IP1

Title Not Available at Time of Publication - Fugate

Abstract not available at time of publication.

William Craig Fugate
Federal Emergency Management Agency
wcfugate@gmail.com

IP2

Title Not Available at Time of Publication - Hidalgo

Abstract not available at time of publication

Cesar Hidalgo
Massachusetts Institute of Technology
cesifoti@gmail.com

IP3

Machine Learning under Resource Constraints

Big data are produced by various sources. Most often, they are distributedly stored at computing farms or clouds. Analytics on the Hadoop Distributed File System (HDFS) then follows the MapReduce programming model. According to the Lambda architecture of Nathan Marz and James Warren, this is the batch layer. It is complemented by the speed layer, which aggregates and integrates incoming data streams in real time. When considering big data and small devices, obviously, we imagine the small devices being hosts of the speed layer, only. Analytics on the small devices is restricted by memory and computation resources. The interplay of streaming and batch analytics offers a multitude of configurations. In this talk, we discuss opportunities for using sophisticated models for learning spatio-temporal models. In particular, we investigate graphical models, which generate the probabilities for connected (sensor) nodes. We even approximate likelihood estimates such that they can be computed on very restricted devices.

Katharina Morik
Technische Universität Dortmund
katharina.morik@cs.uni-dortmund.de

IP4

Repeated Choice, Markov Models, and LAMP

In this talk Ill discuss modeling of user consumption data, using the framework of Discrete Choice. Ill start with some background on choice models, survey the key ideas, then give an example modeling repeated consumption of the same item (for instance a song, restaurant, or web page). Ill then show how choice theory can be integrated with first-order markov models to develop lightweight but powerful LAMP sequence models that efficiently learn how to take history into account. Finally, Ill show some comparisons between these models and standard deep network sequence models.

Andrew Tomkins
Google
atomkins@gmail.com

CP1

Risk Clearance with Guaranteed Precision

In real life applications, we often face the following risk clearance problem: given a set of instances with a known numeric outcome Y (e.g., a toxicity level), we want to learn a model to identify new instances that have a low risk, i.e., the probability of the Y value exceeding a certain maximum MAX is less than some threshold t. This problem guarantees that the cleared instances have the minimum precision of $1-t$ for $Y \leq MAX$. By clearing such low risk instances, we can allocate costly resources to the remaining high risk instances. In this work, we formulate this problem as Risk Clearance with a goal of maximizing the clearance of low risk instances. Existing classification models fail to solve Risk Clearance adequately, so we develop algorithms designed specifically for this problem. We then validate that our approach improves on existing work via experiments on an industrial case study.

Ryan Mcbride
Simon Fraser University
rom2@sfu.ca

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

Viswanadh Nekkanti
Simon Fraser University
vnekkant@sfu.ca

Wenyuan Li
Chongqing University
wenyuanl.li@ieee.org

CP1

Pruning Decision Trees Via Max-Heap Projection

The decision tree model has gained great popularity both in academia and industry due to its capability of learning highly non-linear decision boundaries, and at the same time, still preserving interpretability that usually translates into transparency of decision-making. However, it has been a longstanding challenge for learning robust decision tree models since the learning process is usually sensitive to data and many existing tree learning algorithms lead to overfitted tree structures due to the heuristic and greedy nature of these algorithms. Pruning is usually needed as an ad-hoc procedure to prune the tree structure, which is, however, not guided by a rigorous optimization formulation but by some intuitive statistical justification. Motivated by recent developments in sparse learning, in this paper, we propose a novel formulation that recognizes an interesting connection between decision tree post-pruning and sparse learning, where the tree structure can be embedded as constraints in the sparse learning framework via the use of a max-heap constraint. This novel formulation leads to a non-convex optimization problem which can be solved by an iterative shrinkage algorithm in which the proximal operator can be solved by an efficient max-heap projection algorithm. Extensive experimental results demonstrate that our proposed method achieves better predictive performance than many existing benchmark methods across a wide range of real-world datasets.

Zhi Nie

Arizona State University
Arizona State University
Zhi.Nie@asu.edu

Binbin Lin
University of Michigan, Ann Arbor
bilin@umich.edu

Shuai Huang
University of Washington
shuaih@u.washington.edu

Naren Ramakrishnan
Computer Science
Virginia Tech
naren@cs.vt.edu

Wei Fan
Baidu Research Big Data Lab
wei.fan@gmail.com

Jieping Ye
University of Michigan, Ann Arbor
jpye@umich.edu

CP1

Using a Random Forest to Inspire a Neural Network and Improving on It

Neural networks have become very popular in recent years because of the astonishing success of deep learning in various domains such as image and speech recognition. In many of these domains, specific architectures of neural networks, such as convolutional networks, seem to fit the particular structure of the problem domain very well, and can therefore perform in an astonishingly effective way. However, the success of neural networks is not universal across all domains. Indeed, for learning problems without any special structure, or in cases where the data is somewhat limited, neural networks are known not to perform well with respect to traditional machine learning methods such as random forests. In this paper, we show that a carefully designed neural network with random forest structure can have better generalization ability. In fact, this architecture is more powerful than random forests, because the back-propagation algorithm reduces to a more powerful and generalized way of constructing a decision tree. Furthermore, the approach is efficient to train and requires a small constant factor of the number of training examples. This efficiency allows the training of multiple neural networks in order to improve the generalization accuracy. Experimental results on 10 real-world benchmark datasets demonstrate the effectiveness of the proposed framework.

Suhang Wang
Arizona State University
suhang.wang@asu.edu

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

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

CP1

Active Learning of Classification Models with Likert-Scale Feedback

Annotation of classification data by humans can be a time-consuming and tedious process. Finding ways of reducing the annotation effort is critical for building the classification models in practice and for applying them to a variety of classification tasks. In this paper, we develop a new active learning framework that combines two strategies to reduce the annotation effort. First, it relies on label uncertainty information obtained from the human in terms of the Likert-scale feedback. Second, it uses active learning to annotate examples with the greatest expected change. We propose a Bayesian approach to calculate the expectation and an incremental SVM solver to reduce the time complexity of the solvers. We show the combination of our active learning strategy and the Likert-scale feedback can learn classification models more rapidly and with a smaller number of labeled instances than methods that rely on either Likert-scale labels or active learning alone.

Milos Hauskrecht, Yanbing Xue
Computer Science Department
University of Pittsburgh
milos@pitt.edu, yanbing@cs.pitt.edu

CP1

Margin Distribution Logistic Machine

Linear classifier is an essential part of machine learning, and improving its robustness has attracted much effort. Logistic regression (LR) is one of the most widely used linear classifier for its simplicity and probabilistic output. To reduce the risk of overfitting, LR was enhanced by introducing a generalized logistic loss (GLL) with a L2-norm regularization, aiming to maximize the minimum margin. However, the strategy of maximizing minimal margin is less robust to noisy data. In this paper, we incorporate GLL with margin distribution to exploit the statistical information from the training data, and propose a margin distribution logistic machine (MDLM) for better generalization performance and robustness. Furthermore, we extend MDLM to a multi-class version and learn different classes simultaneously by utilizing more information shared across these classes. Extensive experimental results validate the effectiveness of MDLM on both binary classification and multi-class classification.

Yi Ding, Sheng-Jun Huang, Chen Zu, Daoqiang Zhang
College of Computer Science and Technology
Nanjing University of Aeronautics and Astronautics
yiding.nuaa@foxmail.com, huangsj@nuaa.edu.cn,
chenzu@nuaa.edu.cn, dqzhang@nuaa.edu.cn

CP2

Alpine: Progressive Itemset Mining with Definite Guarantees

With increasing demand for efficient data analysis, execution time of itemset mining becomes critical for many large-scale or time-sensitive applications. We propose a dynamic approach for itemset mining that allows us to achieve flexible trade-offs between efficiency and completeness. ALPINE is to our knowledge the first algorithm to progressively mine itemsets and closed itemsets "support-

wise". It guarantees that all itemsets with support exceeding the current checkpoint's support have been found before it proceeds further. Thus, it is very attractive for extremely long mining tasks with very high dimensional data because it can offer intermediate meaningful and complete results. This feature is the most important contribution of ALPINE, which is also fast but not necessarily the fastest algorithm around. Another critical advantage of ALPINE is that it does not require the apriori decided minimum support threshold.

Qiong Hu, Tomasz Imielinski
Rutgers University
qionghu.cs@rutgers.edu, imielins@cs.rutgers.edu

CP2

Computational Drug Discovery with Dyadic Positive-Unlabeled Learning

Computational Drug Discovery, which uses computational techniques to facilitate and improve the drug discovery process, has aroused considerable interests in recent years. Drug Repositioning (DR) and Drug-Drug Interaction (DDI) prediction are two key problems in drug discovery and many computational techniques have been proposed for them in the last decade. Although these two problems have mostly been researched separately in the past, both DR and DDI can be formulated as the problem of detecting positive interactions between data entities (DR is between drug and disease, and DDI is between pairwise drugs). The challenge in both problems is that we can only observe a very small portion of positive interactions. In this paper, we propose a novel framework called Dyadic Positive-Unlabeled learning (DyPU) to solve the problem of detecting positive interactions. DyPU forces positive data pairs to rank higher than the average score of unlabeled data pairs. Moreover, we also derive the dual formulation of the proposed method with the rectifier scoring function and we show that the associated non-trivial proximal operator admits a closed form solution. Extensive experiments are conducted on real drug data sets and the results show that our method achieves superior performance comparing with the state-of-the-art.

Yashu Liu
Computer Science and Engineering, Arizona State University
yashu.liu@asu.edu

Shuang Qiu
University of Michigan, Ann Arbor
qiush@umich.edu

Ping Zhang
IBM T.J. Watson Research Center
pzhang@us.ibm.com

Pinghua Gong
University of Michigan, Ann Arbor
gongp@umich.edu

Fei Wang
IBM T J Watson Research Center
feiwang03@gmail.com

Guoliang Xue
Arizona State University
xue@asu.edu

Jieping Ye
University of Michigan, Ann Arbor
jpye@umich.edu

CP2

Active Learning of Functional Networks from Spike Trains

Learning functional networks from spike trains is a fundamental problem with many critical applications in neuroscience. However, most of existing works focus on inferring the functional network purely from observational data, which could lead to undiscovered or spurious connections. We demonstrate that by adopting experimental data with interventions applied, the accuracy of the inferred network can be significantly improved. Nevertheless, doing interventions in real experiments is often expensive and must be chosen with care. Hence, in this paper, we design an active learning framework to iteratively choose interventions and learn the functional network. In particular, we propose two models, the variance model and the validation model, to effectively select the most informative interventions. The variance model works best to reveal undiscovered connections while the validation model has the advantage of eliminating spurious connections. Experimental results with both synthetic and real datasets show that when these two models are applied, we could achieve substantially better accuracy than using the same amount of observational data or other baseline methods to choose interventions.

Honglei Liu
University of California, Santa Barbara
liuhonglei@gmail.com

Bian Wu
Washington State University
bian.wu@wsu.edu

CP2

Polyadic Regression and Its Application to Chemogenomics

We study the problem of Polyadic Prediction, where the input consists of an ordered tuple of objects, and the goal is to predict a measurement associated with them. Many tasks can be naturally framed as Polyadic Prediction problems. In drug discovery, for instance, it is important to estimate the treatment effect of a drug on various tissue-specific diseases, as it is expressed over the available genes. Thus, we essentially predict the expression value measurements for several (drug, gene, tissue) triads. To tackle Polyadic Prediction problems, we propose a general framework, called Polyadic Regression, predicting measurements associated with multiple objects. Our framework is inductive, in the sense of enabling predictions for new objects, unseen during training. Our model is expressive, exploring high-order, polyadic interactions in an efficient manner. An alternating Proximal Gradient Descent procedure is proposed to fit our model. We perform an extensive evaluation using real-world chemogenomics data, where we illustrate the superior performance of Polyadic Regression over the prior art. Our method achieves an increase of 0.06 and 0.1 in Spearman correlation between the predicted and the actual measurement vectors, for predicting missing polyadic data and predicting polyadic data for new drugs, respectively.

