (i) document ranking with the data from the recently held Yahoo! challenge, and (ii) collaborative filtering with movie data. The results demonstrate that the models are competitive against well-known rivals.

Truyen Tran, Dinh Phung, Svetha Venkatesh
Curtin University
t.tran2@curtin.edu.au, d.phung@curtin.edu.au, s.venkatesh@curtin.edu.au

CP9
On Flow Authority Discovery in Social Networks

A central characteristic of social networks is that it facilitates rapid dissemination of information between large groups of individuals. This paper will examine the problem of determination of *information flow representatives*, a small group of authoritative representatives to whom the dissemination of a piece of information leads to the maximum spread. Clearly, information flow is affected by a number of different structural factors such as the node degree, connectivity, intensity of information flow interaction and the global structural behavior of the underlying network. We will propose a stochastic information flow model, and use it to determine the authoritative representatives in the underlying social network. We will first design an accurate *RankedReplace* algorithm, and then use a Bayes probabilistic model in order to approximate the effectiveness of this algorithm with the use of a fast algorithm. We will examine the results on a number of real social network data sets, and show that the method is more effective than state-of-the-art methods.

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

Arijit Khan
University of California at Santa Barbara
arijitkhan@cs.ucsb.edu

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

CP9
Exploiting Coherence in Reviews for Discovering Latent Facets and Associated Sentiments

Facet-based sentiment analysis of customer reviews involves discovering the latent facets, sentiments and their associations. Traditional facet-based sentiment analysis algorithms typically perform the various tasks in sequence, and fail to take advantage of the mutual reinforcement of the tasks. In this talk, we present a series of probabilistic models that jointly discover latent facets and sentiment topics, and also order the sentiment topics with respect to a multi-point scale, in a language and domain independent manner. This is achieved by simultaneously capturing both short-range syntactic structure and long range semantic dependencies between the sentiment and facet words. The models further incorporate *coherence* in reviews, where reviewers dwell on one facet or sentiment level before moving on, for more accurate facet and sentiment discovery. Lastly, we discuss the extensive experimention carried out on real world review data in order to demonstrate the efficacy of the proposed models.

Himabindu Lakkaraju
IBM Research - India,
Manyata Tech Park, Outer Ring Road, Bangalore - 560045
klakkara@in.ibm.com

Chiranjib Bhattacharyya, Indrajit Bhattacharya
Indian Institute of Science, Bangalore
India - 560012
chiru@csa.iisc.ernet.in, indrajit@csa.iisc.ernet.in

Srujana Merugu
IBM Research - India,
4 Block C, Vasant Kunj, New Delhi, India - 110070
srujanamerugu@in.ibm.com

CP9
Graph-Based Marginal Ranking for Update Summarization

In this paper, we propose a graph-based regularization framework MarginRank for update summarization. MarginRank extends the cost function of Zhou's Manifold Ranking with suppression terms, suppression of the previous documents on the current ones, to fulfil the assumption that users have read previous documents in update summarization. Experiments on the benchmark data sets TAC 2008 and 2009 show the effectiveness of the proposed method.

Xuan Li, Liang Du, Yi-Dong Shen
State Key Lab of Computer Science
Institute of Software, Chinese Academy of Sciences
islxuan@gmail.com, duliang@ios.ac.cn, ydshen@ios.ac.cn

CP9
Semi-Supervised Convolution Graph Kernels for Relation Extraction

We propose a novel Semi-supervised Convolution Graph Kernel (SCGK) method for semantic Relation Extraction (RE) from text. By encoding sentences as dependency graphs of words, SCGK computes kernels between sentences using a convolution strategy. SCGK adds three semi-supervised strategies in the kernel calculation to enable soft-matching between words, grammatical dependencies, and sentences, respectively. Through convolutions and multi-level semi-supervisions, SCGK provides a powerful model to encode both syntactic and semantic evidence .

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

Yanjun Qi
Machine Learning Department, NEC Labs America
yanjun@nec-labs.com

CP9
Sparse Latent Semantic Analysis

Latent semantic analysis (LSA), as one of the most popular unsupervised dimension reduction tools, has a wide range of applications in text mining and information retrieval. The key idea of LSA is to learn a projection matrix that maps the high dimensional vector space representations of documents to a lower dimensional latent space, i.e. so called latent *topic* space. In this paper, we propose a new model called *Sparse LSA*, which produces a sparse projection matrix via the ℓ_1 regularization. Compared to the traditional LSA, Sparse LSA selects only a small number of relevant words for each topic and hence provides a compact representation of topic-word relationships. Moreover, Sparse LSA is computationally very efficient with much less memory usage for storing the projection matrix. Furthermore, we propose two important extensions of Sparse LSA: group structured Sparse LSA and non-negative Sparse LSA. We conduct experiments on several benchmark datasets and compare Sparse LSA and its extensions with several widely used methods, e.g. LSA, Sparse Coding and LDA. Empirical results suggest that Sparse LSA achieves similar performance gains to LSA, but is more efficient in projection computation, storage, and also well explain the topic-word relationships.

