IP1
Big Data, Small Models, and Extreme Behaviors in the Real World

Nobody would argue that data are getting bigger, not only in terms of the amount available but also in terms of its resolution in time, in space, and now down toward the level of the individual objects in an otherwise large interconnected population. In the dark ages, there was far too little data available to distinguish between multiple theories of how real-world complex systems work, be they from a biological system such as the brain; a social system such as an insurgency or terrorist campaign; or a financial market. However the tide has turned and available data now far outstrips theoretical understanding in these fields. In this talk I discuss the importance of developing minimal generative models that are consistent with the statistical features of the available big data being produced; I argue that the development of each should go hand in hand; and I argue that extreme real-world behaviors can be used to sort out the good from the bad in terms of what additional data to collect and how to mine it. Big data examples to be discussed range from collective neuronal processes in the brain, through to online global extremist activity and subsecond temporal fractures in global financial exchanges.

Neil F. Johnson
Dept. of Physics
University of Miami
njohnson@physics.miami.edu

IP2
Title Not Available at Time of Publication

Abstract not available at time of publication.

Virgilio Almeida
Federal University of Minas Gerais
virgilio@dcc.ufmg.br

IP3
Data and Algorithmic Bias in the Web

The Web is the largest public big data repository that humankind has created. In this overwhelming data ocean we need to be aware of the quality and in particular, of biases the exist in this data, such as redundancy, spam, etc. These biases affect the algorithms that we design to improve the user experience. This problem is further exacerbated by biases that are added by these algorithms, specially in the context of recommendation systems. We give several examples and their relation to sparsity, novelty, and privacy, stressing the importance of the user context to avoid these biases.

Ricardo Baeza-Yates
Universidad de Chile
Department of Computer Science
ricardo.baeza@upf.edu

IP4
Sum-Product Networks: Deep Models with Tractable Inference

Sum-product networks (SPNs) are a new class of deep probabilistic models where inference remains tractable regardless of the number of hidden layers. I will present generative and discriminative algorithms for learning SPN weights, and an algorithm for learning SPN structure. SPNs have achieved impressive results in a wide variety of domains, including object recognition, image completion, activity recognition, language modeling, collaborative filtering, and click prediction. Our algorithms can easily learn SPNs with many layers of latent variables, making them arguably the most powerful type of deep learning to date. (Joint work with Abe Friesen, Rob Gens and Hoifung Poon.)

Pedro Domingos
University of Washington
Dept. of Computer Science & Engineering
pedrod@cs.washington.edu

CP1
Multi-Domain Manifold Learning for Drug-Target Interaction Prediction

Drug-target interaction (DTI) provides novel insights about the genomic drug discovery, and is critical to the drug discovery. Recently, researchers try to incorporate different information about drugs and targets for prediction. However, the heterogeneous and high-dimensional data is an great challenge for existing machine learning methods. In the last few years, extensive research efforts have been devoted to the utilization of manifold property on high dimensional data, e.g. dimension reduction methods preserving local structures of the manifolds. Motivated by the successes of these studies, we propose a general framework incorporating both manifold structures and known interaction/non-interaction information to predict the drug-target interactions. To overcome the challenges of domain scaling and information inconsistency, we formulate the problem with Semidefinite Programming(SDP), including new constraints to improve the robustness of the learning procedure. A variety of optimization techniques are also designed to enhance the scalability of the problem solver. Effectiveness of the method is evaluated by experiments on the benchmark dataset. Compared with other state-of-the-art methods, the proposed methods can lead to better results for the drug-target interaction prediction.

Ruichu Cai
Guangdong University of Technology
cairuichu@gmail.com

CP1
Regularized Weighted Linear Regression for High-Dimensional Censored Data

Survival analysis aims at modeling time to event data which occurs ubiquitously in many biomedical and healthcare applications. One of the critical challenges with modeling such survival data is the presence of censored outcomes which cannot be handled by standard regression models. In this paper, we propose a regularized linear regression model with weighted least-squares to handle the survival prediction in the presence of censored instances. We also employ the elastic net penalty term for inducing sparsity into the linear model to effectively handle high-dimensional data. As opposed to the existing censored linear models, the parameter estimation of our model does not need any prior estimation of survival times of censored instances. In addition, we propose a self-training framework which is able to improve the prediction performance of our proposed linear model. We demonstrate the performance of the proposed model using several real-world

high-dimensional biomedical benchmark datasets and our experimental results indicate that our model outperforms other related competing methods and attains very competitive performance on various datasets.

Yan Li, Bhanukiran Vinzamuri, Chandan Reddy
Wayne State University
rock_liyan@wayne.edu, bhanukiranv@wayne.edu, reddy@cs.wayne.edu

CP1

Kernelized Sparse Self-Representation for Clustering and Recommendation

Sparse models have demonstrated substantial success in applications for data analysis such as clustering, classification and denoising. However, most of the current work is built upon the assumption that data is distributed in a union of subspaces, whereas limited work has been conducted on nonlinear datasets where data reside in a union of manifolds rather than a union of subspaces. To understand data nonlinearity using sparse models, in this paper, we propose to exploit the self-representation property of nonlinear data in an implicit feature space using kernel methods. We propose a kernelized sparse self-representation model, denoted as KSSR, and a novel Kernelized Fast Iterative Soft-Thresholding Algorithm, denoted as K-FISTA, to recover the underlying nonlinear structure among the data. We evaluate our method for clustering problems on both synthetic and real-world datasets, and demonstrate its superior performance compared to the other state-of-the-art methods. We also apply our method for collaborative filtering in recommender systems, and demonstrate its great potential for novel applications beyond clustering.

Xiao Bian
GE Global Research
xiao.bian@ge.com

Feng Li
Indiana University-Purdue University Indianapolis
lifen@umail.iu.edu

Xia Ning
Indiana University - Purdue University Indianapolis
xning@cs.iupui.edu

CP1

Finding Surprisingly Frequent Patterns of Variable Lengths in Sequence Data

We address the problem of finding 'surprising' patterns of variable length in sequence data, where a surprising pattern is defined as a subsequence of a longer sequence, whose observed frequency is statistically significant with respect to a given distribution. Finding statistically significant patterns in sequence data is the core task in some interesting applications such as Biological motif discovery and anomaly detection. We show that the presence of few 'true' surprising patterns in the data could cause a large number of highly-correlated patterns to stand statistically significant just because of those few significant patterns. Our approach to solving the 'redundant patterns' problem is based on capturing the dependencies between patterns through an 'explain' relationship where a set of patterns can explain the statistical significance of another pattern. This allows us to address the problem of redundancy by choosing a few 'core' patterns which explain the significance of all other significant patterns. We propose a greedy algorithm for efficiently finding an approximate *core* pattern set of minimum size. Using both synthetic and real-world sequential data, chosen from different domains including Medicine and Bioinformatics, we show that the proposed notion of *core* patterns very closely matches the notion of 'true' surprising patterns in data.

Reza Sadoddin
Department of Computing Science
University of Alberta
sadoddin@ualberta.ca

Joerg Sander
Department of Computing Science, University of Alberta
jsander@ualberta.ca

Davood Rafiei
Department of Computing Science
University of Alberta
drafiei@ualberta.ca

CP1

Clustering in the Face of Fast Changing Streams

Clustering is arguably the most important primitive for data mining, finding use as a subroutine in many higher-order algorithms. In recent years, the community has redirected its attention from the batch case to the online case. This need to support online clustering is engendered by the proliferation of cheap ubiquitous sensors that continuously monitor various aspects of our world, from heartbeats as we exercise to the number of mosquitoes visiting a well in a village in Ethiopia. In this work, we argue that current online clustering solutions offer a room for improvement. To some degree they all have at least one of the following shortcomings: they are parameter-laden, only defined for certain distance functions, sensitive to outliers, and/or they are approximate. This last point requires clarification; in some sense almost all clustering algorithms are approximate. For example, in general, k-means only approximately optimizes its objective function. However, streaming versions of the k-means algorithm are further approximating this approximation, potentially leading to very poor solutions. In this work, we introduce an algorithm that mitigates these flaws. It is parameter-lite, defined for any distance function, insensitive to outliers and produces the same output as the batch version of the algorithm. We demonstrate the utility and effectiveness of our ideas with case studies in entomology, cardiology and biological audio processing.

Liudmila Ulanova
University of California, Riverside
lulan001@ucr.edu

Nurjahan Begum, Mohammad Shokoohi-Yekta
UC Riverside
nbegu001@ucr.edu, mshok002@ucr.edu

Eamonn Keogh
University of California, Riverside
eamonn@cs.ucr.edu

CP1

Identifying Connectivity Patterns for Brain Diseases Via Multi-Side-View Guided Deep Architec-

tures

There is considerable interest in mining neuroimage data to discover clinically meaningful connectivity patterns to inform an understanding of neurological and neuropsychiatric disorders. Subgraph mining models have been used to discover connected subgraph patterns. However, it is difficult to capture the complicated interplay among patterns. As a result, classification performance based on these results may not be satisfactory. To address this issue, we propose to learn non-linear representations of brain connectivity patterns from deep learning architectures. This is non-trivial, due to the limited subjects and the high costs of acquiring the data. Fortunately, auxiliary information from *multiple side views* such as clinical, serologic, immunologic, cognitive and other diagnostic testing also characterizes the states of subjects from different perspectives. In this paper, we present a novel **M**ulti-side-**V**iew guided **A**uto**E**ncoder (MVAE) that incorporates multiple side views into the process of deep learning to tackle the bias in the construction of connectivity patterns caused by the scarce clinical data. Extensive experiments show that MVAE not only captures discriminative connectivity patterns for classification, but also discovers meaningful information for clinical interpretation.

Jingyuan Zhang
University of Illinois at Chicago
University of Illinois at Chicago
jzhan8@uic.edu

Bokai Cao, Sihong Xie, Chun-Ta Lu, Philip Yu
University of Illinois at Chicago
caobokai@uic.edu, sxie6@uic.edu, clu29@uic.edu, psyu@uic.edu

Ann Ragin
Northwestern University
ann-ragin@northwestern.edu

CP2

Exploiting Emotional Information for Trust/Distrust Prediction

Trust and distrust networks are usually extremely sparse and the vast majority of the existing algorithms for trust/distrust prediction suffer from the data sparsity problem. In this paper, following the research from psychology and sociology, we envision that users' emotions such as happiness and anger are strong indicators of trust/distrust relations. Meanwhile the popularity of social media encourages the increasing number of users to freely express their emotions; hence emotional information is pervasively available and usually denser than the trust and distrust relations. Therefore incorporating emotional information could have the potentials to alleviate the data sparsity in the problem of trust/distrust prediction. In this study, we investigate how to exploit emotional information for trust/distrust prediction. In particular, we provide a principled way to capture emotional information mathematically and propose a novel trust/distrust prediction framework ETD. Experimental results on the real-world social media dataset demonstrate the effectiveness of the proposed framework and the importance of emotional information in trust/distrust prediction.

Ghazaleh Beigi
Arizona State University
Arizona State University
gbeigi@asu.edu

Jiliang Tang
Yahoo! Labs
jlt@yahoo-inc.com

Suhang Wang, Huan Liu
Arizona State University
suhang.wang@asu.edu, huan.liu@asu.edu

CP2

Integrating Community and Role Detection in Information Networks

Traditional studies treat community detection and role detection in information networks as two orthogonal issues. We propose a novel probabilistic network model, the Mixed Membership Community and Role model (MMCR), which models the latent community and role of each node simultaneously, and the probability of links are defined accordingly. By testing our model on synthetic and two real-world networks, we demonstrate that our approach leads to better performance for both community detection and role detection.

Ting Chen
Northeastern University
tingchen@ccs.neu.edu

Lu-An Tang
NEC Laboratories America, Inc.
ltang@nec-labs.com

Yizhou Sun
Northeastern University
yzsun@ccs.neu.edu

Zhengzhang Chen, Haifeng Chen, Guofei Jiang
NEC Laboratories America, Inc.
zchen@nec-labs.com, haifeng@nec-labs.com, gfj@nec-labs.com

CP2

Uncovering Multiple Diffusion Networks Using the First-Hand Sharing Pattern

The problem of finding hidden diffusion processes in a network is getting more attention since after understanding the process, one can manipulate the diffusion speed of the process. In the real world, most nodes are inclined to share the first-hand information, regarded as the first-hand sharing pattern. We propose a generative model with the pattern and design the corresponding optimization method to infer both the hidden networks and transmission rates between nodes.