Ioakeim Perros

Georgia State University
Georgia State University
elatash1@student.gsu.edu

Fei Wang
IBM T J Watson Research Center
feiwang03@gmail.com

Ping Zhang
IBM T.J. Watson Research Center
pzhang@us.ibm.com

Paul Walker
United States Navy
peter.b.walker.mil@mail.mil

Richard Vuduc
Georgia Institute of Technology
richie@cc.gatech. edu

Jyotishman Pathak
Cornell University
pathak@med.cornell.edu

Jimeng Sun
Georgia Institute of Technology
jimeng.sun@gmail.com

CP2

Multi-Region Neural Representation: A Novel Model for Decoding Visual Stimuli in Human Brains

Multivariate Pattern (MVP) classification holds enormous potential for decoding visual stimuli in the human brain by employing task-based fMRI data sets. This paper proposes a novel model of neural representation, which can automatically detect the active regions for each visual stimulus and then utilize these anatomical regions for visualizing and analyzing the functional activities. Moreover, our method introduces analyzing snapshots of brain image for decreasing sparsity rather than using the whole of fMRI time series.

Muhammad Yousefnezhad, Daoqiang Zhang
College of Computer Science and Technology
Nanjing University of Aeronautics and Astronautics
myousefnezhad@outlook.com, dqzhang@nuaa.edu.cn

CP3

Outlier Detection with Autoencoder Ensembles

In this paper, we introduce autoencoder ensembles for unsupervised outlier detection. One problem with neural networks is that they are sensitive to noise and often require large data sets to work robustly, while increasing data size makes them slow. As a result, there are only a few existing works in the literature on the use of neural networks in outlier detection. This paper shows that neural networks can be a very competitive technique to other existing methods. The basic idea is to randomly vary on the connectivity architecture of the autoencoder to obtain significantly better performance. Furthermore, we combine this technique with an adaptive sampling method to make our approach more efficient and effective. Experimental results comparing the proposed approach with state-of-the- art detectors are presented on several benchmark data sets showing the accuracy of our approach.

Jinghui Chen
University of Virginia
jc4zg@virginia.edu

Saket Sathe, Charu C. Aggarwal
IBM T. J. Watson Research Center
ssathe@us.ibm.com, charu@us.ibm.com

Deepak Turaga
IBM Research
turaga@us.ibm.com

CP3

VolTime: Unsupervised Anomaly Detection on Users' Online Activity Volume

Is it possible to spot review frauds and spamming on social media and online stores? In this paper we analyze the joint distribution of the inter-arrival times and volume of events such as comments and online reviews and show that it is possible to accurately rank and detect suspicious users such as spammers, bots and fraudsters. We propose VolTime, a generative model that fits well the inter-arrival time distribution (IAT) of real users. Thus, VolTime automatically spots and ranks suspicious users. Experiments on several real datasets, ranging from Reddit comments and phone calls to Flipkart product reviews, show that VolTime is able to accurately fit the activity volume and IAT of real data. Additionally, we show that VolTime ranks suspicious users with a precision higher than 90% for a sensitivity of 70%.

Daniel Chino, Alceu Costa
University of Sao Paulo
Institute of Mathematical and Computer Sciences
chinodyt@icmc.usp.br, alceufc@icmc.usp.br

Agma J. Traina
University of Sao Paulo
agma@icmc.usp.br

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

CP3

Efficiently Discovering Unexpected Pattern-Co-Occurrences

Our world is filled with both beautiful and brainy people, but how often does a Nobel Prize winner also wins a beauty pageant? Let us assume that someone who is both very beautiful and very smart is more rare than what we would expect from the combination of the number of beautiful and brainy people. Of course there will still always be some individuals that defy this stereotype; these beautiful brainy people are exactly the class of anomaly we focus on in this paper. They do not posses intrinsically rare qualities, it is the unexpected combination of factors that makes them stand out. In this paper we define the above described class of anomaly and propose a method to quickly identify them in transaction data. Further, as we take a pattern set based approach, our method readily explains why a transaction is anomalous. The effectiveness of our method is thoroughly verified with a wide range of experiments on

both real world and synthetic data.

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

Roel Bertens
Universiteit Utrecht
Department of Information and Computing Sciences
R.Bertens@uu.nl

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

CP3

Gleaning Wisdom from the Past: Early Detection of Emerging Rumors in Social Media

The explosive use of social media, in information dissemination and communication, has also made it a popular platform for the spread of rumors. Rumors could be easily propagated and received by a large number of users in social media, resulting in catastrophic effects in the physical world in a very short period. It is a challenging task, if not impossible, to apply classical supervised learning methods to the early detection of rumors, since the labeling process is time-consuming and labor-intensive. Motivated by the fact that abundant label information of historical rumors is publicly available, in this paper, we propose to investigate whether knowledge learned from historical data could potentially help identify newly emerging rumors. In particular, since a disputed factual claim arouses certain reactions such as curiosity, skepticism, and astonishment, we identify and utilize patterns from prior labeled data to help reveal emergent rumors. Experimental results on real-world data sets demonstrate the effectiveness. Further experiments are conducted to show how much earlier it can detect an emerging rumor than traditional approaches.

Liang Wu, Jundong Li
Arizona State University
wuliang@asu.edu, jundongl@asu.edu

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

Huan Liu
Arizona State University
huanliu@asu.edu

CP3

Detecting Malicious Behavior in Computer Networks Via Cost-Sensitive and Connectivity Constrained Classification

The detection of malicious behavior, that is, judging if a host/domain is malicious or benign (i.e., negative or positive labels), is complicated by the issues of imbalanced label distributions, as well as the limited amount of ground truth available to train supervised models. To tackle these challenges, we propose a novel framework to learn cost-sensitive models on both network hosts and external domains simultaneously, based on a bipartite connectivity graph between them. We also explicitly incorporate behavioral features of the hosts from the network data as well as lexical and reputational features for the external domains into the proposed framework. Specifically, we model the predicted labels, measure the misclassification errors by the Hamming distance between the predicted and true labels, incorporate different costs for false negative or false positive errors, and constrain connected nodes to share the same labels in high probability. The proposed framework is then formulated as an optimization task to minimize the total cost (i.e., the misclassification costs multiplied by the misclassification errors). As the Hamming distance is non-differentiable, we use a continuous loss function to approximate it with performance guaranteed. We develop an effective algorithm with good convergence property via Stochastic Gradient Descent technique. Experiments on both synthetic and a real network data collected from an enterprise show the effectiveness of the proposed framework.

Houping Xiao
SUNY Buffalo
houpingx@buffalo.edu

Jing Gao
University at Buffalo
jing@buffalo.edu

Long Vu
IBM T.J. Watson Research Center
lhvu@us.ibm.com

Deepak Turaga
IBM Research
turaga@us.ibm.com

CP4

Time-Aware Subscription Prediction Model for User Acquisition in Digital News Media

User acquisition is one of the most challenging problems for online news providers. In fact, due to availability of different news media, users have a lot of choices in selecting the news source. To date, most of digital news portals have tried to approach the solution indirectly by targeting the user satisfaction through the recommendation systems. In contrast, we address the problem directly by identifying valuable visitors who are likely potential subscribers in the future. First, we suggest that the decision for subscription is not a sudden, instantaneous action, but is the informed decision based on positive experience with digital medium. As such, we propose effective engagement measures and show that they are effective in building the predictive model for subscription. We design a model that not only predicts the potential subscribers but also answers queries about the subscription occurrence time. The proposed model can be used to predict the subscription time and recommend accurately the "potential users" to the current marketing campaign. We evaluate the proposed model using a real dataset from The Globe and Mail which is a major newspaper in Canada. The experimental results show that the proposed model outperforms the traditional state-of-the-art approaches significantly.

Heidar Davoudi, Morteza Zihayat, Aijun An
York University
davoudi@cse.yorku.ca, zihayatm@cse.yorku.ca,

aan@cse.yorku.ca

CP4

The Power of Certainty: A Dirichlet-Multinomial Model for Belief Propagation

Given a friendship network, how certain are we that Smith is a progressive (vs. conservative)? How can we propagate these certainties through the network? While Belief propagation marked the beginning of principled label-propagation to classify nodes in a graph, its numerous variants proposed in the literature fail to take into account uncertainty during the propagation process. As we show, this limitation leads to counter-intuitive results for even simple graphs. Motivated by these observations, we formalize axioms that any node classification algorithm should obey and propose `NetConf` which satisfies these axioms and handles arbitrary network effects (homophily / heterophily) at scale. Our contributions are: (1) *Axioms*: We state axioms that any node classification algorithm should satisfy; (2) *Theory*: `NetConf` is grounded in a Bayesian-theoretic framework to model uncertainties, has a closed-form solution and comes with precise convergence guarantees; (3) *Practice*: Our method is easy to implement and scales linearly with the number of edges in the graph. On experiments using real world data, we always match or outperform `BP` while taking less processing time.

Dhivya Eswaran
Carnegie Mellon University
deswaran@cs.cmu.edu

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

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

CP4

Predict Land Covers with Transition Modeling and Incremental Learning

Successful land cover prediction can provide promising insights in the applications where manual labeling is extremely difficult. However, traditional machine learning models are plagued by temporal variation and noisy features when directly applied to land cover prediction. Moreover, these models cannot take fully advantage of the spatio-temporal relationship involved in land cover transitions. In this paper, we propose a novel spatio-temporal framework to discover the transitions among land covers and at the same time conduct classification at each time step. Based on the proposed model, we incrementally update the model parameters in the prediction process, thus to mitigate the impact of the temporal variation. Our experiments in two challenging land cover applications demonstrate the superiority of the proposed method over multiple baselines. In addition, we show the efficacy of spatio-temporal transition modeling and incremental learning through extensive analysis.

Xiaowei Jia
University of Minnesota, Twin Cities
jiaxx221@umn.edu

Ankush Khandelw, Guruprasad Nayak, James Gerber
University of Minnesota
ankush@cs.umn.edu, nayak@cs.umn.edu,
jsgerber@umn.edu

Kimberly Carlson
University of Hawai'i Manoa
kimberly.carlson@hawaii.edu

Paul West, Vipin Kumar
University of Minnesota
pcwest@umn.edu, kumar@cs.umn.edu

CP4

Generalized Inverse Classification

Inverse classification is the process of perturbing an instance in a meaningful way such that it is more likely to conform to a specific class. Historical methods that address such a problem are often framed to leverage only a single classifier, or specific set of classifiers. These works are often accompanied by naive assumptions. In this work we propose generalized inverse classification (GIC), which avoids restricting the classification model that can be used. We incorporate this formulation into a refined framework in which GIC takes place. Under this framework, GIC operates on features that are immediately actionable. Each change incurs an individual cost, either linear or nonlinear. Such changes are subjected to occur within a specified level of cumulative change (budget). Furthermore, our framework incorporates the estimation of features that change as a consequence of direct actions taken (indirectly changeable features). To solve such a problem, we propose three real-valued heuristic-based methods and two sensitivity analysis-based comparison methods, each of which is evaluated on two freely available real-world datasets. Our results demonstrate the validity and benefits of our formulation, framework, and methods.

Michael T. Lash, Qihang Lin, Nick Street, Jennifer Robinson, Jeffrey Ohlmann
University of Iowa
michael-lash@uiowa.edu, qihang-lin@uiowa.edu, nick-street@uiowa.edu, jennifer-g-robinson@uiowa.edu, jeffrey-ohlmann@uiowa.edu

CP4

From Theory to Practice: Efficient Active Cost-Sensitive Classification with Expected Error Reduction

In many classification tasks, the data distribution is imbalanced and different misclassifications involve different costs. In addition, the data collected are often lack in labels and it is expensive and tedious to label them manually. Motivated by these two problems, we propose a novel active cost-sensitive classification algorithm based on the Expected Error Reduction (EER) framework, aiming to selectively label examples which can directly optimize the expected misclassification costs. However, the native EER (N-EER) framework is inefficient and impractical due to the considerable requirement for model retraining. In this paper, we propose an efficient EER (E-EER) to overcome the inefficiency of N-EER with the application of cost-sensitive classification which is realized by incorporating the cost information into the expected loss calculation. We first present a formal formulation for EER, then the active cost-sensitive classification algorithm is derived. In order to achieve E-EER, we derive an efficient model update rule for logistic regression (LR) and cost-sensitive support vector machines (C-SVM), respectively, to avoid

model retraining, which are employed as the base learners. Furthermore, we theoretically analyze the error bound of our algorithm to provide a guarantee for its generalization performance. Extensive experiments demonstrate the effectiveness and efficiency of our method.