Xi Chen
Carnegie Mellon University
School of Computer Science
xichen@cs.cmu.edu

Yanjun Qi, Bing Bai
Machine Learning Department, NEC Labs America
yanjun@nec-labs.com, bbai@nec-labs.com

Qihang Lin
Carnegie Mellon University
qihanglin@gmail.com

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

CP10
Multidimensional Association Rules in Boolean Tensors

Many datasets are Boolean tensors that denote relations with three dimensions or more. Generalizing association rule mining to such a framework is both a declarative challenge ("what" are interesting rules in tensors) and a procedural one ("how" to efficiently though completely list those rules?). Our proposal tackles both issues within a constraint-based mining setting. A case study on dynamic network analysis thanks to descriptive rules illustrates its added-value.

Loïc Cerf
Department of Computer Science
Universidade Federal de Minas Gerais
lcerf@dcc.ufmg.br

Kim-Ngan Nguyen
Université de Lyon
INSA-Lyon
thi-kim-ngan.nguyen@insa-lyon.fr

Marc Plantevit
Université de Lyon
Université Lyon 1
marc.plantevit@liris.cnrs.fr

Jean-François Boulicaut
Université de Lyon
INSA-Lyon
jean-francois.boulicaut@insa-lyon.fr

CP10
Fast Rule Mining Over Multi-Dimensional Windows

Association rule mining is an indispensable tool for discovering insights from large databases and data warehouses. The data in a warehouse being multi-dimensional, it is often useful to mine rules over subsets of data defined by selections over the dimensions. Such interactive rule mining over multi-dimensional query windows is difficult since rule mining is computationally expensive. Current methods using pre-computation of frequent itemsets require counting of some itemsets by revisiting the transaction database at query time, which is very expensive. We develop a method (RMW) that identifies the minimal set of itemsets to compute and store for each cell, so that rule mining over any query window may be performed without going back to the transaction database. We give formal proofs that the set of itemsets chosen by RMW is sufficient to answer any query and also prove that it is the optimal set to be computed for 1 dimensional queries. We demonstrate through an extensive empirical evaluation that RMW achieves extremely fast query response time compared to existing methods, with only moderate overhead in pre-computation and storage.

Mahashweta Das
University of Texas at Arlington
mahashweta.das@mavs.uta.edu

Deepak Padmanabhan, Prasad Deshpande
IBM Research Laboratory
deepak.s.p@in.ibm.com, prasdesh@in.ibm.com

Ramakrishnan Kannan
IBM India Software Laboratory
rkrishnan@in.ibm.com

CP10
From Black and White to Full Colour: Extending Redescription Mining Outside the Boolean World

Redescription mining is a powerful data analysis tool that is used to find multiple descriptions of the same entities. The current redescription mining methods cannot handle other than Boolean data, making discretization a prerequisite when using non-Boolean data. We extend redescription mining to real-valued data using a surprisingly simple and efficient approach. We demonstrate our algorithm with a real-world task of niche-finding, an important problem in biology.

Esther Galbrun
Department of Computer Science
University of Helsinki, Finland
esther.galbrun@cs.helsinki.fi

Pauli Miettinen

Max-Plank Institute for Informatics
Saarbruecken, Germany
pmiettin@mpi-inf.mpg.de

CP10
A Structure Function for Transaction Data

The ultimate goal of descriptive data mining - in fact of descriptive data analysis in general - is to gain insight in the structure of the data. While the best model may reflect all important structure of D, this is not true for a good model and algorithms often only return a good model rather than the best. Data sets, however, have many models. Different models of the same data set D highlight different aspects of the structure of D. Hence, it makes sense to consider multiple good models of D. The question is: which good models? In this paper we propose a solution for the case were the data is a transaction database [**?**] and the models are code tables [**?**]. More in particular, we introduce a *structure* function, based on the Minimum Description Length (MDL) principle [**?**]. This is a partial function from the set of natural numbers to the set of code tables for D; higher natural numbers are mapped to more complex code tables. Computing the structure function exactly is, unfortunately, too complex. Therefore we introduce the heuristic Groei algorithm, which approximates the true structure function. Through experiments we show that Groei produces a set of good code tables that together provide more insight than any of them alone.

Arno Siebes, Rene Kersten
Dept. of Information and Computing Sciences
Universiteit Utrecht
arno@cs.uu.nl, r.m.kersten1@students.uu.nl

CP10
Characterizing Uncertain Data Using Compression

We study the problem of discovering characteristic patterns in uncertain data. Adopting the possible worlds interpretation of probabilistic data and a compression scheme based on MDL, we formalize the problem of mining patterns that compress the database well in expectation. We devise three methods and empirically compare these on synthetic and real data. Results show that we can extract a small set of meaningful patterns that accurately characterize the data distribution of any probable world.