Pei-Lun Liao, Chung-Kuang Chou, Ming-Syan Chen
National Taiwan University
plliao@arbor.ee.ntu.edu.tw, ckchou@arbor.ee.ntu.edu.tw, mschen@cc.ee.ntu.edu.tw

CP2

Online Prediction of User Actions Through An Ensemble Vote from Vector Representation and Frequency Analysis Models

The ability to predict users next actions is useful to many applications. We propose an online algorithm that combines scores from two models. In the frequency model, the score of an action is calculated based on the frequency that

the action has occurred right after a context. In the vector representation model, a vector for each action is learned, and a score for an action is calculated based on the similarity of vectors of actions.

Changsung Moon, Dakota Medd, Paul Jones, Steve Harenberg
North Carolina State University
cmoon2@ncsu.edu, drmedd@ncsu.edu, pjones@ncsu.edu, sdharenb@ncsu.edu

William Oxbury
Heilbronn Institute
wm.oxbury@bristol.ac.uk

Nagiza Samatova
North Carolina State University
Oak Ridge National Laboratory
samatova@csc.ncsu.edu

CP2

FairPlay: Fraud and Malware Detection in Google Play

Fraudulent behaviors in Google's Android app market fuel search rank abuse and malware proliferation. We present FairPlay, a novel system that uncovers both malware and search rank fraud apps, by picking out trails that fraudsters leave behind. To identify suspicious apps, FairPlay's PCF algorithm correlates review activities and uniquely combines detected review relations with linguistic and behavioral signals gleaned from longitudinal Google Play app data. We contribute a new longitudinal app dataset to the community, which consists of over 87K apps, 2.9M reviews, and 2.4M reviewers, collected over half a year. FairPlay achieves over 95% accuracy in classifying gold standard datasets of malware, fraudulent and legitimate apps. We show that 75% of the identified malware apps engage in search rank fraud. FairPlay discovers hundreds of fraudulent apps that currently evade Google Bouncer's detection technology, and reveals a new type of attack campaign, where users are harassed into writing positive reviews, and install and review other apps.

Bogdan Carbunar, Mahmudur Rahman, Mizanur Rahman
FIU
carbunar@gmail.com, mrahm004@fiu.edu, md.mizanur.rahman@csebuet.org

Duen Horng Chau
Georgia Tech
polo@gatech.edu

CP2

Node Classification in Signed Social Networks

Node classification in social networks has been proven to be useful in many real-world applications. The vast majority of existing algorithms focus on unsigned social networks (or social networks with only positive links), while little work exists for signed social networks. It is evident from recent developments in signed social network analysis that negative links have added value over positive links. Therefore, the incorporation of negative links has the potential to benefit various analytical tasks. In this paper, we study the novel problem of node classification in signed social networks. We provide a principled way to mathematically model positive and negative links simultaneously and propose a novel framework NCSSN for node classification in signed social networks. Experimental results on real-world signed social network datasets demonstrate the effectiveness of the proposed framework NCSSN. Further experiments are conducted to gain a deeper understanding of the importance of negative links for NCSSN.

Jiliang Tang
Yahoo! Labs
jlt@yahoo-inc.com

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

Huan Liu
Arizona State University
huan.liu@asu.edu

CP3

Vocal Competence Based Karaoke Recommendation: A Maximum-Margin Joint Model

in this paper, we develop a karaoke recommender system by incorporating vocal competence. We extract vocal (i.e., pitch, volume, and rhythm) ratings of a user for a song from his/her singing records. Then, we propose a joint modeling method named CBNTF by exploiting the mutual enhancement between non-negative tensor factorization (NTF) and support vector machine (SVM).

Chu Guan
University of Science and Technology of China
guanchu@mail.ustc.edu.cn

Yanjie Fu
Rutgers University
yanjie.fu@rutgers.edu

Xinjiang Lu
Northwestern Polytechnical University
xjlu@mail.nwpu.edu.cn

Hui Xiong
Rutgers, the State University of New Jersey
hxiong@rutgers.edu

Enhong Chen, Yingling Liu
University of Science and Technology of China
cheneh@ustc.edu.cn, ylliu22@ustc.edu.cn

CP3

Top-N Recommendation with Novel Rank Approximation

The importance of accurate recommender systems has been widely recognized by academia and industry. However, the recommendation quality is still rather low. Recently, a linear sparse and low-rank representation of the user-item matrix has been applied to produce Top-N recommendations. This approach uses the nuclear norm as a convex relaxation for the rank function and has achieved better recommendation accuracy than the state-of-the-art methods. In the past several years, solving rank minimization problems by leveraging nonconvex relaxations has received increasing attention. Some empirical results demonstrate that it can provide a better approximation to original problems than convex relaxation. In this paper, we propose a novel rank approximation to enhance the performance

of Top-N recommendation systems, where the approximation error is controllable. Experimental results on real data show that the proposed rank approximation improves the Top-N recommendation accuracy substantially.

Zhao Kang, Qiang Cheng
Southern Illinois University Carbondale
zhao.kang@siu.edu, qcheng@cs.siu.edu

CP3

A Confidence-Based Approach for Balancing Fairness and Accuracy

We study three classical machine learning algorithms in the context of fairness: AdaBoost, SVM, and logistic regression. Our goal is to maintain classification accuracy while reducing the degree to which these algorithms discriminate against a protected group. We introduce a method for achieving fairness by shifting the decision boundary for the protected group. Our method outperforms most previous algorithms in terms of accuracy and low discrimination, while simultaneously allowing for a fast and transparent quantification of the trade-off between bias and error.

Benjamin Fish, Jeremy Kun, Adam D. Lelkes
University of Illinois at Chicago
bfish3@uic.edu, jkun2@uic.edu, alelke2@uic.edu

CP3

A Spatial-Temporal Probabilistic Matrix Factorization Model for Point-of-Interest Recommendation

Most of approaches do not successfully incorporate both geographical influence and temporal effect together into latent factor models for POI recommendation. Hence, we propose a new Spatial-Temporal Probabilistic Matrix Factorization (STPMF) model that models a user's preference for POI as the combination of his geographical preference and other general interest in POI. The temporal dynamics of user's interest is also captured by modeling checkin data in a unique way. The experimental demonstrates the models effectiveness.

Huayu Li
University of North Carolina at Charlotte
hli38@uncc.edu

Richang Hong
Hefei University of Technology
hongrc@hfut.edu.cn

Zhiang Wu
Nanjing University of Finance and Economics
zawu@seu.edu.cn

Yong Ge
UNC Charlotte
yong.ge@uncc.edu

CP3

Differentially Private Significance Testing on Paired-Sample Data

Rigorous data mining results require measures of the statistical significance of the outcomes. The complexity of the data and models makes this a challenge; methods to protect privacy further complicate the issue. We demonstrate how to estimate statistical significance of results in paired-sample data (i.e., with "before' and "after' measurements); the impact of the noise required to provide differential privacy is included in the significance measure. As a result, providing privacy does not complicate the use of the analysis.

Christine M. Task
Knexus Research
christinemarietask@gmail.com

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

CP3

Synergies That Matter: Efficient Interaction Selection Via Sparse Factorization Machine

Collaborative filtering has been widely used in modern recommender systems to provide accurate recommendations by leveraging historical interactions between users and items. The presence of cold-start items and users has imposed a huge challenge to recommender systems based on collaborative filtering, because of the unavailability of such interaction information. The factorization machine is a powerful tool designed to tackle the cold-start problems by learning a bilinear ranking model that utilizes content information about users and items, exploiting the interactions with such content information. While a factorization machine makes use of all possible interactions between all content features to make recommendations, many of the features and their interactions are not predictive of recommendations, and incorporating them in the model will deteriorate the generalization performance of the recommender systems. In this paper, we propose an efficient Sparse Factorization Machine (SFM), that simultaneously identifies relevant user and item content features, models interactions between these relevant features, and learns a bilinear model using only these synergistic interactions. We have carried out extensive empirical studies on both synthetic and real-world datasets, and compared our method to other state-of-the-art baselines, including Factorization Machine. Experimental results show that SFM can greatly outperform other baselines.

Jianpeng Xu, Kaixiang Lin, Pang-Ning Tan
Computer Science and Engineering Department
Michigan State University
xujianpe@msu.edu, linkaixi@msu.edu, ptan@msu.edu

Jiayu Zhou
MSU
jiayuz@msu.edu

CP4

A General Framework to Increase the Robustness of Model-Based Change Point Detection Algorithms to Outliers and Noise

The autonomous identification of time-steps where the behavior of a time-series significantly deviates from a predefined model, or time-series change point detection, is an active field of research with notable applications in finance, health, and advertising. One family of time-series change detection algorithms, referred to as "model-based methods', although useful for many applications, performs poor when the data are noisy and have outliers. We introduce a new framework that enables existing model-based methods

to be more robust to these data challenges. We demonstrate the effectiveness of our approach on remote sensing and mobile health data. Our method introduces two new concepts: (i) a random sampling procedure allows us to overcome outliers, and (ii) a matrix-based representation of anomaly scores provides a flexible and intuitive way to identify multiple types of changes and test their significance. We show that our method performs better than several baseline methods, including application-specific algorithms, and provide all data and open-source code.

XI Chen
University of Minnesota
Department of Computer Science and Engineering
chen@cs.umn.edu

Yuanshun Yao
Univerisity of California, Santa Barbara
Department of Computer Science
yao@cs.ucsb.edu

Sichao Shi
Univerisity of California, San Diego
Department of Computer Science
sis031@ucsd.edu

Snigdhansu Chatterjee
University of Minnesota
Department of Statistics
chatt019@umn.edu

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

James Faghmous
Icahn School of Medicine at Mount Sinai
Department of Population Health Science and Policy
james.faghmous@mssm.edu

CP4

R1stm: One-Class Support Tensor Machine with Randomised Kernel

Identifying unusual or anomalous patterns in an underlying dataset is an important but challenging task in many applications. The focus of the unsupervised anomaly detection literature has mostly been on vectorised data. However, many applications are more naturally described using higher-order tensor representations. Approaches that vectorise tensorial data can destroy the structural information encoded in the high-dimensional space, and lead to the problem of the curse of dimensionality. In this paper we present the first unsupervised tensorial anomaly detection method, along with a randomised version of our method. Our anomaly detection method, the One-class Support Tensor Machine (1STM), is a generalisation of conventional one-class Support Vector Machines to higher-order spaces. 1STM preserves the multiway structure of tensor data, while achieving significant improvement in accuracy and efficiency over conventional vectorised methods. We then leverage the theory of nonlinear random projections to propose the Randomised 1STM (R1STM). Our empirical analysis on several real and synthetic datasets shows that our R1STM algorithm delivers comparable or better accuracy to a state-of-the-art deep learning method and traditional kernelised approaches for anomaly detection, while being approximately 100 times faster in training and testing.

Sarah M. Erfani
The University of Melbourne
The University of Melbourne
sarah.erfani@unimelb.edu.au

CP4

Scalable Anomaly Ranking of Attributed Neighborhoods

In this work we propose normality, a new quality measure for attributed neighborhoods. Normality utilizes structure and attributes together to quantify both internal consistency and external separability. It exhibits two key advantages over other measures: (1) It allows many boundary-edges as long as they can be "exonerated"; i.e., either (i) are expected under a null model, and/or (ii) the boundary nodes do not exhibit the subset of attributes shared by the neighborhood members. Existing measures, in contrast, penalize boundary edges irrespectively. (2) Normality can be efficiently maximized to automatically infer the shared attribute subspace (and respective weights) that characterize a neighborhood. This efficient optimization allows us to process graphs with millions of attributes. We capitalize on our measure to present a novel approach for Anomaly Mining of Entity Neighborhoods (AMEN). Experiments on real-world attributed graphs illustrate the effectiveness of our measure at anomaly detection, outperforming popular approaches including conductance, density, OddBall, and SODA. In addition to anomaly detection, our qualitative analysis demonstrates the utility of normality as a powerful tool to contrast the correlation between structure and attributes across different graphs.

Bryan Perozzi, Leman Akoglu
Stony Brook University
bperozzi@cs.stonybrook.edu, leman@cs.stonybrook.edu

CP4

Routine Mining Based Anomaly Detection in Mobile Phone Data

Previous works related to anomaly detection in mobile phone data rely heavily on manually selected features or statistics, which weakens the generalization of the model. In addition, traditional methods only allow anomaly to appear in certain fixed time interval with predefined duration. In this work, based on an unsupervised probabilistic topic model, we propose a Routine Mining Based Anomaly Detection (RMBAD) approach that learns the pattern of *normal* from naturally existing human routines, free of any manually extracted features. Taking a generative approach, the RMBAD model combines group anomaly detection and topic segmentation method, segmenting mobile subscriber's sequence of behaviors into explainable routines and report those unexplainable segments as anomalies simultaneously. Furthermore, The RMBAD model allows routines of different durations to coexist, thus achieving a more realistic modeling of human activity pattern, which ultimately leads to higher anomaly detection accuracy. Extensive experiments are conducted on both synthetic and real world mobile datasets, and the empirical results show that the RMBAD model can effectively discover hidden routines of human activity and identify those groups of behaviors that collectively appear anomalous.