Yexun Zhang, Yanfeng Wang
Cooperative Medianet Innovation Center
Shanghai Jiao Tong University
zhyxun@sjtu.edu.cn, wangyanfeng@sjtu.edu.cn

Wenbin Cai
Applications & Services Group East Asia
Microsoft, Beijing
wenbca@microsoft.com

Siyuan Zhou, Ya Zhang
Cooperative Medianet Innovation Center
Shanghai Jiao Tong University
zhousiyuan@sjtu.edu.cn, ya_zhang@sjtu.edu.cn

CP5

T-Bne: Tensor-Based Brain Network Embedding

Brain network embedding is the process of converting brain network data to discriminative representations of subjects, so that patients with brain disorders and normal controls can be easily separated. Computer-aided diagnosis based on such representations is potentially transformative for investigating disease mechanisms and for informing therapeutic interventions. However, existing methods either limit themselves to extracting graph-theoretical measures and subgraph patterns, or fail to incorporate brain network properties and domain knowledge in medical science. In this paper, we propose t-BNE, a novel Brain Network Embedding model based on constrained tensor factorization. t-BNE incorporates 1) symmetric property of brain networks, 2) side information guidance to obtain representations consistent with auxiliary measures, 3) orthogonal constraint to make the latent factors distinct with each other, and 4) classifier learning procedure to introduce supervision from labeled data. The Alternating Direction Method of Multipliers (ADMM) framework is utilized to solve the optimization objective. We evaluate t-BNE on three EEG brain network datasets. Experimental results illustrate the superior performance of the proposed model on graph classification tasks with significant improvement 20.51%, 6.38% and 12.85%, respectively. Furthermore, the derived factors are visualized which could be informative for investigating disease mechanisms under different emotion regulation tasks.

Bokai Cao
University of Illinois at Chicago
caobokai@uic.edu

Lifang He
South China University of Technology
lifanghescut@gmail.com

Xiaokai Wei
University of Illinois at Chicago
xwei2@uic.edu

Mengqi Xing
UIC
mxing3@uic.edu

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

Heide Klumpp
UIC
hklumpp@psych

Alex Leow
University of Illinois, Chicago
alexfeuillet@gmail.com

CP5

Clustering with Domain-Specific Usefulness Scores

Clustering is a challenging problem because given the same data set, it can be grouped in multiple different ways. Which of these clustering solutions is interesting depends on its domain application. Thus, incorporating domain expert input often improves clustering performance. However, most existing semi-supervised clustering techniques can only incorporate instance-level constraints (a few labels or must-link/cannot-link constraints), which domain experts may not be comfortable providing in knowledge discovery problems because categories are not known. Fortunately, domain experts often have an idea regarding properties that clustering solutions should have in order to be useful in domain application based on domain relevant scores. In this paper, we provide a framework for jointly optimizing the usefulness and quality of a clustering solution. Experiments on a synthetic data, a benchmark data, and a real-world disease subtyping problem demonstrate the usefulness of our proposed approach.

Yale Chang
Northeastern University
ychang@coe.neu.edu

Junxiang Chen
Northeastern University
Electrical & Computer Engineering Department
jchen@ece.neu.edu

Michael Cho, Peter Castaldi, Edwin Silverman
Brigham and Women's Hospital
Harvard Medical School
remhc@channing.harvard.edu,
repjc@channing.harvard.edu, reeks@channing.harvard.edu

Jennifer Dy
Northeastern University
jdy@ece.neu.edu

CP5

An Rnn Architecture with Dynamic Temporal Matching for Personalized Predictions of Parkinson's Disease

Parkinson's disease (PD) is a chronic disease that develops over years and varies dramatically in its clinical manifestations. A preferred strategy to resolve this heterogeneity and thus enable better prognosis and targeted therapies is to segment out more homogeneous patient subpopulations. However, it is challenging to evaluate the clinical similarities among patients because of the longitudinality and temporality of their records. To address this issue, we propose a deep model that directly learns patient similarity from longitudinal and multi-modal patient records with an Recurrent Neural Network (RNN) archi-

tecture, which learns the similarity between two longitudinal patient record sequences through dynamically matching temporal patterns in patient sequences. Evaluations on real world patient records demonstrate the promising utility and efficacy of the proposed architecture in personalized predictions.

Chao Che
Key Laboratory of Advanced Design and Intelligent Computing
Ministry of Education, Dalian University
chechao@gmail.com

Cao Xiao
Univ. of Washington and IBM T.J. Watson Research Center
danicaxiao@gmail.com

Jian Liang
Dept. of Automation, Tsinghua University
liangjian_work@126.com

Bo Jin
Computer Science, Dalian University of Technology
jinbo@dlut.edu.cn

Jiayu Zhou
Computer Science and Eng., Michigan State University
dearjiayu@gmail.com

Fei Wang
IBM T J Watson Research Center
feiwang03@gmail.com

CP5

Brainzoom: High Resolution Reconstruction from Multi-Modal Brain Signals

How close can we zoom in to observe brain activity? Our understanding is limited by the resolution of imaging modalities that exhibit good spatial but poor temporal resolution, or vice-versa. In this paper, we propose BRAINZOOM, an efficient algorithm that cross-leverages multi-modal brain signals. BRAINZOOM *(a) constructs high resolution* brain images from multi-modal signals, *(b) is scalable*, and *(c) is flexible* in that it can easily incorporate various priors on the brain activity, such as sparsity, low rank, or smoothness. We carefully formulate the problem to tackle nonlinearity in the measurements (via *variable splitting*) and *auto-scale* between different modal signals, and judiciously design an *inexact alternating optimization*-based algorithmic framework to handle the problem with provable convergence guarantees. Our experiments, using a realistic brain simulator to generate fMRI and MEG signals, demonstrate that high spatio-temporal resolution brain imaging is possible from these two modalities. The experiments also suggest that smoothness seems to be the best prior, among several we tried.

Xiao Fu
University of Minnesota - Twin Cities
University of Minnesota
xfu@umn.edu

Kejun Huang
University of Minnesota
huang663@umn.edu

Otilia Stretcu, Hyun Ah Song
Carnegie Mellon University
ostretcu@cs.cmu.edu, hyunahs@cs.cmu.edu

Evangelos Papalexakis
University of California, Riverside
epapalex@cs.ucr.edu

Partha Talukdar
Indian Institute of Science
partha@talukdar.net

Tom Mitchell
Carnegie Mellon University
tom.mitchell@cs.cmu.edu

Nicholas Sidiropoulos
University of Minnesota
nikos@umn.edu

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

CP5

Unified and Contrasting Graphical Lasso for Brain Network Discovery

The analysis of brain imaging data has attracted much attention recently. A popular analysis is to discover a network representation of brain from the neuroimaging data, where each node denotes a brain region and each edge represents a functional association or structural connection between two brain regions. Motivated by the multi-subject and multi-collection settings in neuroimaging studies, in this paper, we consider brain network discovery under two novel settings: 1) *unified setting:* Given a collection of subjects, discover a single network that is good for all subjects. 2) *contrasting setting*: Given two collections of subjects, discover a *single* network that best discriminates two collections. We show that the existing formulation of graphical Lasso (GLasso) cannot address above problems properly. Two novel models, UGLasso (Unified Graphical Lasso) and CGLasso (Contrasting Graphical Lasso), are proposed to address these two problems respectively. We evaluate our methods on synthetic data and two real-world functional magnetic resonance imaging (fMRI) datasets. Empirical results demonstrate the effectiveness of the proposed methods.

Xinyue Liu, Xiangnan Kong
Worcester Polytechnic Institute
xliu4@wpi.edu, xkong@wpi.edu

Ann Ragin
Northwestern University
ann-ragin@northwestern.edu

CP6

A Method to Accelerate Human in the Loop Clustering

Abstract not available at time of publication.

Anni Coden, Marina Danilevsky, Daniel Gruhl, Linda Kato, Meenakshi Nagarajan
IBM Research
anni@us.ibm.com, mdanile@us.ibm.com, dgruhl@us.ibm.com, kato@us.ibm.com, meenanagara-

jan@us.ibm.com

CP6

Uncovering Group Level Insights with Accordant Clustering

Clustering is a widely-used data mining tool, which aims to discover partitions of similar items in data. We introduce a new clustering paradigm, *accordant clustering*, which enables the discovery of (predefined) group level insights. Unlike previous clustering paradigms that aim to understand relationships amongst the individual members, the goal of accordant clustering is to uncover insights at the group level through the analysis of their members. Group level insight can often support a call to action that cannot be informed through previous clustering techniques. We propose the first accordant clustering algorithm, and prove that it finds near-optimal solutions when data possesses inherent cluster structure. The insights revealed by accordant clusterings enabled experts in the field of medicine to isolate successful treatments for a neurodegenerative disease, and those in finance to discover patterns of unnecessary spending.

Amit Dhurandhar
IBM Research
adhuran@us.ibm.com

Margareta Ackerman
San Jose State University
margareta.ackerman@sjsu.edu

Xiang Wang
Google
xiangwa@google.com

CP6

Active Positive-Definite Matrix Completion

In many applications, e.g., recommender systems and biological data analysis, the datasets of interest are positive definite (PD) matrices. Such matrices are usually similarity matrices, obtained by the multiplication of a matrix of preferences or observations with its transpose. Oftentimes, such real-world matrices are missing many entries and a fundamental data-analysis task, known by the term PD-matrix completion, is the inference of these missing entries. In this paper, we introduce the active version of PD-matrix completion, in which we assume access to an oracle that, at a given cost, returns the value of an unobserved entry of the PD matrix. In this setting, we consider the following question: given a fixed budget, which entries should we query so that the completion of the new matrix is much more indicative of the underlying data?. The main contribution of the paper is the formalization of the above question as the ActivePDCompletion problem and the design of novel and effective algorithms for solving it in practice.

Charalampos Mavroforakis
Boston University
Boston University
cmav@cs.bu.edu

Dora Erdos, Mark Crovella, Evimari Terzi
Boston University
edori@cs.bu.edu, crovella@bu.edu, evimaria@cs.bu.edu

CP6

Targeted Matrix Completion

Matrix completion is a problem that arises in many data-analysis settings where the input consists of a partially-observed matrix (e.g., recommender systems, traffic matrix analysis etc.). Classical approaches to matrix completion assume that the input partially-observed matrix is low rank. The success of these methods depends on the number of observed entries and the rank of the matrix; the larger the rank, the more entries need to be observed in order to accurately complete the matrix. In this paper, we deal with matrices that are not necessarily low rank themselves, but rather they contain low-rank submatrices. We propose Targeted, which is a general framework for completing such matrices. In this framework, we first extract the low-rank submatrices and then apply a matrix-completion algorithm to these low-rank submatrices as well as the remainder matrix separately. Although for the completion itself we use state-of-art completion methods, our results demonstrate that Targeted achieves significantly smaller other classical matrix-completion methods. One of the key technical contributions of the paper lies in the identification of the low-rank submatrices from the input partially-observed matrices.

Natali Ruchansky
Univiersity of Southern California
natalir@bu.edu

Evimaria Terzi, Mark Crovella
Boston University
evimaria@cs.bu.edu, crovella@cs.bu.edu

CP6

Model-Based Von Mises-Fisher Co-Clustering with a Conscience

Co-clustering has proven effective to deal with high dimensional sparse data, such as document-term matrices encountered in text mining. Apart from being high dimensional and sparse, the data sets from the aforementioned domain are also directional in nature. Most existing co-clustering approaches are, however, based on popular modeling assumptions, such as Gaussian or Multinomial, which are inadequate for directional data. Moreover, it is well known that, due to high dimensionality and sparsity, co-clustering approaches, like one-sided clustering methods, tend to generate highly skewed solutions with very unbalanced or even empty clusters, especially when the number of required clusters is large. In this paper, we rely on the recently proposed block von Mises-Fisher mixture model (dbmovMFs), which constitutes a general framework for co-clustering directional data distributed on the surface of a unit hypersphere, i.e, L_2 normalized data. In order to overcome the above difficulties, we propose to modify dbmovMFs in a principled way by introducing a conscience mechanism which discourages bad local solutions having empty or very small/large clusters. This gives rise to a new scalable co-clustering algorithm which is guaranteed to increase monotonically a spherical k-means like criterion by intertwining row and column clusterings at each step. Moreover, empirical results, on several real-world datasets, provide strong support for the effectiveness of the proposed approach.