Francesco Bonchi
Yahoo! Research
bonchi@yahoo-inc.com

Matthijs Van Leeuwen
Universiteit Utrecht
Department of Information and Computing Sciences
mleeuwen@cs.uu.nl

Antti Ukkonen
Yahoo! Research
aukkonen@yahoo-inc.com

CP11
A Probabilistic Hierarchical Approach for Pattern Discovery in Collaborative Filtering Data

This paper presents a hierarchical probabilistic approach to collaborative filtering which allows the discovery and analysis of both global patterns (i.e., tendency of some products of being 'universally appreciated') and local patterns (tendency of users within a community to express a common preference on the same group of items). We reformulate the collaborative filtering approach as a clustering problem in a high-dimensional setting, and propose a probabilistic approach to model the data. The core of our approach is a co-clustering strategy, arranged in a hierarchical fashion: first, user communities are discovered, and then the information provided by each user community is used to discover topics, grouping items into categories. The resulting probabilistic framework can be used for detecting interesting relationships between users and items within user communities. The experimental evaluation shows that the proposed model achieves a competitive prediction accuracy with respect to the state-of-art collaborative filtering approaches.

Nicola Barbieri
DEIS - University of Calabria
nicolabarbieri1@gmail.com

Giuseppe Manco, Ettore Ritacco
ICAR-CNR
manco@icar.cnr.it, ritacco@icar.cnr.it

CP11
Multi-Instance Mixture Models and Semi-Supervised Learning

Multi-instance (MI) learning is a variant of supervised learning where labeled examples consist of bags of feature vectors. Under standard assumptions, MI learning tasks can be approximated as semi-supervised learning tasks. To give insight into this connection we introduce multi-instance mixture models (MIMMs). The cost of the semi-supervised approximation to multi-instance learning is explored, both theoretically and empirically, by analyzing the properties of MIMMs relative to semi-supervised mixture models.

James R. Foulds, Padhraic Smyth
UC Irvine
jrfoulds@gmail.com, smyth@ics.uci.edu

CP11
Online Max-Margin Weight Learning for Markov Logic Networks

Most of the existing weight-learning algorithms for Markov Logic Networks (MLNs) use batch training which becomes computationally expensive and even infeasible for very large datasets since the training examples may not fit in main memory. To overcome this problem, previous work has used online learning algorithms to learn weights for MLNs. However, this prior work has only applied existing online algorithms, and there is no comprehensive study of online weight learning for MLNs. In this paper, we derive a new online algorithm for structured prediction using the primal-dual framework, apply it to learn weights for MLNs, and compare against existing online algorithms on three large, real-world datasets. The experimental results show that our new algorithm generally achieves better accuracy than existing methods, especially on noisy datasets.

Tuyen N. Huynh, Raymond Mooney
The University of Texas at Austin
hntuyen@cs.utexas.edu, mooney@cs.utexas.edu

CP11
Multi-Label Collective Classification

Collective classification in relational data has become an important and active research topic in the last decade, where class labels for a group of linked instances are correlated and need to be predicted simultaneously. Collective classification has a wide variety of real world applications, *e.g.* hyperlinked document classification, social networks analysis and collaboration networks analysis. Current research on collective classification focuses on single-label settings, which assumes each instance can only be assigned with exactly one label among a finite set of candidate classes. However, in many real-world relational data, each instance can be assigned with a set of multiple labels simultaneously. In this paper, we study the problem of multi-label collective classification and propose a novel solution, called Icml (Iterative Classification of Multiple Labels), to effectively assign *a set* of multiple labels to each instance in the relational dataset. The proposed Icml model is able to capture the dependencies among the label sets for a group of related instances and the dependencies among the multiple labels within each label set simultaneously. Empirical studies on real-world tasks demonstrate that the proposed multi-label collective classification approach can effectively boost classification performances in multi-label relational datasets.

Xiangnan Kong, Xiaoxiao Shi, Philip Yu
University of Illinois at Chicago
xkong4@uic.edu, xiaoxiao@cs.uic.edu, psyu@cs.uic.edu

CP11
Famer: Making Multi-Instance Learning Better and Faster

we propose a FAst kernel for Multi-instancE leaRning named as FAMER. FAMER constructs a Locally Sensitive Hashing (LSH) based similarity measure for multi-instance framework, and represents each bag as a histogram by embedding instances within the bag into an auxiliary space, which captures the correspondence information between two bags. By designing a bin-dependent weighting scheme, we not only impose different weights on instances according to their discriminative powers, but also exploit co-occurrence relations according to the joint statistics of instances. Without directly computing in a pairwise manner, the time complexity of FAMER is much smaller compared to other typical multi-instance kernels.

Ye Xu
Dartmouth College
Ye.Xu@Dartmouth.edu

Wei Ping, Jianyong Wang
Tsinghua Univ.
weiping.thu@gmail.com, jywang@tsinghua.edu.cn

Xian-Sheng Hua
Microsoft Research Asia
xshua@microsoft.com

CP12
On Classification of Graph Streams

In this paper, we will examine the problem of classification of massive graph streams. The problem of classification has been widely studied in the database and data mining community. The graph domain poses significant challenges because of the structural nature of the data. The stream scenario is even more challenging, and has not been very well studied in the literature. This is because the underlying graphs can be scanned only once in the stream setting, and it is difficult to extract the relevant structural information from the graphs. In many practical applications, the graphs may be defined over a very large set of nodes. The massive domain size of the underlying graph makes it difficult to learn summary structural information for the classification problem, and particularly so in the case of data streams. In order to address these challenges, we proposed a probabilistic approach for constructing an "in-memory' summary of the underlying structural data. This summary determines the discriminative patterns in the underlying graph with the use of a 2-dimensional hashing scheme. We provide probabilistic bounds on the quality of the patterns determined by the process, and experimentally demonstrate the quality on a number of real data sets.