Tian Qin, Guojie Song, Sizhen Du
Peking University

qintian@pku.edu.cn, gjsong@pku.edu.cn, dusizhen@126.com

CP4

A Scalable Approach for Outlier Detection in Edge Streams Using Sketch-Based Approximations

We present a method for performing outlier detection in graph edge streams. Sketches are employed to retain an approximate graph model, which enables provable error bounds on our outlier scoring functions. Experimental results show our method is scalable, with updates done in constant time, and outperforms the baseline.

Stephen Ranshous, Steve Harenberg, Kshitij Sharma
North Carolina State University
smransho@ncsu.edu, sdharenb@ncsu.edu, ksharma3@ncsu.edu

Nagiza Samatova
North Carolina State University
Oak Ridge National Laboratory
samatova@csc.ncsu.edu

CP4

LODES: Local Density Meets Spectral Outlier Detection

The problem of outlier detection has been widely studied in existing literature because of its numerous applications in fraud detection, medical diagnostics, fault detection, and intrusion detection. A large category of outlier analysis algorithms have been proposed, such as proximity-based methods and local density-based methods. These methods are effective in finding outliers distributed along linear manifolds. Spectral methods, however, are particularly well suited to finding outliers when the data is distributed along manifolds of arbitrary shape. In practice, the underlying manifolds may have varying density, as a result of which a direct use of spectral methods may not be effective. In this paper, we show how to combine spectral techniques with local density-based methods in order to discover interesting outliers. We present experimental results demonstrating the effectiveness of our approach with respect to well-known competing methods.

Saket Sathe
IBM, Australia
ssathe@us.ibm.com

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

CP5

Tersesvm : A Scalable Approach for Learning Compact Models in Large-Scale Classification

For large-scale multi-class classification problems, consisting of tens of thousand target categories, recent works have emphasized the need to store billions of parameters. For instance, the classical l_2-norm regularization employed by a state-of-the-art method results in the model size of 17GB for a training set whose size is only 129MB. To the contrary, by using a mixed-norm regularization approach, we show that around 99.5% of the stored parameters is dispensable noise. Using this strategy, we can extract the information relevant for classification, which is constituted in remaining 0.5% of the parameters, and hence demonstrate drastic reduction in model sizes. Furthermore, the proposed method leads to improvement in generalization performance compared to state-of-the-art methods, especially for under-represented categories. Lastly, our method enjoys easy parallelization, and scales well to tens of thousand target categories.

Rohit Babbar
Max-Planck Institute for Intelligent Systems
Tuebingen, Germany
rohit.babbar@tuebingen.mpg.de

Krikamol Muandet, Bernhard Schölkopf
MPI for Intelligent Systems
krikamol@tuebingen.mpg.de, bs@tuebingen.mpg.de

CP5

Pattern Aided Classification

This paper defines a pattern aided classifier (PXC) using several pattern and group-specific-classifier pairs. Importantly, a PXC applies a group-specific classifier only to instances satisfying its associated pattern. The paper also introduces a contrast pattern based algorithm (CPXC) to learn accurate PXCs. Experiments show that CPXC often significantly outperforms traditional classification algorithms. The paper also introduces the opportunity-guided boosting technique and the conditional classifier ensemble concept.

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

Vahid Taslimitehrani
Wright State University
taslimitehrani.2@wright.edu

CP5

Linear and Kernel Classification: When to Use Which?

Kernel methods are known to be a state-of-the-art classification technique. Nevertheless, the training and prediction cost is expensive for large data. On the other hand, linear classifiers can easily scale up, but are inferior to kernel classifiers in terms of predictability. Recent research has shown that for some data sets (e.g., document data), linear is as good as kernel classifiers. In such cases, the training of a kernel classifier is a waste of both time and memory. In this work, we investigate the important issue of efficiently and automatically deciding whether kernel classifiers perform strictly better than linear for a given data set. Our proposed method is based on cheaply constructing a classifier that exhibits nonlinearity and can be automatically trained. Then we make a decision by comparing the performance of our constructed classifier with the linear classifier. We propose two methods: the first one trains the degree-2 feature expansion by a linear-classification method, while the second dissects the feature space into several regions and trains a linear classifier for each region. The design considerations of our methods are very different from past works for speeding up the kernel training. They still aim at obtaining accuracy close to the kernel classifier, but ours would like to give a quick and accurate decision without

worrying about accuracy.

Hsin-Yuan Huang
National Taiwan University
momohuang@gmail.com

CP5

Binary Classifier Calibration Using an Ensemble of Linear Trend Estimation

Learning accurate probabilistic models from data is crucial in many practical tasks in data mining. In this paper we present a new non-parametric calibration method called *ensemble of linear trend estimation* (ELiTE). ELiTE utilizes the recently proposed ℓ_1 trend filtering signal approximation method to find the mapping from uncalibrated classification scores to the calibrated probability estimates. ELiTE is designed to address the key limitations of the histogram binning-based calibration methods which are (1) the use of a piecewise constant form of the calibration mapping using bins, and (2) the assumption of independence of predicted probabilities for the instances that are located in different bins. The method post-processes the output of a binary classifier to obtain calibrated probabilities. Thus, it can be applied with many existing classification models. We demonstrate the performance of ELiTE on real datasets for commonly used binary classification models. Experimental results show that the method outperforms several common binary-classifier calibration methods. In particular, ELiTE commonly performs statistically significantly better than the other methods, and never worse. Moreover, it is able to improve the calibration power of classifiers, while retaining their discrimination power. The method is also computationally tractable for large scale datasets, as it is practically $O(N \log N)$ time, where N is the number of samples.

Mahdi Pakdaman Naeini
Intelligent Systems Program
University of Pittsburgh
pakdaman@cs.pitt.edu

Gregory Cooper
Department of Biomedical Informatics
University of Pittsburgh
gfc@pitt.edu

CP5

Discriminative Training of Structured Dictionaries Via Block Orthogonal Matching Pursuit

It is well established that high-level representations learned via sparse coding are effective for many machine learning applications such as denoising and classification. In addition to being reconstructive, sparse representations that are discriminative and invariant can further help with such applications. In order to achieve these desired properties, this paper proposes a new framework that discriminatively trains structured dictionaries via block orthogonal matching pursuit. Specifically, the dictionary atoms are assumed to be organized into blocks. Distinct classes correspond to distinct blocks of dictionary atoms; however, our algorithm can handle the case where multiple classes share blocks. We provide theoretical justification and empirical evaluation of our method.

Wenling Shang
University of Michigan
shangw@umich.edu

CP5

A Unified View of Localized Kernel Learning

Multiple Kernel Learning, or MKL, extends (kernelized) SVM by attempting to learn not only a classifier/regressor but also the best kernel for the training task, usually from a combination of existing kernel functions. Most MKL methods seek the combined kernel that performs best over *every* training example, sacrificing performance in some areas to seek a global optimum. *Localized* kernel learning (LKL) overcomes this limitation by allowing the training algorithm to match a component kernel to the examples that can exploit it best. Several approaches to the localized kernel learning problem have been explored in the last several years. We unify many of these approaches under one simple system and design a new algorithm with improved performance. We also develop enhanced versions of existing algorithms, with an eye on scalability and performance.

John Moeller, Sarathkrishna Swaminathan, Suresh Venkatasubramanian
University of Utah
moeller@cs.utah.edu, sarath@cs.utah.edu, suresh@cs.utah.edu

CP6

Power Simultaneous Spectral Data Embedding and Clustering

Spectral clustering methods use the Laplacian eigenvalues and eigenvectors to obtain a low-dimensional embedding that can be trivially clustered. For that purpose, spectral clustering is often based on a tandem approach where the two steps: affinity matrix eigendecomposition and k-means clustering, are performed separately. The potential flaw of such common practice is that the obtained relaxed continuous spectral solution can severely deviate from the true discrete clustering solution. Given the high computational cost of such spectral clustering methods, this paper provides the so-called PSDEC framework. PSDEC performs simultaneously the eigendecomposition of the affinity matrix and clustering tasks, and uses the Power method to speed up the unified process convergence. In PSDEC, the selected top eigenvectors of the Laplacian matrix can be of help in detecting a cluster structure of objects and providing simpler and more interpretable solutions. We show that by doing so, our method can learn low-dimensional representations that are better suited to clustering, outperforming not only spectral clustering algorithms but also some NMF variants.

Kais Allab, Lazhar Labiod, Mohamed Nadif
Paris Descartes University
allab.kais@gmail.com, lazhar.labiod@parisdescartes.fr, mohamed.nadif@parisdescartes.fr

CP6

Lagrangian Constrained Clustering

ncorporating background knowledge in clustering problems has attracted much interest in recent years. The knowledge can be represented as pairwise instance-level constraints. There exist applications and scenarios of constrained clustering where satisfying (almost) all of the constraints is required. However, achieving such clusters in a reasonable time is still a challenging problem. This paper presents a

new Lagrangian Constrained Clustering framework (called LCC) for clustering in presence of pairwise constraints. LCC is an iterative optimization procedure which adds penalties for violated constraints and increases the penalty if constraints remain violated from previous iterations to ensure satisfying (almost) all of the constraints. Experiments on UCI data sets shows LCC outperforms other constrained clustering algorithms in scenarios which satisfying (almost) all of the constraints is desired.

Mohadeseh Ganji
University of Melbourne
sghasempour@student.unimelb.edu.au

CP6

Fast Multiplier Methods to Optimize Non-Exhaustive, Overlapping Clustering

To accelerate the convergence of an augmented Lagrangian scheme for optimizing the low rank semidefinite objective of non-exhaustive, overlapping clustering, we consider two fast multiplier methods : a proximal method of multipliers and an alternating direction method of multipliers (ADMM). For the proximal method of multipliers, we show a convergence result for the non-convex case with bound-constrained subproblems. These methods are up to 13 times faster with no change in quality compared with a standard augmented Lagrangian method.

Yangyang Hou
Purdue University
hou13@purdue.edu

Joyce Whang
Sungkyunkwan University
jjwhang@skku.edu

David F. Gleich
Purdue University
dgleich@purdue.edu

Inderjit S. Dhillon
University of Texas at Austin
inderjit@cs.utexas.edu

CP6

Process Trace Clustering: A Heterogeneous Information Network Approach

Process mining is the task of extracting information from event logs, such as ones generated from workflow management or enterprise resource planning systems, in order to discover models of the underlying processes, organizations, and products. As the event logs often contain a variety of process executions, the discovered models can be complex and difficult to comprehend. Trace clustering helps solve this problem by splitting the event logs into smaller subsets and applying process discovery algorithms on each subset, resulting in per-subset discovered processes that are less complex and more accurate. However, the state-of-the-art clustering techniques are limited: the similarity measures are not process-aware and they do not scale well to high-dimensional event logs. In this paper, we propose a conceptualization of process's event logs as a heterogeneous information network, in order to capture the rich semantic meaning, and thereby derive better process-specific features. In addition, we propose SeqPathSim, a meta path-based similarity measure that considers node sequences in the heterogeneous graph and results in better clustering. We also introduce a new dimension reduction method that combines event similarity with regularization by process model structure to deal with event logs of high dimensionality. The experimental results show that our proposed approach outperforms state-of-the-art trace clustering approaches in both accuracy and structural complexity metrics.

Phuong Nguyen
University of Illinois at Urbana-Champaign
pvnguye2@illinois.edu

Aleksander Slominski, Vinod Muthusamy
IBM T.J. Watson Research Center
aslom@us.ibm.com, vmuthus@us.ibm.com

Vatche Ishakian
IBM T J Watson Research Center
vishaki@us.ibm.com

Klara Nahrstedt
University of Illinois at Urbana-Champaign
klara@illinois.edu

CP6

Stochastic Co-Clustering for Document-Term Data

Co-clustering is more useful than one-sided clustering when dealing with high dimensional sparse data. We propose to address the aim of document clustering with a generative model-based co-clustering approach. To this end, we rely on a particular mixture of von Mises-Fisher distributions and propose a new parsimonious model allowing to reveal a block diagonal structure as well as a good partitioning of documents and terms. Then, by setting the estimate of the model parameters under the maximum likelihood (ML) approach, we derive three novel co-clustering algorithms: a soft one and two stochastic variants. Empirical results on numerous simulated and real-world datasets, demonstrate the advantages of our approach to model and co-cluster high dimensional sparse data.