Aghiles Salah

University of Paris Descartes
aghiles-salah@hotmail.fr

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

CP7

A Dual-Tree Algorithm for Fast k-Means Clustering With Large k

k-means is a widely used clustering algorithm, but for k clusters and a dataset size of N, each iteration of Lloyd's algorithm costs $O(kN)$ time. This is problematic because increasingly, applications of k-means involve both large N *and* large k, and there are no accelerated variants that handle this situation. To this end, we propose a dual-tree algorithm that gives the *exact* same results as standard k-means; when using cover trees, we bound the single-iteration runtime of the algorithm as $O(N + k \log k)$, under some assumptions. To our knowledge these are the first sub-$O(kN)$ bounds for exact Lloyd iterations. The algorithm performs competitively in practice, especially for large N and k in low dimensions. Further, the algorithm is tree-independent, so any type of tree may be used.

Ryan Curtin
Symantec Corporation
ryan@ratml.org

CP7

Specious Rules: An Efficient and Effective Unifying Method for Removing Misleading and Uninformative Patterns in Association Rule Mining

Abstract not available at time of publication.

Wilhelmiina Hämäläinen
Aalto University, Department of Computer Science
wilhelmiina.hamalainen@aalto.fi

Geoff Webb
Monash University
geoff.webb@monash.edu

CP7

A Sparse Nonlinear Classifier Design Using Auc Optimization

AUC (Area under the ROC curve) is an important performance measure for applications where the data is highly imbalanced. Efficient AUC optimization is a challenging research problem as the objective function is non-decomposable and non-continuous. Using a max-margin based surrogate loss function, AUC optimization problem can be approximated as a pairwise RankSVM learning problem. Batch learning algorithms for solving the kernelized version of this problem suffer from scalability issues. Therefore, recent years have witnessed an increased interest in the development of online or single-pass algorithms that design a nonlinear classifier by maximizing the AUC performance. However, on many real-world datasets, the AUC performance of these classifiers was observed to be inferior to that of the classifiers designed using batch learning algorithms. Further, many practical imbalanced data classification problems demand fast inference, which underlines the need for designing sparse nonlinear classifiers. Motivated by these observations, we design a scalable algorithm for maximizing the AUC performance by greedily adding the required number of basis functions into the classifier model. The resulting sparse classifier performs faster inference and its AUC performance is comparable with that of the classifier designed using batch mode. Our experimental results show that the level of sparsity achievable can be an order of magnitude larger than that achieved by the Kernel RankSVM model.

Vishal Kakkar
vishal.kakkar@csa.iisc.ernet.in
vishalkakkar90@gmail.com

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

S Sundararajan
Microsoft Research, India
ssrajan@microsoft.com

Dinesh Garg
IIT Gandhinagar, India
dgarg@iitgn.ac.in

CP7

Multi-Core K-Means

Today's microprocessors consist of multiple cores each of which can perform multiple additions, multiplications, or other operations simultaneously in one clock cycle. To maximize performance, two types of parallelism must be applied in a data mining algorithm: MIMD (Multiple Instruction Multiple Data) where different CPU cores execute different code and follow different threads of control, and SIMD (Single Instruction Multiple Data) where within a core, the same operation is executed at once on various data. In this paper, we consider the wide-spread clustering algorithm K-means as a highly relevant use-case for knowledge discovery on big data. We propose Multi-core K-Means (MKM), a completely re-engineered clustering algorithm which applies MIMD and SIMD parallelism. MKM uses a sophisticated strategy for the access of data vectors and cluster representatives to minimize data transfer between main memory, cache, and registers. For SIMD parallelism it is also essential to avoid branching operations like if-then: we propose to code cluster IDs and distances in joint variables to perform the argmin operation SIMD-parallel and without any branching. Our experiments demonstrate a speed-up which is almost linear in the number of cores. On a pair of shared-memory quad-core processors, MKM is between 95 and 140 times faster than non-parallel K-means, 4-6 times faster than auto-vectorized fully parallel standard K-means, and 2.1 times faster than K-means based on BLAS.

Christian Böhm
Ludwig-Maximilians-Universität München
boehm@dbs.ifi.lmu.de

Martin Perdacher, Claudia Plant
University of Vienna
martin.perdacher@univie.ac.at, claudia.plant@univie.ac.at

CP7

Indexing and Classifying Gigabytes of Time Series

under Time Warping

Time series classification maps time series to labels. The nearest neighbour algorithm (NN) using the Dynamic Time Warping (DTW) similarity measure is a leading algorithm for this task. NN compares each time series to be classified to every time series in the training database. With a training database of N time series of lengths L, each classification requires $\vartheta(N \cdot L^2)$ computations. The databases used in almost all prior research have been relatively small (with less than $10,000$ samples) and much of the research has focused on making DTW's complexity linear with L, leading to a runtime complexity of $O(N \cdot L)$. As we demonstrate with an example in remote sensing, real-world time series databases are now reaching the million-to-billion scale. This wealth of training data brings the promise of higher accuracy, but raises a significant challenge because N is becoming the limiting factor. As DTW is not a metric, indexing objects induced by its space is extremely challenging. We tackle this task in this paper. We develop TSI, a novel algorithm for Time Series Indexing which combines a hierarchy of K-means clustering with DTW-based lower-bounding. We show that, on large databases, TSI makes it possible to classify time series orders of magnitude faster than the state of the art.

Chang Wei Tan
Monash University
chang.tan@monash.edu

Geoffrey I Webb
Faculty of Information Technology
Monash University
geoff.webb@monash.edu

Francois Petitjean
Monash University
francois.petitjean@monash.edu

CP8

Subnetworks MIning with Spatial and Temporal Smoothness

For many real-world applications, data is represented in the form of networks with dynamic structures and attributes. The dynamic changes not only happen locally at nodes/edges of the network but manifest as coordinated changes over subnetworks, thus forming network processes. The need to understand what these local network processes are, how they evolve and consequently impact the global network states has become increasingly important. In this paper, we explore these questions and design a novel multinomial logistic regression algorithm for mining a succinct set of subnetworks that are predictive of the progression of global network states. We characterize each global state of networks by a parameterized function whose coefficients are learnt subject to both spatial and temporal network constraints. The L_1 norm is further imposed to remove irrelevant edges that have little or no impact on global network states, and we prove that the combined objective function is convex, thus can be efficiently solved via steepest gradient descent. Extensive experimental analysis on both synthetic and real work datasets demonstrates the effectiveness of our algorithm against competing methods, not only in the prediction accuracy but also in terms of domain relevance of the discovered brain subnetworks.

Xuan-Hong Dang, Hongyuan You
Department of Computer Science
University of California Santa Barbara
xdang@cs.ucsb.edu, hyou@cs.ucsb.edu

Ambuj Singh
UCSB
ambuj@cs.ucsb.edu

Scott Grafton
Department of Psychological and Brain Sciences
University of California, Santa Barbara, CA
scott.grafton@psych.ucsb.edu

CP8

Finding Low-Tension Communities

Motivated by applications that arise in online social media and collaboration networks, there has been a lot of work on community-search. In this class of problems, the goal is to find a subgraph that satisfies a certain connectivity requirement and contains a given collection of seed nodes. In this paper, we extend the community-search problem by associating each individual with a profile. The profile is a numeric score that quantifies the position of an individual with respect to a topic. We adopt a model where each individual starts with a latent profile and arrives to a conformed profile through a dynamic conformation process, which takes into account the individual's social interaction and the tendency to conform with one's social environment. In this framework, social tension arises from the differences between the conformed profiles of neighboring individuals as well as from the differences between individuals' conformed and latent profiles. Given a network of individuals, their latent profiles and this conformation process, we extend the community-search problem by requiring the output subgraphs to have low social tension. From the technical point of view, we study the complexity of this problem and propose algorithms for solving it effectively. Our experimental evaluation in a number of social networks reveals the efficacy and efficiency of our methods.

Esther Galbrun
Department of Computer Science
University of Helsinki, Finland
esther.galbrun@inria.fr

Behzad Golshan
Recruit Institute of Technology
CA, USA
behzad@recruit.ai

Aristides Gionis
Aalto University
Finland
aristides.gionis@aalto.fi

Evimaria Terzi
Boston University
evimaria@cs.bu.edu

CP8

Sensitivity of Community Structure to Network Uncertainty

Community detection constitutes an important task for investigating the internal structure of networks, with a plethora of applications in a wide range of disciplines. A particularly important point, which is rarely taken into account while developing community detection algorithms, is their sensitivity (or stability) to network uncertainty. In

many cases, the input graph data is incomplete or noisy (e.g., due to noise introduced during the collection of the data or for privacy preserving reasons). Then, the following question arises: how stable are the results produced by an algorithm with respect to the uncertainty (i.e., noise level) of the input data? In this paper, we propose a quantitative way to address this problem. We have considered several graph perturbation models to introduce uncertainty to the graph. Then, we examine the sensitivity of an algorithm, with respect to functional and structural characteristics of the detected communities under various perturbation levels. We have studied the performance of some of the most widely used community detection algorithms in practice, and our experimental results indicate that random walk based community detection algorithms tend to be robust under various conditions of network uncertainty.

Marc Mitri
Ecole Polytechnique
marc.mitri@polytechnique.edu

Fragkiskos D. Malliaros
UC San Diego
fmalliaros@ucsd.edu

Michalis Vazirgiannis
Ecole Polytechnique
mvazirg@lix.polytechnique.fr

CP8

Signed Network Embedding in Social Media

Network embedding is to learn low-dimensional vector representations for nodes of a given social network, facilitating many tasks in social network analysis such as link prediction. The vast majority of existing embedding algorithms are designed for unsigned social networks or social networks with only positive links. However, networks in social media could have both positive and negative links, and little work exists for signed social networks. From recent findings of signed network analysis, it is evident that negative links have distinct properties and added value besides positive links, which brings about both challenges and opportunities for signed network embedding. In this paper, we propose a deep learning framework SiNE for signed network embedding. The framework optimizes an objective function guided by social theories that provide a fundamental understanding of signed social networks. Experimental results on two real-world datasets of social media demonstrate the effectiveness of the proposed framework SiNE.

Suhang Wang
Arizona State University
suhang.wang@asu.edu

Jiliang Tang
Michigan State University
tangjili@cse.msu.edu

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

Yi Chang
Huawei Research America
yichang@am.org

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

CP8

MeiKe: Influence-Based Communities in Networks

Given a social network, how to find communities of nodes based on their diffusive characteristics? There exist two important types of nodes, for information propagation: nodes that are influential ('kernel nodes'), and nodes that serve as 'bridges' to boost the diffusion ('media nodes'). How to find these nodes and uncover connections between them? In addition, it is also important to discover the hidden community structure of these nodes, which can help study their interactions, predict links and also understand the information flow in such networks. In this paper, we give an intuitive and novel optimization-based formulation for this task, which aims to discover media nodes as well as community structures of kernel nodes. We prove our task is computationally challenging, and develop an effective and practical algorithm MEIKE (pronounced as 'Mike'). It first obtains media nodes via a new successive summarization based approach, and then finds kernel nodes including their community structures. Experimental results show that MEIKE finds high-quality media and kernel communities which match our expectations and ground-truth (sometimes outperforming non-trivial baselines by 40% in F1-score). Our case studies also demonstrate the applicability of MEIKE on a variety of datasets.

Yao Zhang, Bijaya Adhikari, Steve Jan, B. Aditya Prakash
Virginia Tech
yaozhang@vt.edu, bijaya@cs.vt.edu, tekang@cs.vt.edu, badityap@cs.vt.edu

CP9

HBGG: a Hierarchical Bayesian Geographical Model for Group Recommendation

Location-based social networks such as Foursquare and Plancast have gained increasing popularity. On those sites, users can organize and participate in group activities; hence, recommending venues to a group is of practical importance. In this paper, we study the problem of recommending venues to groups of users and propose a Hierarchical Bayesian Model (HBGG) for this purpose. First, a generative group geographical topic model (GG) which exploits group membership, group mobility regions and group preferences is proposed. And we integrate social structure into one-class collaborative filtering as social-based collaborative filtering (SOCF) to leverage social wisdom. Through the shared latent group features, HBGG connects the group geographical model with SOCF framework for group recommendation. Experimental results on two real datasets show that our methods outperforms the state-of-the-art group recommenders, especially on cold-start user groups.