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

CP12
A Complexity-Invariant Distance Measure for Time Series

We introduce the first complexity invariant distance for time series, and show that it generally produces significant improvements in accuracy. This improvement does not compromise efficiency, since we can lower bound and use a modification of triangular inequality, thus making use of most existing indexing algorithms. We evaluate our ideas with the largest set of time series classification experiments ever attempted, and show that complexity invariant distance measures can produce improvements in accuracy in the vast majority of cases.

Gustavo E. Batista, Xiaoyue Wang, Eamonn Keogh
University of California, Riverside
gbatista@cs.ucr.edu, xwang@cs.ucr.edu, eamonn@cs.ucr.edu

CP12
Time Series Motifs Statistical Significance

Most of the existing algorithms to discover motifs in time series data do not focus on motif evaluation. We present an approach to calculate time series motifs statistical significance. We estimate the expected frequency of a motif using Markov Chain models and compare it to the actual frequency in order to assess a motifs p-value. Our contribution gives means to the application of a powerful technique - statistical tests - to a time series setting.

Nuno C. Castro, Paulo Azevedo
CCTC - Department of Informatics
University of Minho
castro@di.uminho.pt, pja@di.uminho.pt

CP12
Temporal Structure Learning for Clustering Massive Data Streams in Real-Time

Data stream clustering algorithms disregard the information represented by the temporal order of the data points, an important part of the data stream. We propose a new framework which allows us to learn the temporal structure while clustering a data stream by most state-of-the-art data stream clustering algorithm with only minimal overhead. Experiments with intrusion detection show that we are able to considerably improve the results over pure clustering.

Michael Hahsler
Southern Methodist University
mhahsler@lyle.smu.edu

CP12
Significance of Patterns in Time Series Collections

The complexity of time series makes it difficult to assess the significance of patterns found in the data. We propose a new well-grounded null model for time series collections, compare it to the null models of common resampling methods and introduce a new compatible randomization method. Our experiments compare the behavior of the various methods and reflect the results to the differences in their null models.

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

Petteri Kaski
Aalto University
petteri.kaski@tkk.fi

PP0
The Infinite Push: A New Support Vector Ranking Algorithm That Directly Optimizes Accuracy at the Absolute Top of the List

I will describe a new ranking algorithm called the 'Infinite Push' that directly optimizes ranking accuracy at the absolute top of the list. The algorithm is a support vector style algorithm, but due to the different objective, it no longer leads to a quadratic programming problem. Instead, the dual optimization problem involves $l_{1,\infty}$ constraints; we solve this using an efficient gradient projection method. Experiments on real-world data sets confirm the algorithm's focus on accuracy at the absolute top of the list.

Shivani Agarwal
Indian Institute of Science
shivani@csa.iisc.ernet.in

PP0
On Anonymization of Multi-Graphs

The problem of privacy-preserving data mining has attracted considerable attention in recent years because of increasing concerns about the privacy of the underlying data. In recent years, an important data domain which has emerged is that of graphs and structured data. Many data sets such as XML data, transportation networks, traffic in IP networks, social networks and hierarchically structured data are naturally represented as graphs. Existing work on graph privacy has focussed on the problem of anonymizing nodes or edges of a single graph, in which the identity is assumed to be associated with individual nodes. In this paper, we examine the more complex case, where we have a collection of graphs, and the identity is associated with individual graphs rather than nodes or edges. In such cases, the problem of identity anonymization is extremely difficult, since we need to not only anonymize the labels on the nodes, but also the underlying global structural information. In such cases, both the global and local structural information can be a challenge to the anonymization process, since any combination of such information can be used in order to de-identify the underlying graphs. In order to achieve this goal, we will create synthesized representations of the underlying graphs based on aggregate structural analytics of the collection of graphs. The synthesized graphs retain the properties of the original data while satisfying the k-anonymity requirement. Our experimental results show that the synthesized graphs maintain a high level of structural information and compatible classification accuracies with the original data.

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

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

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

PP0
Semi-Supervised Variable Weighting for Clustering

We focus on semi-supervised variable weighting for clustering. Besides exploiting both labeled and unlabeled data to effectively identify the real importance of variables, our method embeds variable weighting in the process of semi-spervised clustering, rather than calculating variable weights separately, to ensure the computation efficiency. Our experimental results demonstrate that semi-supervised variable weighting significantly improves the clustering accuracy of existing semi-supervised k-means without variable weighting, or with unsupervised vari-able weighting.