Aghiles Salah, Nicoleta Rogovschi
University of Paris Descartes
aghiles-salah@hotmail.fr,
nicoleta.rogovschi@parisdescartes.fr

Mohamed Nadif
Paris Descartes University
mohamed.nadif@parisdescartes.fr

CP6

On Finding the Maximum Edge Biclique in a Bipartite Graph: A Subspace Clustering Approach

Bipartite graphs have been proven useful in modeling a wide range of relationship networks. Finding the maximum edge biclique within a bipartite graph is a well-known problem. We propose a probabilistic algorithm for finding the maximum edge biclique. Extensive experimentation shows that the algorithm is significantly better than the state-of-the-art technique. We prove that there are solid theoretical reasons for the algorithms efficacy that manifest in a polynomial complexity of time and space.

Eran Shaham, Honghai Yu
Institute for Infocomm Research (I2R), A*STAR, Singapore
eran.shaham@gmail.com, yuhh@i2r.a-star.edu.sg

Xiao-Li Li
Institute for Infocomm Research, Singapore
xlli@i2r.a-star.edu.sg

CP7

k-Means for Streaming and Distributed Big Sparse Data

We provide the first streaming algorithm for computing a provable approximation to the k-means of sparse Big data. Our main technical result is a deterministic algorithm for computing a sparse (k, ϵ)-coreset, which is a weighted subset of $k^{O(1)}$ input points that approximates the sum of squared distances from the n input points to every set of k centers, up to $(1 \pm \epsilon)$ factor, for any given constant $\epsilon > 0$. This is the first such coreset of size independent of both d and n.

Artem Barger, Danny Feldman
Haifa University
artem@bargr.net, dannyf.post@gmail.com

CP7

Pivot-Based K-Means Algorithm for Numerous-Class Data Sets

This paper presents an accelerated k-means clustering algorithm suitable for a large-scale and numerous-class data set. The proposed iterative algorithm avoids unnecessary exact distance calculations, especially in the early and the last stage in its convergence process, keeping the same result as the standard algorithm. This is efficiently performed by two distinct components. One uses the lower bounds on exact distances between objects and centroids for the acceleration in the early stage. The lower bound is calculated by the triangle inequality with newly introduced fixed points called pivots. The other for the last stage skips a distance calculation between an invariant centroid and an object satisfying the condition of the cluster to which the object is assigned remains unchanged, i.e., its centroid is also invariant. For much further speed-up, we can incorporate an existing algorithm, which works well in the middle stage, into the proposed algorithm as its extension. The experimental results on large-scale and high-dimensional image data sets demonstrate that given a large k value, the proposed algorithm outperforms the existing algorithms regarding both the reduction rate of distance calculations and elapsed time.

Takashi Hattori
NTT Corporation
hattori.takashi@lab.ntt.co.jp

CP7

Geometric Methods to Accelerate k-Means Algorithms

Most implementations of k-means use Lloyd's algorithm, which does many unnecessary distance calculations. Several accelerated algorithms produce exactly the same answer. They avoid redundant work using the triangle inequality paired with a set of bounds on point-centroid distances. Our proposals allow those algorithms to perform even better, giving up to eight times further speedup. Our methods give tighter lower bound updates and efficiently skip centroids that cannot possibly be close to a set of points.

Petr Ryav
Baylor University
petr_rysavy@alumni.baylor.edu

Greg Hamerly
Baylor University
Department of Computer Science
greg_hamerly@baylor.edu

CP8

$Halite_{ds}$: Fast and Scalable Subspace Clustering for Multidimensional Data Streams

Given a data stream with many attributes and high frequency of events, how to cluster similar events? Can it be done in real time? For example, how to cluster decades of frequent measurements of tens of climatic attributes to aid real time alert systems in forecasting extreme climatic events, such as hurricanes? The task of clustering data with many attributes is known as subspace clustering. Today, there exists a need for algorithms of this type well-suited to streams, for which real time processing is highly desirable. This paper proposes the new algorithm $Halite_{ds}$ - a fast, scalable and highly accurate subspace clustering algorithm for multidimensional streams. It improves upon a technique that was originally designed to process static (not streams) data. Our contributions are: (1) Analysis of Streams: the new algorithm takes advantage of the knowledge obtained from clustering past data to easy clustering data in the present. This fact allows our $Halite_{ds}$ to be considerably faster than its base algorithm, yet obtaining top-quality results; (2) Real Time Processing: as opposed to the state-of-the-art, $Halite_{ds}$ is fast and scalable, making it feasible to analyze streams with many attributes and high frequency of events in real time; (3) Experiments: we ran experiments on datasets including a real stream with almost one century of climatic data. Our $Halite_{ds}$ was up to 217 times faster than 5 representative works, always presenting top-quality results.

Afonso E. Da Silva, LUCAS L. Sanches, ANTONIO C. Fraideinberze
University of São Paulo, Brazil
aexpedito@grad.icmc.usp.br,
lucas.lancellotti.sanches@usp.br, antoniocf@usp.br

Robson L. F. Cordeiro
University of São Paulo
robson@icmc.usp.br

CP8

Online Clustering of Multivariate Time-Series

The intrinsic nature of streaming data requires algorithms that are capable of fast data analysis to extract knowledge. Most current unsupervised data analysis techniques rely on the implementation of known batch techniques over a sliding window, which can hinder their utility for the analysis of evolving structure in applications involving large streams of data. This research presents a novel data clustering algorithm, which exploits the correlation between data points in time to cluster the data, while maintaining a set of decision boundaries to identify noisy or anomalous data. We illustrate the proposed algorithm for online clustering with numerical results on both real-life and simulated datasets, which demonstrate the efficiency and accuracy of our ap-

proach compared to existing methods.

Masud Moshtaghi, Christopher Leckie, James Bezdek
The University of Melbourne
masud.moshtgahi@unimelb.edu.au,
caleckie@unimelb.edu.au, jcbezdek@gmail.com

CP8

Learning A Task-Specific Deep Architecture For Clustering

While sparse coding-based clustering methods have shown to be successful, their bottlenecks in both efficiency and scalability limit the practical usage. In recent years, deep learning has been proved to be a highly effective, efficient and scalable feature learning tool. In this paper, we propose to emulate the sparse coding-based clustering pipeline in the context of deep learning, leading to a carefully crafted deep model benefiting from both. A feed-forward network structure, named TAGnet, is constructed based on a graph-regularized sparse coding algorithm. It is then trained with task-specific loss functions from end to end. We discover that connecting deep learning to sparse coding benefits not only the model performance, but also its initialization and interpretation. Moreover, by introducing auxiliary clustering tasks to the intermediate feature hierarchy, we formulate DTAGnet and obtain a further performance boost. Extensive experiments demonstrate that the proposed model gains remarkable margins over several state-of-the-art methods.

Zhangyang Wang, Shiyu Chang
UIUC
masterwant@gmail.com, chang87@illinois.edu

Jiayu Zhou
MSU
jiayuz@msu.edu

Meng Wang
HFUT
wangmeng@hfut.edu.cn

Thomas Huang
UIUC
t-huang1@illinois.edu

CP9

K-Nearest Neighbor Search and Outlier Detection Via Minimax Distances

We study Minimax distance measures for K-nearest neighbor search and classification. Recently, the use of this distance measure is shown to improve the K-nearest neighbor classification results. We consider the computational aspects of this problem and propose an *efficient* and *general-purpose* algorithm for computing Minimax neighbors which requires a significantly lower runtime and is applicable with any arbitrary distance measure. We study the computational optimality of our approach and its connection to the Prim's algorithm, and then, generalize our analysis to computing *one-to-all* Minimax distances. In the following, we investigate in detail the edges selected by Minimax distances and thereby explore the ability of Minimax distances in detecting outlier objects. We evaluate the performance of our methods on a variety of real-world datasets, e.g. text documents and images.

Morteza Haghir Chehreghani
Xerox Research Centre Europe - XRCE
morteza.chehreghani@gmail.com

CP9

Euclidean Co-Embedding of Ordinal Data for Multi-Type Visualization

Embedding deals with reducing the high-dimensional representation of data into a low-dimensional representation. Previous work mostly focuses on preserving similarities among objects. Here, not only do we explicitly recognize multiple types of objects, but we also focus on the ordinal relationships across types. Collaborative Ordinal Embedding or COE is based on generative modelling of ordinal triples. Experiments show that COE outperforms the baselines on objective metrics, revealing its capacity for information preservation for ordinal data.

Dung D. Le
School of Information Systems,
Singapore Management University, Singapore
ddle.2015@phdis.smu.edu.sg

Hady W. Lauw
Singapore Management University
hadywlauw@smu.edu.sg

CP9

Robust Unsupervised Feature Selection on Networked Data

Feature selection has shown its effectiveness to prepare high-dimensional data for many data mining and machine learning tasks. Traditional feature selection algorithms are mainly based on the assumption that data instances are independent and identically distributed. However, this assumption is invalid in networked data since instances are not only associated with high dimensional features but also inherently interconnected with each other. In addition, obtaining label information for networked data is time consuming and labor intensive. Without label information to direct feature selection, it is difficult to assess the feature relevance. In contrast to the scarce label information, link information in networks are abundant and could help select relevant features. However, most networked data has a lot of noisy links, resulting in the feature selection algorithms to be less effective. To address the above mentioned issues, we propose a robust unsupervised feature selection framework NetFS for networked data, which embeds the latent representation learning into feature selection. Therefore, content information is able to help mitigate the negative effects from noisy links in learning latent representations, while good latent representations in turn can contribute to extract more meaningful features. In other words, both phases could cooperate and boost each other. Experimental results on real-world datasets demonstrate the effectiveness of the proposed framework.

Jundong Li
Arizona State University
jundongl@asu.edu

Xia Hu
Texas A&M University
hu@cse.tamu.edu

Liang Wu, Huan Liu
Arizona State University

wuliang@asu.edu, huan.liu@asu.edu

CP9

Kernelized Matrix Factorization for Collaborative Filtering

Matrix factorization (MF) methods have shown great promise in collaborative filtering (CF). Conventional MF methods usually assume that the correlated data is distributed on a linear hyperplane, which is not always the case. Kernel methods are used widely in SVMs to classify linearly non-separable data, as well as in PCA to discover the non-linear embeddings of data. In this paper, we present a novel method to kernelize matrix factorization for collaborative filtering, which is equivalent to performing the low-rank matrix factorization in a possibly much higher dimensional space that is implicitly defined by the kernel function. Inspired by the success of multiple kernel learning (MKL) methods, we also explore the approach of learning multiple kernels from the rating matrix to further improve the accuracy of prediction. Since the right choice of kernel is usually unknown, our proposed multiple kernel matrix factorization method helps to select effective kernel functions from the candidates. Through extensive experiments on real-world datasets, we show that our proposed method captures the nonlinear correlations among data, which results in improved prediction accuracy compared to the state-of-art CF models.

Xinyue Liu
Worcester Polytechnic Institute
xliu4@wpi.edu

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

Yu-Feng Li
Nanjing University, China
liyf@nju.edu.cn

Xiangnan Kong, Xinyuan Sun
Worcester Polytechnic Institute
xkong@wpi.edu, xsun2@wpi.edu

Saket Sathe
IBM, Australia
ssathe@us.ibm.com

CP9

A Framework to Adjust Dependency Measure Estimates for Chance

Estimating the strength of dependency between two variables is fundamental for exploratory analysis and many other applications in data mining. For example: non-linear dependencies between two continuous variables can be explored with the Maximal Information Coefficient (MIC); and categorical variables that are dependent to the target class are selected using Gini gain in random forests. Nonetheless, because dependency measures are estimated on finite samples, the interpretability of their quantification and the accuracy when ranking dependencies become challenging. In this paper, we propose a framework to adjust dependency measure estimates on finite samples. We demonstrate that our approach enhances the interpretability of MIC when used as a proxy for the amount of noise between variables, and to gain accuracy when ranking variables during the splitting procedure in random forests.

Simone Romano, Vinh Nguyen, James Bailey, Karin Verspoor
The University of Melbourne
me@simoneromano.com, vinh.nguyen@unimelb.edu.au, baileyj@unimelb.edu.au, karin.verspoor@unimelb.edu.au

CP9

Nonlinear Joint Unsupervised Feature Selection

In the era of big data, one is often confronted with the problem of high dimensional data for many machine learning or data mining tasks. Feature selection, as a dimension reduction technique, is useful for alleviating the curse of dimensionality while preserving interpretability. In this paper, we focus on unsupervised feature selection, as class labels are usually expensive to obtain. Unsupervised feature selection is typically more challenging than its supervised counterpart due to the lack of guidance from class labels. Recently, regression-based methods with $L_{2,1}$ norms have gained much popularity as they are able to evaluate features jointly which, however, consider only linear correlations between features and pseudo-labels. In this paper, we propose a novel nonlinear joint unsupervised feature selection method based on kernel alignment. The aim is to find a succinct set of features that best aligns with the original features in the kernel space. It can evaluate features jointly in a nonlinear manner and provides a good '0/1' approximation for the selection indicator vector. We formulate it as a constrained optimization problem and develop a Spectral Projected Gradient (SPG) method to solve the optimization problem. Experimental results on several real-world datasets demonstrate that our proposed method outperforms the state-of-the-art approaches significantly.