Ziyu Lu, Hui Li, Nikos Mamoulis, David Cheung
The University of Hong Kong
lvziyuzju@gmail.com, hli2@cs.hku.hk, nikos@cs.hku.hk, dcheung@cs.hku.hk

CP9

Redundancies in Data and Their Effect on the Evaluation of Recommendation Systems: A Case Study

on the Amazon Reviews Datasets

A collection of datasets crawled from Amazon, 'Amazon reviews', is popular in the evaluation of recommendation systems. These datasets, however, contain redundancies (duplicated recommendations for variants of certain items). These redundancies went unnoticed in earlier use of these datasets and thus incurred to a certain extent wrong conclusions in the evaluation of algorithms tested on these datasets. We analyze the nature and amount of these redundancies and their impact in the evaluation of recommendation methods. While the general and obvious conclusion is that redundancies should be avoided and datasets should be carefully preprocessed, we observe more specifically that their impact can be different for different methods. With this work, we also want to raise the awareness of the importance of data quality, model understanding, and appropriate evaluation.

Daniel Basaran
LMU Munich
basaran@cip.ifi.lmu.de

Eirini C. Ntoutsi
Ludwig-Maximilians-Universität München (LMU)
ntoutsi@kbs.uni-hannover.de

Arthur Zimek
University of Southern Denmark,
zimek@imada.sdu.dk

CP9

Price Recommendation on Vacation Rental Websites

Abstract not available at time of publication.

Yang Li
Northwestern Polytechnical University
liyangnpu@mail.nwpu.edu.cn

Suhang Wang
Arizona State University
suhang.wang@asu.edu

Tao Yang, Quan Pan
Northwestern Polytechnical University
yangtao107@nwpu.edu.cn, quanpan@nwpu.edu.cn

Jiliang Tang
Michigan State University
tangjili@cse.msu.edu

CP9

Selection of Negative Samples for One-Class Matrix Factorization

Many recommender systems have only implicit user feedback. The two possible ratings are positive and negative, but only part of positive entries are observed. One-class matrix factorization (MF) is a popular approach for such scenarios by treating some missing entries as negative. Two major ways to select negative entries are by sub-sampling a set with similar size to that of observed positive entries or by including all missing entries as negative. They are referred to as "subsampled' and "full' approaches in this work, respectively. Currently detailed comparisons between these two selection schemes on large-scale data are still lacking. One important reason is that the "full' approach leads to a hard optimization problem after treating all missing entries as negative. In this paper, we successfully develop efficient optimization techniques to solve this challenging problem so that the "full' approach becomes practically viable. We then compare in detail the two approaches "subsampled' and "full' for selecting negative entries. Results show that the "full' approach of including much more missing entries as negative yields better results.

Hsiang-Fu Yu
Amazon
rofuyu@cs.utexas.edu

Mikhail Bilenko
Microsoft
mbilenko@microsoft.com

Chih-Jen Lin
National Taiwan University
cjlin@csie.ntu.edu.tw

CP9

Collaborative User Network Embedding for Social Recommender Systems

To address the issue of data sparsity and cold-start in recommender system, social information (e.g. user-user trust links) has been introduced to complement rating data for improving the performances of traditional model-based recommendation techniques such as matrix factorization (MF) and Bayesian personalized ranking (BPR). Although effective, the utilization of the explicit user-user relationships extracted directly from such social information has three main limitations. First, it is difficult to obtain explicit and reliable social links. Only a small portion of users indicate explicitly their trusted friends in recommender systems. Second, the 'cold-start' users are 'cold' not only on rating but also on socializing. There is no significant amount of explicit social information that can be useful for 'cold-start' users. Third, an active user can be socially connected with others who have different taste/preference. Direct usage of explicit social links may mislead recommendation. To address these issues, we propose to extract implicit and reliable social information from user feedbacks and identify top-k semantic friends for each user. We incorporate the top-k semantic friends information into MF and BPR frameworks to solve the problems of ratings prediction and items ranking, respectively. The experimental results on three real-world datasets show that our methods achieve better results than the state-of-the-art MF with explicit social links and social BPR.

Chuxu Zhang
Rutgers University
cz201@cs.rutgers.edu

Lu Yu
King Abdullah University of Science and Technology
lu.yu@kaust.edu.sa

Yan Wang, Chirag Shah
Rutgers University
yw298@cs.rutgers.edu, chirags@rutgers.edu

Xiangliang Zhang
King Abdullah University of Science and Technology

xiangliang.zhang@kaust.edu.sa

CP10

Condensing Temporal Networks Using Propagation

Modern networks are very large in size and also evolve with time. As their size grows, the complexity of performing network analysis grows as well. Getting a smaller representation of a temporal network with similar properties will help in various data mining tasks. In this paper, we study the novel problem of getting a smaller diffusion-equivalent representation of a set of time-evolving networks. We first formulate a well-founded and general temporal-network condensation problem based on the so-called system-matrix of the network. We then propose NetCondense, a scalable and effective algorithm which solves this problem using careful transformations in sub-quadratic running time, and linear space complexities. Our extensive experiments show that we can reduce the size of large real temporal networks (from multiple domains such as social, co-authorship and email) significantly without much loss of information. We also show the wide-applicability of NetCondense by leveraging it for several tasks: for example, we use it to understand, explore and visualize the original datasets and to also speed-up algorithms for the influence-maximization problem on temporal networks.

Bijaya Adhikari, Yao Zhang, Aditya Bharadwaj, B. Aditya Prakash
Virginia Tech
bijaya@cs.vt.edu, yaozhang@vt.edu, adb@cs.vt.edu, badityap@cs.vt.edu

CP10

Ranking in Heterogeneous Networks with Geo-Location Information

Entity ranking by importance or authority through relational information is an important problem in network science. Most existing work addresses the problem for homogeneous networks. With the emergence of richer networks, with various types of entities and meta-data, it becomes essential to build models that can leverage all available data in a meaningful way. In this work, we consider the ranking problem in heterogeneous information networks (HIN) with side information. Specifically, we introduce a new model called HINside that has two key properties: (i) it explicitly represents the interactions (i.e., authority transfer rates or ATR) between different types of nodes, and (ii) it carefully incorporates the geo-location information of the entities to account for the distance and the competition between them. Besides an intuitive local formula, our model has a matrix form for which we derive a closed-form solution. Thanks to its closed form, HINside lends itself to be used within various learning-to-rank objectives, for the estimation of its parameters (the ATR) provided training data. We formulate two kinds of objective functions for parameter learning with efficient estimation procedures. We validate the effectiveness of our proposed model and the learning procedures on samples from two real-world graphs, where we show the advantages of HINside over popular existing models, including Pagerank and degree centrality.

Abhinav Mishra
Stony Brook University
abhmishra@cs.stonybrook.edu

Leman Akoglu
Carnegie Mellon University
lakoglu@cs.cmu.edu

CP10

Toward Personalized Relational Learning

Relational learning exploits relationships among instances manifested in a network to improve the predictive performance of many network mining tasks. Due to its empirical success, it has been widely applied in myriad domains. In many cases, individuals in a network are highly idiosyncratic. They not only connect to each other with a composite of factors but also are often described by some content information of high dimensionality specific to each individual. For example in social media, as user interests are quite diverse and personal; posts by different users could differ significantly. Moreover, social content of users is often of high dimensionality which may negatively degrade the learning performance. Therefore, it would be more appealing to tailor the prediction for each individual while alleviating the issue related to the curse of dimensionality. In this paper, we study a novel problem of Personalized Relational Learning and propose a principled framework PRL to personalize the prediction for each individual in a network. Specifically, we perform personalized feature selection and employ a small subset of discriminative features customized for each individual and some common features shared by all to build a predictive model. On this account, the proposed personalized model is more human interpretable. Experiments on real-world datasets show the superiority of the proposed PRL framework over traditional relational learning methods.

Jundong Li, Liang Wu
Arizona State University
jundongl@asu.edu, wuliang@asu.edu

Osmar Zaiane
University of Alberta
zaiane@ualberta.ca

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

CP10

Graph-Based Semi-Supervised Learning for Relational Networks

We address the problem of semi-supervised learning in relational networks, networks in which nodes are entities and links are the relationships or interactions between them. Typically this problem is confounded with the problem of graph-based semi-supervised learning (GSSL), because both problems represent the data as a graph and predict the missing class labels of nodes. However, not all graphs are created equally. In GSSL a graph is constructed, often from independent data, based on similarity. As such, edges tend to connect instances with the same class label. Relational networks, however, can be more heterogeneous and edges do not always indicate similarity. For instance, instead of links tending to connect nodes with the same class label, they may tend to connect nodes with different class labels (*link-heterogeneity*). Or having the same class label might not imply the same type of connectivity across the whole network (*class-heterogeneity*). Performing classification in networks with different types of heterogeneity is a hard problem that is made harder still by the fact we do not know a-priori the type or level of heterogeneity. In this

work we present two scalable approaches for graph-based semi-supervised learning for the more general case of relational networks. Our methods perform well on a diverse set of networks and do so without prior knowledge of how classes interact.

Leto Peel
Universite catholique de Louvain
leto.peel@uclouvain.be

CP10

Community-Aware Network Sparsifiction

Network sparsification aims to reduce the number of edges of a network while maintaining its structural properties; such properties include shortest paths, cuts, spectral measures, or network modularity. Sparsification has multiple applications, such as, speeding up graph-mining algorithms, graph visualization, as well as identifying the important network edges. In this paper we consider a novel formulation of the network-sparsification problem. In addition to the network, we also consider as input a set of communities. The goal is to sparsify the network so as to preserve the network structure with respect to the given communities. We introduce two variants of the community-aware sparsification problem, leading to sparsifiers that satisfy different connectedness community properties. From the technical point of view, we prove hardness results and devise effective approximation algorithms. Our experimental results on a large collection of datasets demonstrate the effectiveness of our algorithms.

Aristides Gionis
Aalto University
Finland
aristides.gionis@aalto.fi

Rozenshtein Polina
Aalto University
polina.rozenshtein@aalto.fi

Nikolaj Tatti
HIIT, Aalto University
nikolaj.tatti@aalto.fi

Evimaria Terzi
Boston University
evimaria@bu.edu

CP11

A Graduated Non-Convexity Relaxation for Large Scale Seriation

In this work we propose a highly scalable algorithm for solving the combinatorial data analysis problem of seriation. Seriation is a technique for optimizing a permutation of data instances, with respect to some proximity measure such that nearby instances in the linear arrangement are more similar. One consistent objective function for seriation is the 2-SUM minimization problem, which uses the 2-norm between instance locations to penalize non-zero similarity values, and can be written as a quadratic function of the permutation vector. Recently, two convex relaxations of the 2-SUM problem have been proposed, which can be solved as constrained quadratic programs using interior point methods; however, the interior point solvers become expensive when the problem size increases. In this paper we present a graduated non-convexity method for vector-based relaxations of the 2-SUM that yields better approximate solutions and scales to very large problem sizes. We conduct a number of experiments on real and synthetic datasets. The experimental results demonstrate that our proposed algorithm outperforms other approaches that solve the 2-SUM, and is the only competitive approach that can scale to large problem sizes.

Xenophon Evangelopoulos, Austin Brockmeier
Department of Computer Science,
University of Liverpool, U.K.
X.Evangelopoulos@liverpool.ac.uk,
a.j.brockmeier@liverpool.ac.uk

Tingting Mu
School of Computer Science,
University of Manchester, U.K.
tingting.mu@manchester.ac.uk

John Goulermas
Department of Computer Science,
University of Liverpool, U.K.
j.y.goulermas@liverpool.ac.uk

CP11

Outlier Detection for Text Data

The problem of outlier detection is extremely challenging in many domains such as text, in which the attribute values are typically non-negative, and most values are zero. In such cases, it often becomes difficult to separate the outliers from the natural variations in the patterns in the underlying data. In this paper, we present a matrix factorization method, which is naturally able to distinguish the anomalies with the use of low rank approximations of the underlying data. Our iterative algorithm TONMF is based on Block Coordinate Descent (BCD) framework. Our approach has significant advantages over traditional methods for text outlier detection. Finally, we present experimental results illustrating the effectiveness of our method over competing methods.