Ling Chen, Chengqi Zhang
University of Technology, Sydney
ling.chen@uts.edu.au, ling.chen@uts.edu.au

PP0
Sparse Solutions for Single Class Svms: A Bi-Criterion Approach

We propose an innovative learning algorithm - a variation of One-class $\nu-$SVMs learning algorithm to produce sparser solutions with a reduced run-time complexity. The proposed technique returns an approximate solution, nearly as good as the solution obtained by the classical approach, by minimizing the original risk function along with a regularization term. The proposed algorithm closely preserves the accuracy of standard one-class $\nu-$SVMs while reducing both training time and test time by several factors.

Santanu Das
UARC/UCSC
santanu.das-1@nasa.gov

Nikunj Oza
Nasa Ames Research Center

nikunj.c.oza@nasa.gov

PP0

Maximising the Quality of Influence

In percolation theory, graph vertices are either active or inactive, and a percolation process decides how activation spreads. Firstly, we propose and analyse a simple data-driven percolation process. Secondly, we generalise the following problem considered in [Kempe et al., 2003]: which k vertices maximise the number of active vertices at the end of percolation process? In our case, activations are considered in $[0, 1]$ measuring the "quality' of percolation, and percolation decays occur along edges.

Charanpal S. Dhanjal, Stephan Clemencon
Telecom ParisTech
charanpal.dhanjal@telecom-paristech.fr,
stephan.clemencon@telecom-paristech.fr

PP0

Active Teaching for Inductive Learners

We propose and study a new intelligent teaching paradigm called *active teaching.* In contrast to active learning, we assume that the learner can only passively conduct inductive learning from the given examples, but the teacher (oracle) can actively provide "good" examples to the learner, in order to speed up the teaching (learning) process. We establish a framework with four specific paradigms of active teaching, and develop the corresponding teaching strategies.

Jun Du
The University of Western Ontario
jdu42@csd.uwo.ca

Charles Ling
The Universtiy of Western Ontario
cling@csd.uwo.ca

PP0

An Algorithm for Sparse Pca Based on a New Sparsity Control Criterion

Sparse principal component analysis (PCA) imposes extra constraints or penalty terms to the standard PCA to achieve sparsity. In this paper, we first introduce an efficient algorithm for finding a single sparse principal component (PC) with a specified cardinality. Moreover, combining our algorithm for computing a single sparse PC with the Schur complement deflation scheme, we develop an algorithm which sequentially computes multiple PCs by greedily maximizing the *adjusted variance* explained by them. On the other hand, to address the difficulty of choosing the proper sparsity and parameter in various sparse PCA algorithms, we propose a new PCA formulation whose aim is to minimize the sparsity of the PCs while requiring that their *relative adjusted variance* is larger than a given fraction. We also show that a slight modification of the aforementioned multiple component PCA algorithm can also find sharp solutions of the latter formulation.

Yunlong He, Renato D.C. Monteiro, Haesun Park
Georgia Institute of Technology
he.yunlong@gmail.com, renato.monteiro@isye.edu,
hpark@cc.gatech.edu

PP0

Multi-Task Multiple Kernel Learning

Two novel formulations for learning shared feature representations across multiple tasks are presented. The idea is to learn shared kernels that are sparse linear combinations of given base kernels. The second formulation further searches for low-dimensional subspaces in the space induced by the selected kernels. One main contribution is an efficient mirror-descent based algorithm for solving this formulation which is an instance of a mixed Schatten-norm regularized problem. Simulations illustrate the efficacy of the formulations.

Sakethanath Jagarlapudi
Indian Institute of Technology, Bombay
saketh@cse.iitb.ac.in

Pratik Jawanpuria
Indian Institute of Technology, Bombay
INDIA
pratik.j@cse.iitb.ac.in

PP0

One-Class-Based Uncertain Data Stream Learning

This paper presents a novel approach to one-class-based uncertain data stream learning. Our proposed approach works in three steps. Firstly, we put forward a local kernel-density-based method to generate a bound score for each instance, which refines the location of the corresponding instance. Secondly, we construct an uncertain one-class classifier by incorporating the generated bound score into a one-class SVM-based learning phase. Thirdly, we devise an ensemble classifier, integrated from uncertain one-class classifiers built on the current and historical chunks, to cope with the concept drift involved in the uncertain data stream environment. Our proposed method explicitly handles the uncertainty of the input data and enhances the ability of one-class learning in reducing the sensitivity to noise. Extensive experiments on uncertain data streams demonstrate that our proposed approach can achieve better performance and is highly robust to noise in comparison with state-of-the-art one-class learning method.

Bo Liu, Yanshan Xiao, Longbing Cao
UTS
csbliu@gmail.com, ysxiao@it.uts.edu.au,
lbcao@it.uts.edu.au

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

PP0

Unsupervised Disaggregation of Low Frequency Power Measurements

Fear of increasing prices and concern about climate change are motivating residential power conservation efforts. We investigate the effectiveness of several unsupervised disaggregation methods on low frequency power measurements collected in real homes. Specifically, we consider variants of the factorial hidden Markov model. Our results indicate that a conditional factorial hidden semi-Markov model, which integrates additional features related to when and how appliances are used in the home and more accurately represents the power use of individual appliances, outper-

forms the other unsupervised disaggregation methods. Our results show that unsupervised techniques can provide per-appliance power usage information in a non-invasive manner, which is ideal for enabling power conservation efforts.