Xiaokai Wei, Bokai Cao
University of Illinois at Chicago
xwei2@uic.edu, caobokai@uic.edu

Philip S. Yu
University of Illinois at Chicago
Chicago, USA
psyu@uic.edu

CP10

Risk Prediction with Electronic Health Records: A Deep Learning Approach

Electronic Health Records (EHR) is one of the major carriers for make this data-driven healthcare revolution successful. There are many challenges on working directly with EHR, such as temporality, sparsity, noisiness, bias, etc. Thus effective feature extraction, or phenotyping from patient EHRs is a key step before any further applications. In this paper, we propose a deep learning approach for phenotyping from patient EHRs. We first represent the EHRs for every patient as a temporal matrix with time on one dimension and event on the other dimension. Then we build a four-layer convolutional neural network model for extracting phenotypes and perform prediction. The first layer is composed of those EHR matrices. The second layer is a one-side convolution layer that can extract phenotypes from the first layer. The third layer is a max pooling layer introducing sparsity on the detected phenotypes, so that only those significant phenotypes will remain. The fourth layer is a fully connected softmax prediction layer. In order to incorporate the temporal smoothness of the patient EHR, we also investigated three different temporal fusion

mechanisms in the model: early fusion, late fusion and slow fusion. Finally the proposed model is validated on a real world EHR data warehouse under the specific scenario of predictive modeling of chronic diseases.

Yu Cheng
IBM T.J. Watson Research Center
chengyu@us.ibm.com

CP10

Uncovering Latent Behaviors in Ant Colonies

Many biological systems have collective behaviors that endow them with immense adaptive advantages compared to more solitary species. Describing these behaviors is challenging yet necessary in order to understand these biological systems. We propose a probabilistic model that enables us to uncover the collective behaviors observed in a colony of ants. This model is based on the assumption that the be- havior of an individual ant is a time-dependent mixture of latent behaviors that are specific to the whole colony. We apply this model to a large-scale dataset obtained by ob- serving the mobility of nearly 1000 Camponotus fellah ants from six different colonies. Our results indicate that a colony typically exhibits three classes of behaviors, each character- ized by a specific spatial distribution and a level of activity. Moreover, these spatial distributions, which are uncovered automatically by our model, match well with the ground truth as manually annotated by domain experts. We fur- ther explore the evolution of the behavior of individual ants and show that it is well captured by a second order Markov chain that encodes the fact that the future behavior of an ant depends not only on its current behavior but also on its preceding one.

Mohamed Kafsi
École Polytechnique Fédérale de Lausanne
mohamed.kafsi@epfl.ch

CP10

IPath: **Forecasting the Pathway to Impact**

Forecasting the success of scientific work has been attracting extensive research attention in the recent years. It is often of key importance to foresee the pathway to impact for scholarly entities for (1) tracking research frontier, (2) invoking an early intervention and (3) proactively allocating research resources. Many recent progresses have been seen in modeling the long-term scientific impact for *point prediction*. However, challenges still remain when it comes to *forecasting the impact pathway*. In this paper, we propose a novel predictive model to collectively achieve a set of design objectives to address these challenges, including prediction consistency and parameter smoothness. Extensive empirical evaluations on real scholarly data validate the effectiveness of the proposed model.

Liangyue Li, Hanghang Tong
Arizona State University
liangyue@asu.edu, hanghang.tong@asu.edu

Jie Tang
Tsinghua University
jietang@tsinghua.edu.cn

Wei Fan
Big Data Labs - Baidu USA
fanwei03@baidu.com

CP10

Predicting the Popularity of News Articles

Consuming news articles is an integral part of our daily lives and news agencies such as The Washington Post (WP) expend tremendous effort in providing high quality reading experiences for their readers. Journalists and editors are faced with the task of determining which articles will become popular so that they can efficiently allocate resources to support a better reading experience. The reasons behind the popularity of news articles are typically varied, and might involve contemporariness, writing quality, and other latent factors. In this paper, we cast the problem of popularity prediction problem as regression, engineer several classes of features (metadata, contextual or content-based, temporal, and social), and build models for forecasting popularity. The system presented here is deployed in a real setting at The Washington Post; we demonstrate that it is able to accurately predict article popularity with an $R^2 \approx 0.8$ using features harvested within 30 minutes of publication time.

Yaser Keneshloo
Virginia Tech
yaserkl@vt.edu

Shuguang Wang, Eui-Hong (Sam) Han
The Washington Post
shuguang.wang@washpost.com, sam.han@washpost.com

Naren Ramakrishnan
Virginia Tech
naren@vt.edu

CP10

Cost-Sensitive Batch Mode Active Learning: Designing Astronomical Observation by Optimizing Telescope Time and Telescope Choice

Astronomers and telescope operators must make decisions about what to observe given limited telescope time. To optimize this decision-making process, we present a batch, cost-sensitive, active learning approach that exploits structure in the unlabeled dataset, accounts for label uncertainty, and minimizes annotation costs. We first cluster the unlabeled instances in feature space. We next introduce an uncertainty-reducing selection criterion that encourages the batch of selected instances to span multiple clusters, in addition to taking into account annotation cost. Finally, we extend this criterion to incorporate the fact that nearby astronomical objects may be observed at the same time. On two large astronomical data sets, our approach balances the trade-offs among FOV, aperture, and time cost and, therefore, helps astronomers design effective experiments.

Xide Xia
Harvard University
Harvard University
xidexia@g.harvard.edu

Finale Doshi-Velez, Pavlos Protopapas
Harvard University

finale@seas.harvard.edu, pavlos@seas.harvard.edu

CP10

The Impact of Community Safety on House Ranking

Community safety has considerable impacts on housing investments. Housing investors can make more informed decisions if they are fully aware of safety related factors. We developed a safety-aware house ranking method by extracting and incorporating spatio-temporal community safety knowledge into house assessment. To test the proposed method, a comprehensive evaluation was conducted on real-world crime and house data.

Zijun Yao, Yanjie Fu, Bin Liu
Rutgers University
zijun.yao@rutgers.edu, yanjie.fu@rutgers.edu, binben.liu@rutgers.edu

Hui Xiong
Rutgers, the State University of New Jersey
hxiong@rutgers.edu

CP11

Distributed Representations of Expertise

Collaborative networks are common in real life, where domain experts work together to solve tasks issued by customers. How to model the proficiency of experts is critical for us to understand and optimize collaborative networks. Traditional expertise models, such as topic model based methods, cannot capture two aspects of human expertise simultaneously: Specialization (what area an expert is good at?) and Proficiency Level (to what degree?). Specifically, in our first model, we assume that each expert will only handle tasks whose difficulty level just matches his/her proficiency level, while experts in the second model accept tasks whose levels are equal to or lower than his/her proficiency level. Experiments on real world datasets show that both models outperform topic model based approaches and standard classifiers such as logistic regression and support vector machine in terms of prediction accuracy. The learnt vector representations can be used to compare expertise in a large organization and optimize expert allocation.

Fangqiu Han, Shulong Tan
University of California, Santa Barbara
fhan@cs.ucsb.edu, laos1984@gmail.com

Huan Sun
Computer Science Department
University of California, Santa Barbara
huansun@cs.ucsb.edu

Mudhakar Srivatsa
IBM Research
msrivats@us.ibm.com

Deng Cai
Zhejiang University
dengcai@gmail.com

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

CP11

Birdnest: Bayesian Inference for Ratings-Fraud Detection

Review fraud is a pervasive problem in online commerce, in which fraudulent sellers write or purchase fake reviews to manipulate perception of their products and services. Fake reviews are often detected based on several signs, including 1) they occur in short bursts of time; 2) fraudulent user accounts have skewed rating distributions. In this paper, we combine these 2 approaches in a principled manner, allowing successful detection even when one of these signs is not present. To combine these two approaches, we formulate our Bayesian Inference for Rating Data (BIRD) model, a flexible Bayesian model of user rating behavior. Based on our model we formulate a likelihood-based suspiciousness metric, Normalized Expected Surprise Total (NEST). We propose a linear-time algorithm for performing Bayesian inference using our model and computing the metric. Experiments on real data show that BIRDNEST successfully spots review fraud in large, real-world graphs: the 50 most suspicious users of the Flipkart platform flagged by our algorithm were investigated and all identified as fraudulent by domain experts at Flipkart.

Bryan Hooi, Neil Shah, Alex Beutel
Carnegie Mellon University
bhooi@andrew.cmu.edu, neilshah@cs.cmu.edu, abeutel@cs.cmu.edu

Stephan Gunneman
Technical University of Munich
guennemann@in.tum.de

Leman Akoglu
Stony Brook University
leman@cs.stonybrook.edu

Mohit Kumar, Disha Makhija
Flipkart
k.mohit@flipkart.com, disha.makhiji@flipkart.com

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

CP11

Query-Driven Maximum Quasi-Clique Search

Quasi-cliques are an elegant way to model dense subgraphs, with each node adjacent to at least a fraction $\lambda \in (0, 1]$ of other nodes in the subgraph. In this paper, we focus on a new graph mining problem, the *query-driven maximum quasi-clique* (QMQ) search, which aims to find the largest λ-quasi-clique containing a given query node set S. This problem has many applications and is proved to be NP-Hard and inapproximable. To solve the problem efficiently in practice, we propose the notion of *core tree* to organize dense subgraphs recursively, which reduces the search space and effectively helps find the solution within a few tree traversals. To optimize a solution to a better solution, we introduce three refinement operations: *Add*, *Remove* and *Swap*. We propose two iterative maximization algorithms, DIM and SUM, to approach QMQ by deterministic and stochastic means respectively. With extensive experiments on three real datasets, we demonstrate that our algo-

rithms significantly outperform several baselines in running time and the quality.

Pei Lee, Laks V.S. Lakshmanan
University of British Columbia
peil@cs.ubc.ca, laks@cs.ubc.ca

CP11

Unstable Communities in Network Ensembles

Ensembles of graphs arise in several natural appli- cations, such as mobility tracking, computational biology, social networks, and epidemiology. A common problem addressed by many existing mining techniques is to identify subgraphs of interest in these ensembles. In contrast, in this paper, we propose to quickly discover maximally variable regions of the graphs, i.e., sets of nodes that induce very different subgraphs across the ensemble. We first develop two intuitive and novel definitions of such node sets, which we then show can be efficiently enumerated using a level-wise algorithm. Finally, using extensive experiments on multiple real datasets, we show how these sets capture the main structural variations of the given set of networks and also provide us with interesting and relevant insights about these datasets.

Aditya Prakash
Virginia Tech
badityap@cs.vt.edu

CP11

A Fast Kernel for Attributed Graphs

As a fundamental technique for graph analysis, graph kernels have been successfully applied to a wide range of problems. Unfortunately, the high computational complexity of existing graph kernels is limiting their further applications to larger-scale graph datasets. In this paper, we propose a fast graph kernel, the *descriptor matching* (DM) kernel, for graphs with both categorical and numerical attributes. The computation time of the DM kernel is linear with respect to graph size. On graphs with n nodes and m edges, the kernel computation for two graphs can be done in $O(n + m)$ time. Although there are other linear-time graph kernels, most of them are restricted to graphs with only categorical attributes; their efficiency mainly comes from the sparseness of the feature space resulted from the mutually orthogonal categorical attributes. Extensive experiments on both synthetic and real-world graph datasets show promising performance of DM in both accuracy and efficiency.

Yu Su
University of California Santa Barbara
ysu@cs.ucsb.edu

Fangqiu Han
University of California, Santa Barbara
fhan@cs.ucsb.edu

Richard Harang
U.S. Army Research Lab
richard.e.harang.civ@mail.mil

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

CP11

Camlp: Confidence-Aware Modulated Label Propagation

How can we tell if Alice is a talkative person or a silent person? In this paper, we focus on the node classification problem on networked data such as social networks and the web. There are two open challenges with this problem: (1) we want to handle various kinds of label correlations in real-world networks such as homophily (i.e., love of the same) and heterophily (i.e., love of the different), and (2) we want to exploit the confidence of the inference results to enhance the accuracy. There is no algorithm that solves these two challenges at the same time. We tackle with these two challenges by proposing CAMLP, a novel node classification algorithm. Our contributions are three-fold: (a) Novel algorithm; our algorithm is confidence-aware and is applicable to both homophily and heterophily networks, (b) Theory; we give theoretical analyses of our algorithm, and (c) Practice; we perform extensive experiments on 5 different network datasets including homophily and heterophily networks. Our experiments show that the proposed algorithm improves the precision of major competitors not only on heterophily networks, but also on homophily networks.