Ramakrishnan Kannan
Oak Ridge National Laboratories
kannanr@ornl.gov

Hyenkyun Woo
Korea University of Technology and Education
hyenkyun@koreatech.ac.kr

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

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

CP11

Exploring Latent Semantic Factors to Find Useful Product Reviews

Online reviews provided by consumers are a valuable asset for e-Commerce platforms, influencing potential consumers in making purchasing decisions. However, these reviews are of varying quality, with the *useful* ones buried deep within a heap of non-informative reviews. In this work, we attempt to automatically identify review quality in terms of its *helpfulness* to the end consumers. In contrast to previ-

ous works in this domain exploiting a variety of syntactic and community-level features, we delve deep into the *semantics* of reviews as to what makes them useful, providing *interpretable* explanation for the same. We identify a set of *consistency* and *semantic* factors, all from the *text, ratings, and timestamps* of user-generated reviews, making our approach generalizable across all communities and domains. We explore review semantics in terms of several latent factors like the *expertise* of its author, his judgment about the fine-grained *facets* of the underlying product, and his *writing style*. These are cast into a Hidden Markov Model – Latent Dirichlet Allocation (HMM-LDA) based model to *jointly* infer: (i) reviewer expertise, (ii) item facets, and (iii) review helpfulness. Large-scale experiments on *five* real-world datasets from *Amazon* show significant improvement over state-of-the-art baselines in predicting and ranking useful reviews.

Subhabrata Mukherjee
Max-Planck-Institut für Informatik
smukherjee@mpi-inf.mpg.de

Kashyap Popat, Gerhard Weikum
Max Planck Institute for Informatics
kpopat@mpi-inf.mpg.de, weikum@mpi-inf.mpg.de

CP11

User-Guided Cross-Domain Sentiment Classification

Sentiment analysis has been studied for decades, and it is widely used in many real applications such as media monitoring. In sentiment analysis, when addressing the problem of limited labeled data from the target domain, transfer learning, or domain adaptation, has been successfully applied, which borrows information from a relevant source domain with abundant labeled data to improve the prediction performance in the target domain. The key to transfer learning is how to model the relatedness among different domains. For sentiment analysis, a common practice is to assume similar sentiment polarity for the common keywords shared by different domains. However, existing methods largely overlooked the human factor, i.e., the users who expressed such sentiment. In this paper, we address this problem by explicitly modeling the human factor related to sentiment classification. In particular, we assume that the content generated by the same user across different domains is biased in the same way in terms of the sentiment polarity. To this end, we propose a new graph-based approach named U-Cross, which models the relatedness of different domains via both the shared users and keywords. It is non-parametric and semi-supervised in nature. Furthermore, we also study the problem of shared user selection to prevent negative transfer. In the experiments, we demonstrate the effectiveness of U-Cross by comparing it with existing state-of-the-art techniques on three real data sets.

Arun Reddy Nelakurthi, Hang Hang Tong, Ross Maciejewski, Nadya Bliss, Jingrui He
Arizona State University
anelakur@asu.edu, hanghang.tong@asu.edu, rmacieje@asu.edu, nadya.bliss@asu.edu, jingrui.he@asu.edu

CP11

Biclustering: An Application of Dual Topic Models

Biclustering is a data mining technique that allows simultaneous clustering of two variables. A common biclustering task for categorical variables is to find 'heavy' biclusters, i.e., biclusters with high co-occurrence values. Although algorithms have been proposed to extract heavy biclusters, they provide little information about relative importance of each bicluster, as well as importance of the variables for each bicluster. To address these problems, there have been attempts to apply mixture models using information theory or Bayesian method. Although they are able to rank the biclusters and the variables for each bicluster, they do not target at extracting heavy biclusters. Furthermore, these models constrain the search for biclusters in such a way that every cell in the matrix must participate in some bicluster. We attempt to alleviate these limitations using dual topic models. First of all, we develop a *generalized* LDA topic model that extracts dual topics, i.e., topics in opposite directions – row- and column-topics. To obtain better topics, it applies mutual reinforcement, i.e., considering column-topics while constructing row-topics, and vice versa. Heavy biclusters, the high co-occurred relationship, are extracted using thresholds. We show that our model Dual Topic to Biclusters (DT2B) is effective in extracting heavy biclusters by experimenting over a simulated data, a text corpus and a microarray gene expression data.

Daniel Rugeles
Nanyang Technological University
daniel007@e.ntu.edu.sg

Kaiqi Zhao
Nanyang Technological University
kzhao002@e.ntu.edu.sg

Cong Gao
Nanyang Technological University
gaocong@ntu.edu.sg

Manoranjan Dash, Shonali Krishnaswamy
Agency for Science Technology and Research
dashm@i2r.a-star.edu.sg, spkrishna@i2r.a-star.edu.sg

CP12

Correlation by Compression

Discovering correlated variables is one of the core problems in data analysis. Many measures for correlation have been proposed, yet it is surprisingly ill-defined in general. That is, most, if not all, measures make very strong assumptions on the data distribution or type of dependency they can detect. In this work, we provide a general theory on correlation, without making any such assumptions. Simply put, we propose correlation by compression. To this end, we propose two correlation measures based on solid information theoretic foundations, i.e. Kolmogorov complexity. The proposed correlation measures possess interesting properties desirable for any sensible correlation measure. However, Kolmogorov complexity is not computable, and hence we propose practical and computable instantiations based on the Minimum Description Length (MDL) principle. In practice, we can apply the proposed measures on any type of data by instantiating them with any lossless real-world compressors that reward pairwise dependencies. Extensive experiments show that the correlation measures works well in practice, have high statistical power, and find meaningful correlations on binary data, while they are easily extendible to other data types.

Kailash Budhathoki
Max Planck Institute for Informatics
Max Planck Institute for Informatics

kbudhath@mpi-inf.mpg.de

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

CP12

Embedded Supervised Feature Selection for Multi-Class Data

Supervised multi-class learning arises in many application domains such as biology, computer vision, social network analysis, and information retrieval. These applications often involve high-dimensional data, which not only significantly increase the time and space requirement of the underlying algorithms but also degrade their performance due to the curse of dimensionality. Feature selection has been proven effective and efficient for preparing high-dimensional data for many learning tasks. Traditional feature selection algorithms for multi-class data assume the independence of label categories and select features with the capability to distinguish samples from different classes. However, class labels in multi-class data may be correlated and little work exists for exploiting label correlation in multi-class feature selection. In this paper, we investigate label correlation in feature selection for multi-class data. In particular, we provide a principled approach for capturing label correlation and propose an Embedded Supervised Feature Selection (ESFS) framework, which embeds label correlation modeling in supervised feature selection for multi-class data. Experiments on both synthetic data and various types of public benchmark datasets show that the proposed framework effectively captures the multi-class label correlation and significantly outperforms existing state-of-the-art baseline methods.

Lin Chen
Arizona State University
lin.chen.cs@asu.edu

Jiliang Tang
Michigan State University
tangjili@msu.edu

Baoxin Li
Arizona State University
baoxin.li@asu.edu

CP12

A Deflation Method for Structured Probabilistic PCA

Modern treatments of structured PCA often focus on the estimation of a single component under various assumptions or priors, such as sparsity and smoothness, and then the procedure is extended to multiple components by sequential estimation interleaved with deflation. While prior work has highlighted the importance of proper deflation for ensuring the quality of the estimated components, to our knowledge, proposed techniques have only been developed and applied to non-probabilistic PCA, and are not trivially extended to probabilistic analyses. This work introduces a novel and efficient deflation method for Probabilistic PCA using tools recently developed for constrained probabilistic estimation via information projection. The components estimated using the proposed deflation regain some of the interpretability of classic PCA such as straightforward estimates of variance explained, while retaining the ability to incorporate rich prior structure. Moreover, sequential estimation allows for scaling probabilistic techniques to be at par with their deterministic counterparts. Experimental results on simulated data demonstrate the utility of the proposed deflation in terms of component recovery, and evaluation on neuroimaging data show both qualitative and quantitative improvements in the quality of the estimated components. We also present timing experiments on real data to illustrate the importance of sequential estimation with proper deflation for scalability.

Rajiv Khanna, Joydeep Ghosh
UT Austin
rajivak@utexas.edu, jghosh@utexas.edu

Russell Poldrack
Stanford University
poldrack@stanford.edu

Oluwasanmi Koyejo
University of Illinois at Urbana Champaign
sanmi@illinois.edu

CP12

Roflmao: Robust Oblique Forests with Linear Matrix Operations

Random Forests (RF) remains one of the most widely used general purpose classification methods. Two recent large-scale empirical studies demonstrated it to be the best overall classification method among a variety of methods evaluated. One of its main limitations, however, is that it is restricted to making only axis-aligned recursive partitions of the feature space. Consequently, RF is particularly sensitive to the orientation of the data. Several studies have proposed "oblique" decision forest methods to address this limitation. However, these methods either have a time and space complexity significantly greater than RF, are sensitive to unit and scale, or empirically do not perform as well as RF on real data. One promising oblique method that was proposed alongside the canonical RF method, called Forest-RC (F-RC), has not received as much attention by the community. Despite it being just as old as RF, virtually no studies exist investigating its theoretical or empirical performance. In this work, we demonstrate that F-RC empirically outperforms RF and another recently proposed oblique method called Random Rotation Random Forest on a large number of datasets, while approximately maintaining the same computational complexity. Furthermore, a variant of F-RC which rank transforms the data prior to learning is especially invariant to affine transformations and robust to data corruption.

Tyler M. Tomita
Johns Hopkins University
Center for Imaging Science
ttomita@jhu.edu

CP12

Exploiting Hierarchical Structures for Unsupervised Feature Selection

Feature selection has been proven to be effective and efficient in preparing high-dimensional data for many mining and learning tasks. Features of real-world high-dimensional data such as words of documents, pixels of images and genes of microarray data, usually present inherent hierarchical structures. In a hierarchical structure, features could

share certain properties. Such information has been exploited to help supervised feature selection but it is rarely investigated for unsupervised feature selection, which is challenging due to the lack of labels. Since real world data is often unlabeled, it is of practical importance to study the problem of feature selection with hierarchical structures in an unsupervised setting. In particular, we provide a principled method to exploit hierarchical structures of features and propose a novel framework HUFS, which utilizes the given hierarchical structures to help select features without labels. Experimental study on real-world datasets is conducted to assess the effectiveness of the proposed framework.

Suhang Wang, Yilin Wang
Arizona State University
suhang.wang@asu.edu, ywang370@asu.edu

Jiliang Tang
Michigan State University
tangjili@cse.msu.edu

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

Suhas Ranganath, Huan Liu
Arizona State University
srangan8@asu.edu, huan.liu@asu.edu

CP13

BreachRadar: Automatic Detection of Points-of-Compromise

Bank transaction fraud results in over $13B annual losses for banks, merchants, and card holders worldwide. Much of this fraud starts with a Point-of-Compromise (a data breach or a "skimming" operation) where credit and debit card digital information is stolen, resold, and later used to perform fraud. We introduce this problem and present an automatic Points-of-Compromise (POC) detection procedure. BreachRadar is a distributed alternating algorithm that assigns a probability of being compromised to the different possible locations. We implement this method using Apache Spark and show its linear scalability in the number of machines and transactions. BreachRadar is applied to two datasets with *billions* of real transaction records and fraud labels where we provide multiple examples of real Points-of-Compromise we are able to detect. We further show the effectiveness of our method when injecting Points-of-Compromise in one of these datasets, simultaneously achieving over 90% precision and recall when only 10% of the cards have been victims of fraud.

Miguel Araujo
Carnegie Mellon University and INESC-TEC
maraujo@cs.cmu.edu

Miguel Almeida, Jaime Ferreira, Luis Silva, Pedro Bizarro
Feedzai
miguel.almeida@feedzai.com, jaime.ferreira@feedzai.com, luis.silva@feedzai.com, pedro.bizarro@feedzai.com

CP13

Identifying Deep Contrasting Networks from Time Series Data: Application to Brain Network Analysis

The analysis of multiple time series data, which are generated from a networked system, has attracted much attention recently. This technique has been used in a wide range of applications including functional brain network analysis of neuroimaging data and social influence analysis. In functional brain network analysis, the activity of different brain regions can be represented as multiple time series. An important task in the analysis is to identify the latent network from the observed time series data. In this network, the edges (functional connectivity) capture the correlation between different time series (brain regions). Conventional network extraction approaches usually focus on capturing the connectivity through linear measures under unsupervised settings. In this paper, we study the problem of identifying deep nonlinear connections under group-contrasting settings, where we have two groups of time series samples, and the goal is to identify nonlinear connections that are discriminative across the two groups. We propose a method called GCC (Graph Construction CNN) which is based on deep convolutional neural networks. The CNN in our model learns a nonlinear edge-weighting function to assign discriminative values to the edges of a network. Experiments on a real-world ADHD dataset show the effectiveness of the proposed method. We also demonstrate the extensibility of our proposed framework by combining it with an autoencoder.