Hyungsul Kim
University of Illinois at Urbana-Champaign
hkim21@illinois.edu

Manish Marwah, Martin Arlitt, Geoff Lyon
HP Labs
manish.marwah@hp.com, martin.arlitt@hp.com,
geoff.lyon@hp.com

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

PP0
Weighted Graph Compression for Parameter-Free Clustering With Pacco

Object similarities are more and more characterized by connectivity information available in form of graph data. Therein, we propose a novel clustering algorithm for weighted graphs, called *PaCCo* (*Pa*rameter-free *C*lustering by *Co*ding costs). *PaCCo* is based on the Minimum Description Length (MDL) principle in combination with a bisecting k-Means strategy. By relating the clustering problem to data compression, good graph cluster structures enable strong graph compressions. Resulting groups of highly connected nodes reveal valuable knowledge.

Nikola S. Mueller
Max Planck Institute of Biochemistry
nimuell@biochem.mpg.de

Katrin Haegler, Junming Shao
University of Munich, Germany
katrin.haegler@med.uni- muenchen.de,
shao@dbs.ifi.lmu.de

Claudia Plant
Florida State University
cplant@fsu.edu

Christian Boehm
University of Munich, Germany
boehm@dbs.ifi.lmu.de

PP0
Human Dynamics in Large Communication Networks

How often humans communicate with each other? What are the mechanisms that explain how human actions are distributed over time? Here we answer these questions by studying the time interval between calls and SMS messages in an anonymized, large mobile network, with 3.1 million users, over 200 million phone calls and 300 million SMS messages,spanning 70 GigaBytes. Our first contribution is the Truncated Autocatalytic Process (TAP) model, that explains the time between communication events (ie., times between phone-initiations) for a single individual. The novelty is that the model is autocatalytic, in the sense that the parameters of the model change, depending on the latest inter-event time: long periods of inactivity in the past result in long periods of inactivity in the future, and vice-versa. We show that the TAP model mimics the inter-event times of the users of our dataset extremely well, despite its parsimony and simplicity. Our second contribution is the TAP-classifier , a classification method based on the inter event times and in addition to other features. We showed that the inferred sleep intervals and the reciprocity between outgoing and incoming calls are good features to classify users. Finally, analyze the network effects of each class of users and we found surprising results. Moreover, all of our methods are fast, and scale linearly with the number of customers.

Pedro Olmo Vaz de Melo
UFMG
pedro.olmo@gmail.com

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

Antonio Loureiro
UFMG
loureiro@dcc.ufmg.br

PP0
Scalable Software-Defect Localisation by Hierarchical Mining of Dynamic Call Graphs

The localisation of defects in computer programmes is essential in software engineering and is important in domain-specific data mining. Existing techniques which build on call-graph mining localise defects well, but do not scale for large software projects. This paper presents a hierarchical approach with good scalability characteristics. It makes use of novel call-graph representations, frequent subgraph mining and feature selection. It first analyses call graphs of a coarse granularity, before it *zooms-in* into more fine-grained graphs. We evaluate our approach with defects in the Mozilla Rhino project: In our setup, it narrows down the code a developer has to examine to about 6% only.

Frank Eichinger, Christopher Oßner, Klemens Böhm
Karlsruhe Institute of Technology (KIT)
Institute for Program Structures and Data Organization (IPD)
eichinger@kit.edu, ossner@kit.edu,
klemens.boehm@kit.edu

PP0
Discovering Bucket Orders from Data

The problem of obtaining an order on a set of entities which contain inherent ties among them arises in many applications. Bucket order has turned out to be a useful tool in ordering entities in such applications. A bucket order is a partition of the set of entities into “buckets’. There is a total order on the buckets, but the entities within a bucket are treated as tied. The bucket order problem is, given a set of input rankings, compute a bucket order that best captures the ordering preferences in the input. In many applications, the discrepancies in the input rankings are “local’. We present a formal model to capture such settings and consider the following question: “how many input rankings need to be sampled to discover the underlying bucket order on n entities?’. We prove an upperbound of $O(\sqrt{\log n})$. We present a new approach for the bucket order problem which exploits a connection between the discovery of buckets and the correlation clustering problem. We present empirical evaluation of our algorithms on real

and artificially generated datasets.

Vinayaka Pandit
IBM India Research Lab
pvinayak@in.ibm.com

Sreyash Kenkre
IBM Research - India
srkenkre@in.ibm.com

Arindam Khan
Georgia University of Technology
arindamkhan.cs.iitkgp@gmail.com

PP0

WindMine: Fast and Effective Mining of Web-Click Sequences

Given a large stream of users clicking on web sites, how can we find trends and anomalies? We have developed a novel method, WindMine, to find patterns and anomalies in such datasets. Our approach has the following advantages: (a) it is effective in discovering meaningful patterns, (b) it automatically determines suitable window sizes, and (c) it is efficient, in terms of computation time. We examine the effectiveness and scalability by performing experiments on real datasets.