Yuto Yamaguchi
University of Tsukuba
yuto_ymgc@kde.cs.tsukuba.ac.jp

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

Hiroyuki Kitagawa
University of Tsukuba
kitagawa@cs.tsukuba.ac.jp

CP12

Fast Lossless Frequent Itemset Mining in Data Streams Using Crucial Patterns

We study the problem of mining exact frequent itemsets from data streams. However, the frequent itemsets are often quite large in number and contain redundant information. In this paper, we introduce the novel concept of crucial patterns and show that they provide a significantly better lossless compression for the frequent itemsets than other condensed representations. Lastly, we propose our new Crucial Pattern Mining (CPM) algorithm for data streams that includes several significant optimization strategies.

Ariyam Das, Carlo Zaniolo
University of California, Los Angeles
ariyam@cs.ucla.edu, zaniolo@cs.ucla.edu

CP12

Modelling Recurrent Events for Improving Online Change Detection

The task of online change point detection in sensor data streams is often complicated due to presence of noise that can be mistaken for real changes and therefore affecting performance of change detectors. Most of the existing change detection methods assume that changes are independent from each other and occur at random in time. In

this paper we study how performance of detectors can be improved in case of recurrent changes.

Alexandr Maslov
Eindhoven University of Technology
maslov314@gmail.com

MYKOLA PECHENIZKIY
Eindhoven University of Technology
Department of Mathematics and Computer Science
m.pechenizkiy@tue.nl

Indrė Žliobaitė
Aalto University
Department of Computer Science
zliobaite@gmail.com

Tommi Kärkkäinen
University of Jyvaskyla
Department of Mathematical Information Technology
tommi.karkkainen@jyu.fi

CP12

Temporal Kernel Descriptors for Learning with Time-Sensitive Patterns

Detecting temporal patterns is a challenging data mining task. Often, timestamps of events occurred can provide critical information to recognize temporal patterns. Unfortunately, most existing techniques are not able to fully extract useful temporal information based on the time (especially at different resolutions of time). They miss out on 3 crucial factors: (i) they do not distinguish between timestamp features (which have cyclical or periodic properties) and ordinary features; (ii) they are not able to detect patterns exhibited at different resolutions of time (e.g. different patterns at the annual level, and at the monthly level); and (iii) they are not able to relate different features (e.g. multi-modal features) of instances with different temporal properties (e.g. while predicting stock prices, stock fundamentals may have annual patterns, and at the same time factors like peer stock prices and global markets may exhibit daily patterns). To solve these issues, we offer a novel multiple-kernel learning view and develop Temporal Kernel Descriptors which utilize Kernel functions to comprehensively detect temporal patterns by deriving relationship of instances with the time features. We automatically learn the optimal kernel function, and hence the optimal temporal similarity between two instances. We formulate the optimization as a Multiple Kernel Learning (MKL) problem. We empirically evaluate its performance by solving the optimization using Online MKL.

Doyen Sahoo, Abhishek Sharma
Singapore Management University
doyensahoo.2014@smu.edu.sg,
abhisheksh.2014@smu.edu.sg

Steven Hoi
Singapore Management University
Singapore
chhoi@smu.edu.sg

Peilin Zhao
Institute for Infocomm Research
zhaop@i2r.a-star.edu.sg

CP12

Dpclass: An Effective But Concise Discriminative Patterns-Based Classification Framework

we propose a natural and effective way to resolve pattern-based classification by adopting discriminative patterns which are the prefix paths from root to nodes in tree-based models (e.g., random forest). Moreover, we further compress the number of discriminative patterns by selecting the most effective pattern combinations that fit into a generalized linear model. As a result, our discriminative pattern-based classification framework (DPClass) could perform as good as previous state-of-the-art algorithms, provide great interpretability by utilizing only very limited number of discriminative patterns, and predict new data extremely fast.

Jingbo Shang, Wenzhu Tong, Jian Peng
University of Illinois at Urbana-Champaign
shang7@illinois.edu, wtong8@illinois.edu,
jianpeng@illinois.edu

Jiawei Han
UIUC
hanj@illinois.edu

CP12

Macfp: Maximal Approximate Consecutive Frequent Pattern Mining under Edit Distance

In this paper, we introduce an efficient solution to this problem. We first formulate the Maximal Approximate Consecutive Frequent Pattern Mining (MACFP) problem that identifies substring patterns under edit distance in a long query sequence. Then, we propose a novel algorithm with linear time complexity to check whether the support of a substring pattern is above a predefined threshold in the query sequence, thus greatly reducing the computational complexity of MACFP. With this fast decision algorithm, we can efficiently solve the original pattern discovery problem with several indexing and searching techniques.

Jingbo Shang, Jian Peng
University of Illinois at Urbana-Champaign
shang7@illinois.edu, jianpeng@illinois.edu

Jiawei Han
UIUC
hanj@illinois.edu

CP12

Flexibly Mining Better Subgroups

In subgroup discovery, perhaps the most crucial task is to discover high-quality one-dimensional subgroups, and refinements of these. For nominal attributes, finding such binary features is relatively straightforward, as we can consider individual attribute values as such. For numerical attributes, the task is more challenging as individual numeric values are not reliable statistics. Instead, we can consider combinations of adjacent values, i.e. bins. Existing binning strategies, however, are not tailored for subgroup discovery. That is, the bins they construct do not necessarily facilitate the discovery of high-quality subgroups, therewith potentially degrading the mining result. To address this, we introduce FLEXI. In short, we propose to use

an optimal binning strategy for finding high-quality binary features for both numeric and ordinal attributes. We instantiate FLEXI with various quality measures and show how to achieve efficiency accordingly. Experiments on both synthetic and real-world data sets show that FLEXI outperforms state of the art with up to 25 times improvement in subgroup quality.

Hoang Vu Nguyen
Karlsruhe Institute of Technology
hnguyen@mmci.uni-saarland.de

Jilles Vreeken
Max Planck Institute for Informatics
Saarland University
jilles@mpi-inf.mpg.de

CP13

Deterministic Column Sampling for Low-Rank Matrix Approximation: Nystrom vs. Incomplete Cholesky Decomposition

Computing with large kernel or similarity matrices is essential to many state-ofthe-art machine learning techniques in classification, clustering, and dimensionality reduction. The cost of forming and factoring these kernel matrices can become intractable for large datasets. We introduce an an adaptive column sampling technique called Accelerated Sequential Incoherence Selection (oASIS) that samples columns without computing the entire kernel matrix. Numerical experiments demonstrate that oASIS has performance comparable to state-of-the-art adaptive sampling methods at a fraction of the cost.

Tom Goldstein
Department of Computer Science
University of Maryland
tomg@cs.umd.edu

CP13

A Polynomial Expansion Line Search for Large-Scale Unconstrained Optimization of Smooth L2-Regularized Loss Functions, with Implementation in Apache Spark

In large-scale unconstrained optimization algorithms such as limited memory BFGS (LBFGS) and the Nonlinear Conjugate Gradient (NCG) method, a common subproblem is a univariate line search minimizing the loss function along a descent direction. Commonly used line search methods iteratively compute an approximate solution that satisfies the Wolfe conditions, but typically require multiple function and gradient evaluations and hence have expensive parallel communication costs. In this paper we propose a new line search procedure for cases where the loss function is analytic, as in logistic regression and low-rank matrix factorization. The procedure approximates the loss function by a truncated Taylor polynomial with coefficients that may be computed in a single pass over the dataset, after which the polynomial may be minimized with high accuracy in a neighbourhood of the expansion point. The expansion may be repeated iteratively in a line search until sufficient accuracy is reached. Our Polynomial Expansion Line Search (PELS) method was implemented in the Apache Spark framework, where it accelerated the training of logistic regression models on datasets from the LIBSVM repository with Gradient Descent, NCG, and LBFGS in parallel on a 16 node computing cluster with 256 cores. The PELS method dramatically reduced the number of iterations required by NCG on all datasets, such that it often outperformed LBFGS.

Michael Hynes
University of Waterloo
mbhynes@uwaterloo.ca

CP13

Discovery of Precursors to Adverse Events Using Time Series Data

We develop an algorithm for automatic discovery of precursors in time series data (ADOPT). In a time series setting, a precursor may be considered as any event that precedes and favors an adverse event. In a multivariate time series data, there are exponential number of events which makes a brute force search intractable. ADOPT works by breaking down the problem into two steps - (1) inferring a model of the nominal time series (data without adverse event) by considering the nominal data to be generated by a hidden expert and (2) using the expert's model as a benchmark to evaluate the adverse time series to identify suboptimal events as precursors. For step (1), we use a Markov Decision Process (MDP) framework where value functions and Bellman's optimality are used to infer the expert's actions. For step (2), we define a precursor score to evaluate a given instant of the time series by comparing its utility with that of the expert. Thus, the search for precursor events is transformed to a search for sub-optimal action sequences in ADOPT. As an application case study, we use ADOPT to discover precursors to go-around events in commercial flights using real aviation data.

Vijay Manikanda Janakiraman
UARC/NASA Ames Research Center
vijai@umich.edu

Bryan Matthews
SGT/NASA Ames Research Center
bryan.l.matthews@nasa.gov

Nikunj Oza
Nasa Ames Research Center
nikunj.c.oza@nasa.gov

CP13

Structured Regression on Multilayer Networks

We present a structured regression model for node attribute prediction in multilayer networks. Our model considers the multiple types of nodal connections jointly without averaging and hence maximizes the information gained. It accommodates temporal graphs with missing nodes and unobserved connections. We present evidence that this model outperforms the traditionally used one and that it offers predictive accuracy that increases with the number of layers used, on both synthetic data and challenging real world applications.

Athanasia Polychronopoulou, Zoran Obradovic
Temple University
n.polychr@temple.edu, zoran.obradovic@temple.edu

CP13

Collective Opinion Spam Detection Using Active Inference

Opinion spam has become a widespread problem in the

online review world, where paid or biased reviewers write fake reviews to elevate or relegate a product (or business) to mislead the consumers for profit or fame. In recent years, opinion spam detection has attracted a lot of attention from both the business and research communities. However, the problem still remains challenging as labeled data is scarce and human labeling is expensive which is needed for supervised learning and evaluation. There exist recent work (e.g. FraudEagle [**?**], SpEagle [**?**]) which address the spam detection problem as an unsupervised network inference task on the review network. These methods are also able to incorporate labels (if available), and have been shown to achieve improved performance under the semi-supervised inference setting, in which the labels of a random sample of nodes are consumed. In this work, we address the problem of active inference for the opinion spam detection problem. Active inference is the process of carefully selecting a subset of instances (nodes) whose labels are obtained from an oracle to be used during the (network) inference. As such, our goal is to employ a label acquisition strategy that selects a given number of nodes (a.k.a. the budget) wisely, as opposed to randomly, so as to improve detection performance significantly over the random selection.

Shebuti Rayana
Stony Brook University
srayana@cs.stonybrook.edu

CP13

RelSim: Relation Similarity Search in Schema-Rich Heterogeneous Information Networks

Recent studies have demonstrated the power of modeling real world data as heterogeneous information networks (HINs) consisting of multiple types of entities and relations, where network schema turns out to be a useful high-level description of an HIN. Unfortunately, most of such studies (e.g., similarity search) confine discussions on the networks with only a few entity and relationship types, such as DBLP. In the real world, however, the network schema can be rather complex, such as Freebase. In such HINs with rich schema, it is often too much burden to ask users to provide meta-path(s) explicitly for similarity search. Rather than only focusing on automatic meta-path generation, we study a more interesting problem of relation similarity search in schema-rich HINs, where similar relation instances and the relevant meta-paths are jointly discovered. Under our problem setting, users are only asked to provide some simple relation instance examples (e.g., ¡ Barack Obama, John Kerry ¿ and ¡ George W. Bush, Condoleezza Rice ¿) as a query, and we generate meta-path(s) that best explains the latent semantic relation (LSR) implied by the query (e.g., president vs. secretary-of-state). Such meta-paths will guide to find other similar relation instances (e.g., ¡ Bill Clinton, Madeleine Albright ¿).

Chenguang Wang
Peking University
wangchenguang@pku.edu.cn

CP14

Online Sparse Passive Aggressive Learning with Kernels

Conventional online kernel methods often yield an unbounded large number of support vectors, making them inefficient and non-scalable for large-scale applications. Recent studies on bounded kernel-based online learning have attempted to overcome this shortcoming. Although they can bound the number of support vectors at each iteration, most of them fail to bound the number of support vectors for the final output solution which is often obtained by averaging the series of solutions over all the iterations. In this paper, we propose a novel kernel-based online learning method, Sparse Passive Aggressive learning (SPA), which can output a final solution with a bounded number of support vectors. The key idea of our method is to explore an efficient stochastic sampling strategy, which turns an example into a new support vector with some probability that depends on the loss suffered by the example. We theoretically prove that the proposed SPA algorithm achieves an optimal regret bound in expectation, and empirically show that the new algorithm outperforms various bounded kernel-based online learning algorithms.