John Boaz Lee, Xiangnan Kong, Yihan Bao
Worcester Polytechnic Institute
jtlee@wpi.edu, xkong@wpi.edu, ybao@wpi.edu

Constance Moore
University of Massachusetts Medical School
contance.moore@umassmed.edu

CP13

Cumulative Knowledge-based Regression Models for Next-term Grade Prediction

Grade prediction for courses not yet taken by students is important so as to guide them while registering for next-term courses. Moreover, it can help their advisers for designing personalized degree plans and modifying them based on the students' performance. In this paper, we present cumulative knowledge-based regression models with different course-knowledge spaces for the task of next-term grade prediction. These models utilize historical student-course grade data as well as the information available about the courses that capture the relationships between courses in terms of the knowledge components provided by them. Our experiments on a large dataset obtained from the College of Science and Engineering at University of Minnesota show that our proposed methods achieve better performance than competing methods and that these performance gains are statistically significant.

Sara Morsy
University of Minnesota
University of Minnesota
morsy002@umn.edu

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

CP13

Hidden: Hierarchical Dense Subgraph Detection

with Application to Financial Fraud Detection

Dense subgraphs are fundamental patterns in graphs, and dense subgraph detection is often the key step of numerous graph mining applications. Most of the existing methods aim to find a single subgraph with a high density. However, dense subgraphs at different granularities could reveal more intriguing patterns in the underlying graph. In this paper, we propose to *hierarchically* detect dense subgraphs. The key idea of our method (HiDDen) is to envision the density of subgraphs as a *relative* measure to its background (i.e., the subgraph at the coarse granularity). Given that the hierarchical dense subgraph detection problem is essentially a nonconvex quadratic programming problem, we propose effective and efficient alternative projected gradient based algorithms to solve it. The experimental evaluations on real graphs demonstrate that (1) our proposed algorithms find subgraphs with an up to 40% higher density in almost every hierarchy; (2) the densities of different hierarchies exhibit a desirable variety across different granularities; (3) our projected gradient descent based algorithm scales *linearly* w.r.t the number of edges of the input graph; and (4) our methods are able to reveal interesting patterns in the underlying graphs (e.g., synthetic ID in financial fraud detection).

Si Zhang, Dawei Zhou, Mehmet Yigit Yildirim
Arizona State University
szhan172@asu.edu, dzhou23@asu.edu,
yigityildirim@asu.edu

Scott Alcorn
Early Warnings LLC
scott.alcorn@earlywarning.com

Jingrui He, Hasan Davulcu, Hanghang Tong
Arizona State University
jingrui.he@asu.edu, hasandavulcu@asu.edu, hanghang.tong@asu.edu

CP13

Uplift Modeling with Multiple Treatments and General Response Types

Randomized experiments have been used to assist decision-making in many areas. They help people select the optimal treatment for the test population with certain statistical guarantee. However, subjects can show significant heterogeneity in response to treatments. The problem of customizing treatment assignment based on subject characteristics is known as uplift modeling, differential response analysis, or personalized treatment learning in literature. A key feature for uplift modeling is that the data is unlabeled. It is impossible to know whether the chosen treatment is optimal for an individual subject because response under alternative treatments is unobserved. This presents a challenge to both the training and the evaluation of uplift models. In this paper we describe how to obtain an unbiased estimate of the key performance metric of an uplift model, the expected response. We present a new uplift algorithm which creates a forest of randomized trees. The trees are built with a splitting criterion designed to directly optimize their uplift performance based on the proposed evaluation method. Both the evaluation method and the algorithm apply to arbitrary number of treatments and general response types. Experimental results on synthetic data and industry-provided data show that our algorithm leads to significant performance improvement over other applicable methods.

Yan Zhao, Xiao Fang, David Simchi-Levi
Massachusetts Institute of Technology
zhaoyanmit@gmail.com, fxiao@mit.edu, dslevi@mit.edu

CP13

MultiC^2: An Optimization Framework for Learning from Task and Worker Dual Heterogeneity

Nowadays, crowdsourcing has been commonly used to enlist label information both effectively and efficiently. One major challenge in crowdsourcing is the diverse worker quality, which determines the accuracy of the label information provided by such workers. Motivated by the observation that in many crowdsourcing platforms, the same set of workers typically work on the same set of tasks, we propose to model the diverse worker quality by studying their behaviors across multiple related tasks. To this end, we propose an optimization framework named MultiC^2 for learning from task and worker dual heterogeneity. It uses a weight tensor to represent the workers' behaviors across multiple tasks, and seeks to find the optimal solution of the tensor by exploiting its structured information. We then propose an iterative algorithm to solve the optimization framework and analyze its computational complexity. To infer the true label of an example, we construct a worker ensemble based on the estimated tensor, whose decisions will be weighted using a set of entropy weight. Finally, we test the performance of MultiC^2 on various data sets, and demonstrate its superiority over state-of-the-art crowdsourcing techniques.

Yao Zhou
Arizona State University
Arizona State University
yzhou174@asu.edu

Lei Ying, Jingrui He
Arizona State University
lei.ying.2@asu.edu, jingrui.he@asu.edu

CP14

Near-Optimal and Practical Algorithms for Graph Scan Statistics

Scan statistics have become a popular approach used for detecting "hotspots' and "anomalies' in spatio-temporal and network data. This methodology involves maximizing a score function over all connected subgraphs and are NP-hard in general. A number of heuristics have been proposed for these problems, but they do not provide any quality guarantees. In this paper, we develop a unified framework for designing algorithms for optimizing a large class of parametric and non-parametric scan statistics for networks, subject to connectivity constraints. Our algorithms run in time that scales linearly on the size of the graph and depends on a parameter we call the "effective solution size', while providing rigorous approximation guarantees. In contrast, most prior methods have super-linear running times in terms of graph size. Extensive empirical evidence demonstrates the effectiveness and efficiency of our proposed algorithms in comparison with state-of-the-art methods. Our proposed approach improves on the performance relative to all prior methods, giving up to over 25% increase in the score. Further, our algorithms scale to networks with up to a million nodes, which is 1-2 orders of

magnitude larger than all prior applications.

Jose Cadena
Virginia Tech
jcadena@vbi.vt.edu

Anil Vullikanti
Dept. of Computer Science, and Virginia Bioinformatics Inst.
Virginia Tech
akumar@vbi.vt.edu

Feng Chen
University of Albany-SUNY
fchen5@albany.edu

CP14

Accelerated Attributed Network Embedding

Network embedding is to learn low-dimensional vector representations for nodes in a network. It has shown to be effective in a variety of tasks such as node classification and link prediction. While embedding algorithms on pure networks have been intensively studied, in many real-world applications, nodes are often accompanied with a rich set of attributes or features, aka attributed networks. It has been observed that network topological structure and node attributes are often strongly correlated with each other. Thus modeling and incorporating node attribute proximity into network embedding could be potentially helpful, though non-trivial, in learning better vector representations. Meanwhile, real-world networks often contain a large number of nodes and features, which put demands on the scalability of embedding algorithms. To bridge the gap, in this paper, we propose an accelerated attributed network embedding algorithm AANE, which enables the joint learning process to be done in a distributed manner by decomposing the complex modeling and optimization into many sub-problems. Experimental results on several real-world datasets demonstrate the effectiveness and efficiency of the proposed algorithm.

Xiao Huang
Texas A&M University
xhuang@tamu.edu

Jundong Li
Arizona State University
jundongl@asu.edu

Xia Hu
Texas A&M University
xiahu@tamu.edu

CP14

Absenteeism Detection in Social Media

Event detection in online social media has primarily focused on identifying abnormal spikes, or bursts, in activity. However, disruptive events such as socio-economic disasters, civil unrest, and even power outages, often involve abnormal troughs or lack of activity, leading to absenteeism. We present the first study, to our knowledge, that models absenteeism and uses detected absenteeism instances as a basis for event detection in location-based social networks such as Twitter. The proposed framework addresses the challenges of (i) early detection of absenteeism, (ii) identifying the locus of the absenteeism, and (iii) identifying groups or communities underlying the absenteeism. Our approach uses the formalism of graph wavelets to represent the spatiotemporal structure of user activity in a location-based social network. This formalism facilitates multiscale analysis, enabling us to detect anomalous behavior at different graph resolutions, which in turn allows the identification of event locations and underlying groups. The effectiveness of our approach is evaluated using Twitter activity related to civil unrest events in Latin America.

Fang Jin
Virginia Tech University
jfang8@vt.edu

CP14

Multimodal Network Alignment

A multimodal network encodes relationships between the same set of nodes in multiple settings, and network alignment is a powerful tool for transferring information and insight between a pair of networks. We propose a method for multimodal network alignment that computes a matrix which indicates the alignment, but produces the result as a low-rank factorization directly. We then propose new methods to compute approximate maximum weight matchings of low-rank matrices to produce an alignment. We evaluate our approach by applying it on synthetic networks and use it to de-anonymize a multimodal transportation network.

Huda Nassar, David F. Gleich
Purdue University
hnassar@purdue.edu, dgleich@purdue.edu

CP14

FACETS: Adaptive Local Exploration of Large Graphs

Visualization is a powerful paradigm for exploratory data analysis. Visualizing large graphs, however, often results in excessive edges crossings and overlapping nodes. We propose a new scalable approach called FACETS that helps users *adaptively* explore large million-node graphs from a *local* perspective, guiding them to focus on nodes and neighborhoods that are most subjectively interesting to users. We contribute novel ideas to measure this interestingness in terms of how surprising a neighborhood is given the background distribution, as well as how well it matches what the user has chosen to explore. FACETS uses Jensen-Shannon divergence over information-theoretically optimized histograms to calculate the subjective user interest and surprise scores. Participants in a user study found FACETS easy to use, easy to learn, and exciting to use. Empirical runtime analyses demonstrated FACETS's practical scalability on large real-world graphs with up to 5 million edges, returning results in fewer than 1.5 seconds.

Robert Pienta, Minsuk Kahng
Georgia Institute of Technology
pientars@gatech.edu, kahng@gatech.edu

Zhiyuan Lin
Stanford University
zylin@stanford.edu

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

Partha Talukdar
Indian Institute of Science
ppt@serc.iisc.in

James Abello
Rutgers University
abello@dimacs.rutgers.edu

Ganesh Parameswaran
Yahoo! Inc.
ganeshparameswaran2010@gmail.com

Duen Horng Chau
Georgia Tech
polo@gatech.edu

CP14

Meta-Path Graphical Lasso for Learning Heterogeneous Connectivities

Sparse inverse covariance estimation has attracted lots of interests since it can recover the structure of the underlying Gaussian graphical model. This is a useful tool to demonstrate the connections among objects (nodes). Previous works on sparse inverse covariance estimation mainly focus on learning one single type of connections from the observed activities with a lasso, group lasso or tree-structure penalty. However, in many real-world applications, the observed activities on the nodes can be related to multiple types of connections. In this paper, we consider the problem of learning heterogeneous connectivities from the observed activities by incorporating meta paths extracted from a heterogeneous information network (HIN), an information network with multiple types of nodes and links, into the conventional graphical lasso framework. We aim at extracting the strongest type of relation between any pairs of entities and ignoring other minor relations. Specially, we introduce two novel kinds of constraints: meta path constraints and exclusive constraints, which ensure the unique type of relation among a pair of objects. This problem is highly challenging due to the non-convex optimization. We proposed a method based upon the alternating direction method of multipliers (ADMM) to efficiently solve the problem. The conducted experiments on both synthetic and real-world datasets illustrate the effectiveness of the proposed method.

Yao Zhang, Yun Xiong
Fudan University
zhang_yao15@fudan.edu.cn, yunx@fudan.edu.cn

Xinyue Liu, Xiangnan Kong
Worcester Polytechnic Institute
xliu4@wpi.edu, xkong@wpi.edu

Yangyong Zhu
Fudan University
yyzhu@fudan.edu.cn

CP15

Differentially Private Rank Aggregation

Given a collection of rankings of a set of items, *rank aggregation* seeks to compute a ranking that can serve as a single best representative of the collection. Rank aggregation is a well-studied problem and a number of effective algorithmic solutions have been proposed in the literature. However, when individuals are asked to contribute a ranking, they may be concerned that their personal preferences will be disclosed inappropriately to others. This acts as a disincentive to individuals to respond honestly in expressing their preferences and impedes data collection and data sharing. We address this problem by investigating rank aggregation under differential privacy, which requires that a released output (here, the aggregate ranking computed from individuals' rankings) remain almost the same if any one individual's ranking is removed from the input. We propose a number of differentially-private rank aggregation algorithms: two are inspired by non-private approximate rank aggregators from the existing literature; another uses a novel rejection sampling method to sample privately from a complex distribution. For all the methods we propose, we quantify, both theoretically and empirically, the "cost' of privacy in terms of the quality of the rank aggregation computed.