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

Lei Li
Carnegie Mellon University
leili@cs.cmu.edu

Yasuko Matsubara
Kyoto University
y.matsubara@db.soc.i.kyoto-u.ac.jp

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

PP0

The Odd One Out: Identifying and Characterising Anomalies

In many situations there exists an abundance of positive examples, but only a handful of negatives. In this paper we show how in binary or transaction data such rare cases can be identified and characterised. Our approach uses the Minimum Description Length principle to decide whether an instance is drawn from the training distribution or not. By using frequent itemsets to construct this compressor, we can easily and thoroughly characterise the decisions, and explain what changes in an example would lead to a different verdict. Furthermore, we give a technique through which, given only a few negative examples, the decision landscape and optimal boundary can be predicted—making the approach parameter-free. Experimentation on benchmark and real data shows our method provides very high classification accuracy, thorough and insightful characterisation of decisions, predicts the decision landscape reliably, and can pinpoint observation errors. Moreover, a case study on real MCADD data shows we provide an interpretable approach with state-of-the-art performance for screening newborn babies for rare diseases.

Koen Smets, Jilles Vreeken
Universiteit Antwerpen
koen.smets@ua.ac.be, jilles.vreeken@ua.ac.be

PP0

Computationally Generated Cardiac Biomarkers: Heart Rate Patterns to Predict Death Following Coronary Attacks

We propose novel computational biomarkers to identify patients at an increased risk of mortality following coronary attacks. These computational biomarkers are based on the discovery of approximately conserved heart rate sequences that are significantly overrepresented in either high or low risk patients. We propose a randomized hashing- and greedy centroid selection-based algorithm to efficiently discover such heart rate patterns in large electrocardiographic datasets and present the results of evaluating these ideas in over 3,000 patients.

Zeeshan Syed, Chih-Chun Chia
University of Michigan
zhs@umich.edu, jazzchia@umich.edu

PP0

Sequential Minimal Optimization in Adaptive-Bandwidth Convex Clustering

Convex clustering is an effective approach in computing not the local, but the global optimum of a cluster assignment. The existing Expectation-Maximization algorithm was computationally inefficient, because an extremely large number of iterations is required for the convergence. We propose more efficient optimization algorithm to significantly reduce the required number of iterations, in which accurate pruning while choosing a pair of kernels and an element-wise Newton-Raphson method are combined.

Rikiya Takahashi
IBM Research - Tokyo
rikiya@jp.ibm.com

PP0

Online Discovery of Top-K Similar Motifs in Time Series Data

A motif is a pair of non-overlapping sequences with very similar shapes in a time series. We study the online *top-k* most similar motif discovery problem. A special case of this problem corresponding to $k = 1$ was investigated in the literature by *Mueen and Keogh*. We generalize the problem to any k and propose space-efficient algorithms for solving it. We show that our algorithms are optimal in term of space. In the particular case when $k = 1$, our algorithms achieve better performance both in terms of space and time consumption than the algorithm of *Mueen and Keogh*. We demonstrate our results by both theoretical analysis and extensive experiments with both synthetic and real-life data. Finally, we show possible applications of the *top-k* similar motifs discovery problem.

Hoang Thanh Lam
TU Eindhoven
t.l.hoang@tue.nl

Ninh Dang Pham
Ho Chi Minh University of Technology Vietnam

ninhpham@contentinterface.com

Toon Calders
TU Eindhoven
t.calders@tue.nl

PP0
Gaussian Process for Dimensionality Reduction in Transfer Learning

Dimensionality reduction has been considered as one of the most significant tools for data analysis. In general, supervised information is helpful for dimensionality reduction. However, in typical real applications, supervised information in multiple source tasks may be available, while the data of the target task are unlabeled. An interesting problem of how to guide the dimensionality reduction for the unlabeled target data by exploiting useful knowledge, such as label information, from multiple source tasks arises in such a scenario. In this paper, we propose a new method for dimensionality reduction in the transfer learning setting. Unlike traditional paradigms where the useful knowledge from multiple source tasks is transferred through distance metric, our proposal firstly converts the dimensionality reduction problem into integral regression problems in parallel. Gaussian process is then employed to learn the underlying relationship between the original data and the reduced data. Such a relationship can be appropriately transferred to the target task by exploiting the prediction ability of the Gaussian process model and inventing different kinds of regularizers. Extensive experiments on both synthetic and real data sets show the effectiveness of our method.

Bin Tong
Kyushu University
bintong@i.kyushu-u.ac.jp

Junbin Gao
Charles Sturt University, Australia
jbgao@csu.edu.au

Nguyen Huy Thach, Einoshin Suzuki
Kyushu University, Japan
thachnh@i.kyushu-u.ac.jp, suzuki@i.kyushu-u.ac.jp

PP0
Semantic Graph Kernels for Automated Reasoning

We propose a semantic graph kernel suitable for learning in structured mathematical domains. Our kernel incorporates contextual information about the features and unlike 'random walk'-based graph kernels it is also applicable to sparse graphs. For evaluation we use a subset of the large formal Mizar mathematical library. The experiments demonstrate that the proposed kernel leads to consistent improvement in performance compared to linear, Gaussian, latent semantic, and geometric graph kernels.