Jing Lu
Singapore Management University
Singapore
jing.lu.2014@phdis.smu.edu.sg

Peilin Zhao
Institute for Infocomm Research
zhaop@i2r.a-star.edu.sg

Steven Hoi
Singapore Management University
Singapore
chhoi@smu.edu.sg

CP14

ADMM for Training Sparse Structural SVMs with Augmented ℓ_1 Regularizers

Structural Support Vector Machine (Structural SVM) is a powerful tool for classification problems involving structured outputs. This paper proposes a fast Alternating Direction Method of Multipliers (ADMM) for structural SVM with augmented ℓ_1 regularizers. The designed ADMM alternately solves a sequence of three problems, one of which uses a fast sequential dual optimization method [**?**] developed for training ℓ_2 regularized structural SVM. The other two problems have easy-to-compute closed-form solutions. The algorithm is simple to implement and extensive empirical experiments show that the proposed ADMM is faster than several competing methods on a number of benchmark sequence labeling datasets. In addition to showing the convergence of the proposed ADMM, this paper is the first to prove the global non-asymptotic convergence of the sequential dual optimization method to solve a subproblem of the ADMM algorithm.

Balamurugan Palaniappan
SIERRA Project Team, INRIA, Paris, France.
balamurugan.palaniappan@inria.fr

Anusha Posinasetty
Computer Science and Automation
Indian Institute of Science, Bangalore, India.
anu4iisc@csa.iisc.ernet.in

Shirish Shevade
Indian Institute Of Science
Bangalore

shirish@csa.iisc.ernet.in

CP14

Structural Orthogonal Procrustes Regression for Face Recognition with Pose Variations and Misalignment

Regression based method is a hot topic in the face recognition community and has achieved interesting results when dealing with well-aligned frontal face images. However, most of the existing regression analysis based methods are sensitive to pose variations. In this paper, we firstly introduce the orthogonal Procrustes problem (OPP), which is simple but effective, as a model to handle pose variations in two-dimensional face images. OPP seeks an optimal transformation between two images to correct the pose from one to the other. We integrate OPP into the regression model and propose the structural orthogonal Procrustes regression (SOPR) using the nuclear norm constraint on the error term to keep image's structural information. Moreover, a subject-wise strategy is adopted to address the problem that the gallery images may span over different poses. The proposed model is solved by an efficient iteratively reweighted algorithm and experimental results on popular face databases demonstrate the effectiveness of our method.

Ying Tai, Jian Yang, Fanlong Zhang, Yigong Zhang, Lei Luo, Jianjun Qian
Nanjing University of Science and Technology
tyshiwo@gmail.com, csjyang@njust.edu.cn, csfzhang@126.com, zhangyigong0378@126.com, zzdxpyy3001@163.com, csjqian@njust.edu.cn

CP14

Effective Crowd Expertise Modeling Via Cross Domain Sparsity and Uncertainty Reduction

Characterizations of crowd expertise is vital to online applications where the crowd plays a central role, such as StackExchange for question-answering and LinkedIn as a workforce market. With accurately estimated worker expertise, new jobs can be assigned to the right workers more effectively and efficiently. Most existing methods solely rely on the sparse worker-job interactions, leading to poorly estimated expertise that does not generalize well to a large amount of unseen jobs. Though transfer learning can utilize external domains to mitigate the sparsity, the auxiliary domains can themselves suffer from incomplete information, leading to inferior performance. There is a lack of principled framework to handle the sparse and incomplete data to achieve better expertise modeling. Based on multitask learning, we propose a framework that uses the knowledge learned from one domain to gradually resolve the data sparsity or incompleteness problem in the other alternatively. Experimental results on several question-answering datasets demonstrate the effectiveness and convergence of the iterative framework.

Sihong Xie, Qingbo Hu, Weixiang Shao
University of Illinois at Chicago
sxie6@uic.edu, qhu5@uic.edu, software.shao@gmail.com

Jingyuan Zhang
University of Illinois at Chicago
University of Illinois at Chicago
jzhan8@uic.edu

Jing Gao
University at Buffalo
jing@buffalo.edu

Wei Fan
IBM T.J.Watson Research
wei.fan@gmail.com

Philip S. Yu
University of Illinois at Chicago
Chicago, USA
psyu@uic.edu

CP14

Gspartan: a Geospatio-Temporal Multi-Task Learning Framework for Multi-Location Prediction

This paper presents a novel geospatio-temporal prediction framework called GSpartan to simultaneously build local regression models at multiple locations. The framework assumes that the local models share a common, low-rank representation, which makes them amenable to multi-task learning. GSpartan learns a set of base models to capture the spatio-temporal variabilities of the data and represents each local model as a linear combination of the base models. A graph Laplacian regularization is used to enforce constraints on the local models based on their spatial autocorrelation. We also introduce sparsity-inducing norms to perform feature selection for the base models and model selection for the local models. Experimental results using historical climate data from 37 weather stations showed that, on average, GSpartan outperforms single-task learning and other existing multi-task learning methods in more than 65% of the stations, which increases to 81% when there are fewer training examples.

Jianpeng Xu
Computer Science and Engineering Department
Michigan State University
xujianpe@msu.edu

Pang-Ning Tan
Michigan State University
ptan@cse.msu.edu

Lifeng Luo
Geograpy Department
Michigan State University
lluo@msu.edu

Jiayu Zhou
MSU
jiayuz@msu.edu

CP14

Learning Correlative and Personalized Structure for Online Multi-Task Classification

Multi-Task Learning (MTL) enhances generalization performance by learning multiple related tasks simultaneously. Conventional MTL works under the offline or batch learning setting, which suffers from the expensive re-training cost and poor scalability. To address these inefficiency issues, online learning technique has been applied to solve the MTL problems. However, existing algorithms for online MTL constrain task relatedness into a presumed structure via a single weight matrix, a strict restriction that does not hold in the practical scenarios. In this paper, we propose a general online MTL framework that overcomes

this restriction by decomposing the weight matrix into two components: the first component captures the correlative structure among tasks in a low-rank subspace, whereas the second component identifies the personalized patterns for the outlier tasks. We devise a projected gradient scheme to adaptively learn the components. Theoretical analysis of our solution shows that the proposed algorithm can achieve a sub-linear regret with respect to the best linear model in hindsight. Experimental results on a number of real-world datasets also verify the efficacy of our approach.

Peng Yang
Institute for Infocomm Research
yangp@i2r.a-star.edu.sg

Giangxia Li
Institute for Infocomm Research, Singapore
lig@i2r.a-star.edu.sg

Peilin Zhao
Institute for Infocomm Research
zhaop@i2r.a-star.edu.sg

Xiao-Li Li, Sujatha Das Gollapalli
Institute for Infocomm Research, Singapore
xlli@i2r.a-star.edu.sg, gollapallis@i2r.a-star.edu.sg

CP15

Rank Selection for Non-Negative Matrix Factorization Using Normalized Maximum Likelihood Coding

Non-negative matrix factorization (NMF) is one of the most important technologies in data mining.This is the task of factorizing a matrix into the product of two non-negative low rank matrices. In most of works on NMF, the rank is predetermined in ad hoc. This paper addresses the issue of how we can select the best rank from given data. The problem is that the conventional statistical model selection criteria such as AIC, MDL etc. cannot straightforwardly be applied to this issue because the regularity conditions for the criteria are not fulfilled. We overcome this problem to propose a novel methodology for rank selection. The key ideas are to 1) use the technique of latent variable completion to make the model regular and 2) then to apply the normalized maximum likelihood coding to rank selection for the regular model. We further propose a novel method for rank change detection when rank changes over time.We demonstrate the effectiveness of our methods for rank selection and rank change detection through synthetic data and real data sets.

Yu Ito
The University of Tokyo
dance2982002@gmail.com

Shin-Ichi Oeda
National Institute of Technology, Kisarazu College
oeda@j.kisarazu.ac.jp

Kenji Yamanishi
The University of Tokyo
yamanishi@mist.i.u-tokyo.ac.jp

CP15

Capricorn: An Algorithm for Subtropical Matrix Factorization

Max-times algebra, sometimes known as subtropical algebra, is a semi-ring over the nonnegative real numbers where the addition operation is the max function and the multiplication is the standard one. Factorising a matrix over the max-times algebra allows us to find structures that cannot be easily expressed using standard decompositions. We present an algorithm, Capricorn, that finds such decompositions, and is also capable of identifying when the subtropical structure is not present.

Sanjar Karaev
Max Planck Institute for Informatics
skaraev@mpi-inf.mpg.de

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

CP15

Automatic Unsupervised Tensor Mining with Quality Assessment

Tensor decomposition has been very popular in unsupervised modelling and multi-aspect data mining. In an exploratory setting, where no labels or ground truth are available how can we automatically decide how many components to extract? How can we assess the quality of our results, so that a domain expert can factor this quality measure in the interpretation of our results? In this paper, we introduce AutoTen, a novel automatic unsupervised tensor mining algorithm with minimal user intervention, which leverages and improves upon heuristics that assess the result quality. We extensively evaluate method's performance on synthetic data, outperforming existing baselines on this very hard problem. Finally, we apply AutoTen to a variety of real datasets, providing insights and discoveries.

Evangelos E. Papalexakis
CMU
epapalex@cs.cmu.edu

CP16

Scaling Lifted Probabilistic Inference and Learning Via Graph Databases

Over the past decade, exploiting relations and symmetries within probabilistic models has been proven to be surprisingly effective at solving large scale data mining problems. One of the key operations inside these lifted approaches is counting - be it for parameter/structure learning or for efficient inference. Typically, however, they just count exploiting the logical structure using adhoc operators. This paper investigates whether 'Compilation to Graph Databases' could be a practical technique for scaling lifted probabilistic inference and learning methods. We demonstrate that the proposed approach achieves reasonable speed-ups for both inference and learning, without sacrificing performance.

Mayukh Das
Indiana University, Bloomington
maydas@indiana.edu

Yuqing Wu
Computer Science Department
Pomona College

melanie.wu@pomona.edu

Tushar Khot
Allen Institute for AI
tushark@allenai.org

Kristian Kersting
Computer Science Department
Technical University of Dortmund, Germany
kristian.kersting@cs.tu-dortmund.de

Sriraam Natarajan
Indiana University
School of Informatics and Computing
natarasr@indiana.edu

CP16

Sparse Hybrid Variational-Gibbs Algorithm for Latent Dirichlet Allocation

Topic modeling algorithms such as the latent Dirichlet allocation (LDA) play an important role in machine learning research. Fitting LDA using Gibbs sampler-related algorithms involves a sampling process over K topics. We can use the sparsity in LDA to accelerate this expensive topic sampling process even for very large K values. However, LDA gradually loses sparsity as the number of documents increases. Motivated by the goal of fast LDA inference with large numbers of both topics and documents, in this paper we propose the novel sparse hybrid variational-Gibbs (SHVG) algorithm. The SHVG algorithm divides the topic sampling probability into a sparse term that scales linearly with the number of per-document instantiated topics K_d, and a dense term that uses the Alias method to reduce the time cost to constant $O(1)$ time. This will lead to a significant improvement on efficiency. Using stochastic optimization techniques, we further develop an online version of SHVG for streaming documents. Experimental results on corpora with a wide range of sizes demonstrate the efficiency and effectiveness of the proposed SHVG algorithm.

Ximing Li, Jihong Ouyang, Xiaotang Zhou
Jilin university
liximing86@gmail.com, ouyj@jlu.edu.cn, zhou_xiaotang@126.com

CP16

Estimating Posterior Ratio for Classification: Transfer Learning from Probabilistic Perspective

Transfer learning assumes classifiers of similar tasks share certain parameter structures. Unfortunately, modern classifiers uses sophisticated feature representations with huge parameter spaces which lead to costly transfer. Under the impression that changes from one classifier to another should be "simple', an efficient transfer learning criteria that only learns the "differences' is proposed in this paper. We train a *posterior ratio* which turns out to minimize the upper-bound of the target learning risk. The model of posterior ratio does not have to share the same parameter space with the source classifier at all so it can be easily modelled and efficiently trained. The resulting classifier therefore is obtained by simply multiplying the existing probabilistic-classifier with the learned posterior ratio.