Evgeni Tsivtsivadze, Josef Urban, Herman Geuvers, Tom Heskes
Radboud University
evgeni@science.ru.nl, josef.urban@science.ru.nl, herman.geuvers@science.ru.nl, t.heskes@science.ru.nl

PP0
Exemplar-Based Robust Coherent Biclustering

The biclustering, co-clustering, or subspace clustering problem involves simultaneously grouping the rows and columns of a data matrix to uncover biclusters or submatrices that optimize a desired objective function or coherence measure. We introduce a novel formulation of the coherent biclustering problem that is exemplar-based and robust to background interference, and we use it to derive two algorithms that are shown to be competitive with the current state-of-the-art algorithms for finding coherent biclusters.

Kewei Tu
Department of Computer Science
Iowa State University
tukw@iastate.edu

Xixiu Ouyang, Dingyi Han, Yong Yu
Department of Computer Science and Engineering
Shanghai Jiaotong University
xxouyang@apex.sjtu.edu.cn, handy@apex.sjtu.edu.cn, yyu@apex.sjtu.edu.cn

Vasant Honavar
Computer Science Department
Iowa State University
honavar@cs.iastate.edu

PP0
IMet: Interactive Metric Learning in Healthcare Applications

Patient similarity assessment aims at providing a clinically meaningful distance measure for case retrieval in the context of clinical decision intelligence. Two of the key challenges are how to incorporate physician feedback with regard to the retrieval results and how to interactively update the underlying similarity measure based on the feedback. In this paper, we present the interactive Metric learning (iMet) method that can incrementally adjust the underlying distance metric based on latest supervision information. iMet is designed to scale linearly with the data set size based on matrix perturbation theory, which allows the derivation of sound theoretical guarantees. We show empirical results demonstrating that iMet outperforms the baseline by three orders of magnitude in speed while obtaining comparable accuracy on several benchmark datasets. We also describe the application of the algorithm in a real world physician decision support system.

Fei Wang
Department of Statistical Science, Cornell University
feiwang03@gmail.com

Jimeng Sun
IBM Research
jimeng@us.ibm.com

Jianying Hu, Ebadollahi Shahram
IBM
jyhu@us.ibm.com, ebad@us.ibm.com

PP0
Efficient Document Clustering Via Online Nonneg-

ative Matrix Factorizations

In recent years, Nonnegative Matrix Factorization (NMF) has received considerable interest from the data mining and information retrieval fields. It has already been successfully used in document clustering and image representation. However, despite its empirical success, there are still some limitations that restrict its application to real world problems: (1) NMF needs to hold the whole data matrix in main memory in the whole iteration process, which is impossible for large scale data sets; (2) NMF cannot handle new coming data points efficiently, which makes it incapable of processing streaming data. In this paper, we propose a new on-line NMF approach to address these problems based on stochastic approximations, which scales up gracefully to large datasets with millions of samples. Moreover, on-line NMF proceeds one data point at a time, which can naturally handle the streaming data. Experiments are conducted on several real world document data sets to demonstrate the efficacy and efficiency of the proposed method.

Fei Wang
Department of Statistical Science, Cornell University
feiwang03@gmail.com

Chenhao Tan
Cornell
chenhao@cs.cornell.edu

Christian Konig
Microsoft
chrisko@microsoft.com

Ping Li
Cornell
pingli@cornell.edu

PP0

Diversified Trajectory Pattern Ranking in Geo-Tagged Social Media

Social media such as those residing in the popular photo sharing websites is attracting increasing attention in recent years. As a type of user-generated data, wisdom of the crowd is embedded inside such social media. In particular, millions of users upload to Flickr their photos, many associated with temporal and geographical information. In this paper, we investigate how to rank the trajectory patterns mined from the uploaded photos with geotags and timestamps. The main objective is to reveal the collective wisdom recorded in the seemingly isolated photos and the individual travel sequences reflected by the geo-tagged photos. Instead of focusing on mining frequent trajectory patterns from geo-tagged social media, we put more effort into ranking the mined trajectory patterns and diversifying the ranking results. Through leveraging the relationships among users, locations and trajectories, we rank the trajectory patterns. We then use an exemplar-based algorithm to diversify the results in order to discover the representative trajectory patterns. We have evaluated the proposed framework on 12 different cities using a Flickr dataset and demonstrated its effectiveness.

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

Jiebo Luo
Kodak Research Laboratories
jiebo.luo@kodak.com

Thomas Huang
University of Illinois at Urbana-Champaign
huang@ifp.uiuc.edu

PP0

Extending Consensus Clustering to Explore Multiple Clustering Views

We develop Multiple Consensus Clustering (MCC) to explore multiple clustering views of a given dataset. Instead of generating a single consensus, MCC organizes the different input clusterings into a hierarchical tree structure and allows for interactive exploration of multiple clustering solutions. A dynamic programming algorithm is proposed to obtain a flat partition from the hierarchical tree using the modularity measure. Multiple consensuses are finally obtained by applying consensus clustering algorithms to each clustering cluster.

Yi Zhang, Tao Li
Florida International University
yzhan004@cs.fiu.edu, taoli@cs.fiu.edu