Song Liu, Kenji Fukumizu
The Institute of Statistical Mathematics
liu@ism.ac.jp, fukumizu@ism.ac.jp

CP17

High Dimensional Structured Estimation with Noisy Designs

Structured estimation methods, such as LASSO, have received considerable attention in recent years and substantial progress has been made in extending such methods to general norms and non-Gaussian design matrices. In real world problems, however, covariates are usually corrupted with noise and there have been efforts to generalize structured estimation method for noisy covariate setting. In this paper we first show that without any information about the noise in covariates, currently established techniques of bounding statistical error of estimation fail to provide consistency guarantees. However, when information about noise covariance is available or can be estimated, then we prove consistency guarantees for any norm regularizer, which is a more general result than the state of the art. Next, we investigate empirical performance of structured estimation, specifically LASSO, when covariates are noisy and empirically show that LASSO is not consistent or stable in the presence of additive noise. However, prediction performance improves quite substantially when the noise covariance is available for incorporating in the estimator.

Amir Asiaee T.
University of Minnesota
ataheri@cs.umn.edu

Soumyadeep Chaterjee
Yahoo! Labs
soumyadeep@yahoo-inc.com

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

CP17

Regularized Parametric Regression for High-Dimensional Survival Analysis

Survival analysis aims to predict the occurrence of specific events of interest at future time points. The presence of incomplete observations due to *censoring* brings unique challenges in this domain and differentiates survival analysis techniques from other standard regression methods. In many applications where the distribution of the survival times can be explicitly modeled, parametric survival regression is a better alternative to the commonly used Cox proportional hazards model for this problem of censored regression. However, parametric survival regression suffers from model overfitting in high-dimensional scenarios. In this paper, we propose a *unified* model for regularized parametric survival regression for an arbitrary survival distribution. We employ a generalized linear model to approximate the negative log-likelihood and use the elastic net as a sparsity-inducing penalty to effectively deal with high-dimensional data. The proposed model is then formulated as a penalized iteratively reweighted least squares and solved using a cyclical coordinate descent-based method. We demonstrate the performance of our proposed model on various high-dimensional real-world microarray gene expression benchmark datasets. Our experimental results indicate that the proposed model produces more accurate estimates compared to the other competing state-of-the-art

methods.

Yan Li
Wayne State University
rock_liyan@wayne.edu

Kevin Xu
University of Toledo
kevin.xu@utoledo.edu

Chandan Reddy
Wayne State University
reddy@cs.wayne.edu

CP17

On Skewed Multi-Dimensional Distributions: the *fusionrp* Model, Algorithms, and Discoveries

How do we model and find outliers in Twitter data? Given the number of retweets of each person on a social network, what is their expected number of comments? Real-life data are often very skewed, exhibiting power-law-like behavior. For such skewed multi-dimensional discrete data, the existing models are not general enough to capture various realistic scenarios, and need to be discretized as they often model continuous quantities. We propose *FusionRP*, short for Fusion Restaurant Process, a simple and intuitive model for skewed multi-dimensional discrete distributions, such as number of retweets vs. comments in Twitter-like data. Our model is discrete by design, has provably asymptotic log-logistic sum of marginals , is general enough to capture varied relationships, and most importantly, fits real data very well. We give an effective and scalable maximum-likelihood based fitting approach that is linear in the number of unique observed values and the input dimension. We test *FusionRP* on a twitter-like social network with 2.2M users, a phone call network with 1.9M call records, game data with 45M users and Facebook data with 2.5M posts. Our results show that *FusionRP* significantly outperforms several alternative methods and can detect outliers, such as bot-like behavior in the Facebook data.

Venkata Krishna Pillutla
Carnegie Mellon University
pillutla.krishna@gmail.com

Zhanpeng Fang
Carnegie Mellon University
Tsinghua University
zhanpenf@andrew.cmu.edu

Pravallika Devineni
University of California, Riverside
pdevi002@ucr.edu

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

Danai Koutra
University of Michigan, Ann Arbor
dkoutra@umich.edu

Jie Tang
Tsinghua University
jietang@tsingua.edu.cn

CP17

Copula-HDP-HMM: Non-Parametric Modeling of Temporal Multivariate Data for I/O Efficient Bulk Cache Preloading

Bulk-cache-preloading involves prefetching large batches of relevant data into cache. We address this by analyzing high-level spatio-temporal motifs from I/O-traces by aggregating them into a time-series of multivariate counts. To analyze such data, we propose Copula-HDP-HMM, a Bayesian-non-parametric model based on Gaussian-Copula, suitable for temporal multivariate data with arbitrary marginals, also suitable for multivariate counts. Finally, HULK, our proposed bulk-cache-preloading strategy using Copula-HDP-HMM, shows order of magnitude improvement in hitrates and I/O-efficiency over baselines.

Lavanya S. Tekumalla
CSA, Indian Institute Of Science
lavanya.iisc@gmail.com

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

Anusha Posinasetty
Computer Science and Automation
Indian Institute of Science, Bangalore, India.
anu4iisc@csa.iisc.ernet.in

CP17

Constrained Group Testing to Predict Binding Response of Candidate Compounds

We study the problem of identifying reactive compound(s) in a solution, using a minimal number of chemical tests. To solve this problem, we introduce a new model called *mask-based constrained group testing*, develop an algorithm, and prove that under the right conditions, our algorithm is guaranteed w.h.p. to identify the active compounds of a solution with a small number of tests. We also show that our algorithm performs well empirically.

Paul Quint
Dept. of Computer Science
University of Nebraska
pquint@cse.unl.edu

Stephen Scott
Department of Computer Science
University of Nebraska
sscott@cse.unl.edu

N. V. Vinodchandran
Dept. of Computer Science
University of Nebraska
vinod@cse.unl.edu

BRAD Worley
Dept. of Chemistry
University of Nebraska

bradley.worley@huskers.unl.edu

CP17

Universal Dependency Analysis

Most data is multi-dimensional. Discovering whether any subset of dimensions, or subspaces, shows dependence is a core task in data mining. To do so, we require a measure that quantifies how dependent a subspace is. For practical use, such a measure should be universal in the sense that it captures correlation in subspaces of any dimensionality and allows to meaningfully compare scores across different subspaces, regardless how many dimensions they have and what specific statistical properties their dimensions possess. Further, it would be nice if the measure can non-parametrically and efficiently capture both linear and non-linear correlations. In this paper, we propose UDS, a multivariate dependence measure that fulfils all of these desiderata. In short, we define UDS based on cumulative entropy and propose a principled normalisation scheme to bring its scores across different subspaces to the same domain, enabling universal dependence assessment. UDS is purely non-parametric as we make no assumption on data distributions nor types of correlation. To compute it on empirical data, we introduce an efficient and non-parametric method. Extensive experiments show that UDS outperforms state of the art.

Hoang Vu Nguyen
Karlsruhe Institute of Technology
hnguyen@mmci.uni-saarland.de

Panagiotis Mandros
Max Planck Institute for Informatics
pmandros@mpi-inf.mpg.de

Jilles Vreeken
Max Planck Institute for Informatics
Saarland University
jilles@mpi-inf.mpg.de

CP18

Infusing Geo-Recency Mixture Models for Effective Location Prediction in LBSN

An overwhelming volume of geo-social data is generated daily. In this paper, we propose a novel technique that utilizes matrix factorization to predict the top-k future locations for check-in data. Specifically, we introduce a new approach to capture the unique mobility behaviors of users by crafting Geo-Recency based Gaussian mixture models. In addition, we present a new technique that employs these Geo-Recency Gaussian mixture models to effectively quantify social influence.

Roland Assam, Subramanyam Sathyanarayana
RWTH Aachen University, Germany
assam@cs.rwth-aachen.de,
subramanyam.sathyanarayana@rwth-aachen.de

Thomas Seidl
RWTH Aachen University
seidl@cs.rwth-aachen.de

CP18

Joint Learning of Representation and Structure for Sparse Regression on Graphs

In many applications, including climate science, power systems, and remote sensing, multiple input variables are observed for each output variable and the output variables are dependent. Several methods have been proposed to improve prediction by learning the conditional distribution of the output variables. However, when the relationship between the raw features and the outputs is nonlinear, the existing methods cannot capture both the nonlinearity and the underlying structure well. In this study, we propose a structured model containing hidden variables, which are nonlinear functions of inputs and which are linearly related with the output variables. The parameters modeling the relationships between the input and hidden variables, between the hidden and output variables, as well as among the output variables are learned simultaneously. To demonstrate the effectiveness of our proposed method, we conducted extensive experiments on eight synthetic datasets and three real-world challenging datasets: forecasting wind power, forecasting solar energy, and forecasting precipitation over U.S. The proposed method was more accurate than state-of-the-art structured regression methods.

Chao Han, Shanshan Zhang
Temple University
tuf27898@temple.edu, tuf14438@temple.edu

Mohamed Ghalwash
Temple University
Ain Shams University
mohamed@temple.edu

Slobodan Vucetic, Zoran Obradovic
Temple University
vucetic@temple.edu, zoran.obradovic@temple.edu

CP18

Learning Linear Dynamical Systems from Multivariate Time Series: A Matrix Factorization Based Framework

We propose a generalized LDS framework, gLDS, for learning LDS models from a collection of multivariate time series (MTS) based on matrix factorization. One advantage of gLDS is that various types of constraints can be easily incorporated into the learning process. Furthermore, we propose a temporal smoothing regularization approach for learning the LDS model, which stabilizes the model, its learning algorithm and predictions it makes. We demonstrate the advantages of gLDS on various real-world datasets.

Zitao Liu, Milos Hauskrecht
University of Pittsburgh
ztliu@cs.pitt.edu, milos@cs.pitt.edu

CP18

Spatio-Temporal Tensor Analysis for Whole-Brain Fmri Classification

Owing to its importance in disease diagnosis, whole-brain fMRI image analysis has become an emerging research area. Conventionally, the input fMRI brain images are converted into vectors or matrices and adapted in kernel based classifiers. However, fMRI data are inherently coupled with sophisticated spatio-temporal tensor structure. Such important structural information will be lost if the

tensors are converted into vectors. Furthermore, the time series of fMRI data are often very noisy, involving time shift and with low temporal resolution. To deal with the above challenges, more compact and discriminative representations for kernel modeling are needed. In this paper, we propose a novel spatio-temporal tensor kernel (STTK) approach for whole-brain fMRI image analysis. Specifically, we design a volumetric time series extraction approach to model the temporal data, and propose a spatio-temporal tensor based factorization for feature extraction. We further leverage the tensor structure to encode prior knowledge in the kernel. Extensive experiments on real-world datasets demonstrate that our proposed approach can effectively boost the fMRI classification performance on divergent disease diagnosis (i.e. Alzheimer's disease, ADHD and HIV).

Guixiang Ma
University of Illinois, Chicago
gma4@uic.edu

CP18

Speeding Up All-Pairwise Dynamic Time Warping Matrix Calculation

Dynamic Time Warping (DTW) is certainly the most relevant distance for time series analysis. However, its quadratic time complexity may hamper its use, mainly in the analysis of large time series data. All the recent advances in speeding up the exact DTW calculation are confined to similarity search. However, there is a significant number of important algorithms including clustering and classification that require the pairwise distance matrix for all time series objects. The only techniques available to deal with this issue are constraint bands and DTW approximations. In this paper, we propose the first exact approach for speeding up the all-pairwise DTW matrix calculation. Our method is exact and may be applied in conjunction with constraint bands. We demonstrate that our algorithm reduces the runtime in approximately 50% on average and up to one order of magnitude in some datasets.

Diego Silva
University of Sao Paulo
diegofsilva@gmail.com

Gustavo Batista
University of São Paulo
gbatista@icmc.usp.br

CP18

Linear-Time Detection of Non-Linear Changes in Massively High Dimensional Time Series

Change detection in multivariate time series has applications in many domains, including health care and network monitoring. A common approach to detect changes is to compare the divergence between the distributions of a reference window and a test window. When the number of dimensions is very large, however, such a naive approach has both quality and efficiency issues: to ensure robustness the window size needs to be large, which not only leads to missed alarms but also increases runtime. To this end, we propose Light, a linear-time algorithm for robustly detecting non-linear changes in massively high dimensional time series. Importantly, Light provides high flexibility in choosing the window size, allowing the domain expert to fit the level of details required. To do such, we 1) perform scalable PCA to reduce dimensionality, 2) perform scalable factorisation of the joint distribution, and 3) scalably compute divergences between these lower dimensional distributions. Extensive empirical evaluation on both synthetic and real-world data show that Light outperforms state of the art with up to 100% improvement in both quality and efficiency.

Hoang Vu Nguyen
Karlsruhe Institute of Technology
hnguyen@mmci.uni-saarland.de

Jilles Vreeken
Max Planck Institute for Informatics
Saarland University
jilles@mpi-inf.mpg.de