Michael Hay
Colgate University
mhay@colgate.edu

Liudmila Elagina, Gerome Miklau
University of Massachusetts Amherst
lelagina@cs.umass.edu, miklau@cs.umass.edu

CP15

Private and Right-Protected Big Data Publication: An Analysis

The ease of digital data dissemination has spurred an amplified interest in technologies related to data privacy and right protection. We examine how both goals can be achieved simultaneously by constructing modified data instances that are both differentially private and right protected. The proposed method first produces a sketch of the dataset via random projection and then perturbs the sketch just enough to ensure privacy. The right-protection mechanism inserts small noise in the dataset which subsequently can be used to verify ownership. We provide analytical privacy, right-protection, and utility guarantees. Our utility guarantees ensure approximate preservation of pairwise distances, thus mining operations such as search, classification, and clustering can be performed on the differentially private and right protected dataset.

Michail Vlachos
IBM Research
michalis0@gmail.com

Reinhard Heckel
University of California Berkeley
reinhard.heckel@gmail.com

CP15

Multivariate Confidence Intervals

Confidence intervals are a popular way to visualize and analyze data distributions. Unlike p-values, they can convey information both about statistical significance as well as effect size. However, very little work exists on applying confidence intervals to multivariate data. In this paper we define confidence intervals for multivariate data that extend the one-dimensional definition in a natural way. In our definition every variable is associated with its own confidence interval as usual, but a data vector can be outside of a few of these, and still be considered to be within the confidence area. We analyze the problem and show that the resulting confidence areas retain the good qualities of

their one-dimensional counterparts: they are informative and easy to interpret. Furthermore, we show that the problem of finding multivariate confidence intervals is hard, but provide efficient approximate algorithms to solve the problem.

Jussi Korpela, Emilia Oikarinen, Kai Puolamäki, Antti Ukkonen
Finnish Institute of Occupational Health
jussi.korpela@ttl.fi, emilia.oikarinen@ttl.fi, kai.puolamaki@ttl.fi, antti.ukkonen@ttl.fi

CP15

Statistical Learning Theory Approach for Data Classification with ℓ-Diversity

Corporations are retaining ever-larger corpuses of personal data; the frequency or breaches and corresponding privacy impact have been rising accordingly. One way to mitigate this risk is through use of anonymized data, limiting the exposure of individual data to only where it is absolutely needed. This would seem particularly appropriate for data mining, where the goal is generalizable knowledge rather than data on specific individuals. In practice, corporate data miners often insist on original data, for fear that they might "miss something" with anonymized or differentially private approaches. This paper provides a theoretical justification for the use of anonymized data. Specifically, we show that a support vector classifier trained on anatomized data satisfying ℓ-diversity should be expected to do as well as on the original data. Anatomy preserves all data values, but introduces uncertainty in the mapping between identifying and sensitive values, thus satisfying ℓ-diversity. The theoretical effectiveness of the proposed approach is validated using several publicly available datasets, showing that we outperform the state of the art for support vector classification using training data protected by k-anonymity, and are comparable to learning on the original data.

Koray Mancuhan
Purdue University
kmancuha@purdue.edu

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

CP15

Multi-Task Multiple Kernel Relationship Learning

This paper presents a novel multitask multiple kernel learning framework that efficiently learns the kernel weights leveraging the relationship across multiple tasks. The idea is to automatically infer this task relationship in the *RKHS* space corresponding to the given base kernels. The problem is formulated as a regularization-based approach called *Multi-Task Multiple Kernel Relationship Learning* (*MK-MTRL*), which models the task relationship matrix from the weights learned from latent feature spaces of task-specific base kernels. Unlike in previous work, the proposed formulation allows one to incorporate prior knowledge for simultaneously learning several related tasks. We propose an alternating minimization algorithm to learn the model parameters, kernel weights and task relationship matrix. In order to tackle large-scale problems, we further propose a two-stage *MK-MTRL* online learning algorithm and show that it significantly reduces the computational time, and also achieves performance comparable to that of the joint learning framework. Experimental results on benchmark datasets show that the proposed formulations outperform several state-of-the-art multitask learning methods.

Keerthiram Murugesan
Carnegie Mellon University
kmuruges@cs.cmu.edu

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

CP15

Hash-Based Feature Learning for Incomplete Continuous-Valued Data

Hash-based feature learning is a widely-used data mining approach for dimensionality reduction and for building linear models that are comparable in performance to their nonlinear counterpart. Unfortunately, such an approach is inapplicable to many real-world data sets because they are often riddled with missing values. Substantial data preprocessing is therefore needed to impute the missing values before the hash-based features can be derived. Biases can be introduced during this preprocessing because it is performed independently of the subsequent modeling task, which can result in the models constructed from the imputed hash-based features being suboptimal. To overcome this limitation, we present a novel framework called H-FLIP that simultaneously estimates the missing values while constructing a set of nonlinear hash-based features from the incomplete data. The effectiveness of the framework is demonstrated through experiments using both synthetic and real-world data sets.

Shuai Yuan, Pang-Ning Tan, Kendra Cheruvelil, Emi Last Name, Nicholas Skaff, Patricia Soranno
Michigan State University
yuanshu2@msu.edu, ptan@cse.msu.edu, spencek1@gmail.com, fergusca@msu.edu, nicholas.skaff@gmail.com, soranno@anr.msu.edu

CP16

Learning from Multi-Modality Multi-Resolution Data: An Optimization Approach

Abstract not available at time of publication.

Yada Zhu
IBM T.J. Watson Research Center
yzhu@us.ibm.com

Jianbo Li
Three Bridges Capital
jianboliru@gmail.com

Jingrui He
Arizona State University
jingrui.he@asu.edu

CP16

Limited-Memory Common-Directions Method for Distributed Optimization and Its Application on Empirical Risk Minimization

Distributed optimization has become an important research topic for dealing with extremely large volume of

data available in the Internet companies nowadays. Additional machines make computation less expensive, but inter-machine communication becomes prominent in the optimization process, and efficient optimization methods should reduce the amount of the communication in order to achieve shorter overall running time. In this work, we utilize the advantages of the recently proposed, theoretically fast-convergent common-directions method, but tackle its main drawback of excessive spatial and computational costs to propose a limited-memory algorithm. The result is an efficient, linear-convergent optimization method for parallel/distributed optimization. We further discuss how our method can exploit the problem structure to efficiently train regularized empirical risk minimization (ERM) models. Experimental results show that our method outperforms state-of-the-art distributed optimization methods for ERM problems.

Ching-Pei Lee
University of Wisconsin-Madison
ching-pei@cs.wisc.edu

Po-Wei Wang
Carnegie Mellon University
poweiw@cs.cmu.edu

Weizhu Chen
Microsoft
wzchen@microsoft.com

Chih-Jen Lin
National Taiwan University
cjlin@csie.ntu.edu.tw

CP16

Sparse Graphical Modeling Via Stochastic Complexity

We consider a method for estimating a true sparse model over an exponentially large number of candidates. In view of the minimum description length principle, we propose a novel criterion derived by continuous relaxation of the *stochastic complexity*, together with an efficient algorithm for finding its optimizer. The experimental results on the problem of identifying sparse graphical models indicate that the proposed method consistently discovers underlying true models.

Kohei Miyaguchi, Shin Matsushima
Graduate School of Information Science and Technology
The University of Tokyo
kohei_miyaguchi@mist.i.u-tokyo.ac.jp,
shin_matsushima@mist.i.u-tokyo.ac.jp

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

CP16

Automatic Frankensteining: Creating Complex Ensembles Autonomously

Automating machine learning by providing techniques that autonomously find the best algorithm, hyperparameter configuration and preprocessing is helpful for both researchers and practitioners. Therefore, it is not surprising that automated machine learning has become a very interesting field of research. While current research is mainly focusing on finding good pairs of algorithms and hyperparameter configurations, we will present an approach that automates the process of creating a top performing ensemble of several layers, different algorithms and hyperparameter configurations. These kinds of ensembles are called jokingly Frankenstein ensembles and proved their benefit on versatile data sets in many machine learning challenges. We compare our approach Automatic Frankensteining with the current state of the art for automated machine learning on 80 different data sets and can show that it outperforms them on the majority using the same training time. Furthermore, we compare Automatic Frankensteining on a large scale data set to more than 3,500 machine learning expert teams and are able to outperform more than 3,000 of them within 12 CPU hours.

Martin Wistuba, Nicolas Schilling, Lars Schmidt-Thieme
University of Hildesheim
wistuba@ismll.uni-hildesheim.de, schilling@ismll.uni-hildesheim.de, schmidt-thieme@ismll.uni-hildesheim.de

CP16

A Fast Trust-Region Newton Method for Softmax Logistic Regression

With the emergence of big data, there has been a growing interest in optimization routines that lead to faster convergence of Logistic Regression (LR). Among many optimization methods such as Gradient Descent, Quasi-Newton, Conjugate Gradient, etc., the Trust-region based truncated Newton method (TRON) algorithm has been shown to converge the fastest. The TRON algorithm also forms an important component of the highly efficient and widely used `liblinear` package. It has been shown that the WANBIA-C trick of scaling with the log of the naive Bayes conditional probabilities can greatly accelerate the convergence of LR trained using (first-order) Gradient Descent and (approximate second-order) Quasi-Newton optimization. In this work we study the applicability of the WANBIA-C trick to TRON. We first devise a TRON algorithm optimizing the softmax objective function and then demonstrate that WANBIA-C style preconditioning can be beneficial for TRON, leading to an extremely fast (batch) LR algorithm. Second, we present a comparative analysis of one-vs-all LR and softmax LR in terms of the 0-1 Loss, Bias, Variance, RMSE, Log-Loss, Training and Classification time, and show that softmax LR leads to significantly better RMSE and Log-Loss. We evaluate our proposed approach on 51 benchmark datasets.

Nayyar Zaidi
Monash University
Australia
nayyar.zaidi@monash.edu

Geoff Webb
Monash University
geoff.webb@monash.edu

CP17

Concept Drift Detection with Hierarchical Hypothesis Testing

Abstract not available at time of publication.

Shujian Yu
University of Florida
yusjlcy9011@ufl.edu

Zubin Abraham
Robert Bosch LLC
zubin.abraham@us.bosch.com

CP17

CSTG: An Effective Framework for Cost-Sensitive Sparse Online Learning

Sparse online learning and cost-sensitive learning are two important areas of machine learning and data mining research. Each has been well studied with many interesting algorithms developed. However, very limited published work addresses the joint study of these two fields. In this paper, to tackle the high-dimensional data streams with skewed distributions, we introduce a framework of cost-sensitive sparse online learning. Our proposed framework is a substantial extension of the influential Truncated Gradient (TG) method by formulating a new convex optimization problem, where the two mutual restraint factors, misclassification cost and sparsity, can be simultaneously and favorably balanced. We theoretically analyze the regret and cost bounds of the proposed algorithm, and pinpoint its theoretical merit compared to the existing related approaches. Large-scale empirical comparisons to five baseline methods on eight real-world streaming datasets demonstrate the encouraging performance of the developed method. Algorithm implementation and datasets are available upon request.

Zhong Chen
Xavier Univ. of Louisiana
zchen@xula.edu

Zhide Fang
LSU Health Sciences Center,
zfang@lsuhsc.edu

Wei Fan
Baidu Research Big Data Lab
wei.fan@gmail.com

Andrea Edwards
Xavier Univ. Of Louisiana
aedwards@xula.edu

Kun Zhang
Xavier University of Louisiana
kzhang@xula.edu

CP17

Deep Learning: A Generic Approach for Extreme Condition Traffic Forecasting

Traffic forecasting is a vital part of intelligent transportation systems. However, traffic forecasting is challenging due to short-term (e.g., accidents, constructions) and long-term (e.g., peak-hour, seasonal, weather) impacts. While most previously proposed techniques focus on normal condition forecasting, a single framework for extreme condition traffic forecasting does not exist. In this paper, we propose to take a deep learning approach. We build a deep neural network based on long short term memory (LSTM) units. We apply the deep LSTM to forecast peak-hour traffic and manage to identify unique characteristics of the traffic data. We further improve the architecture for post-accident forecasting and propose the mixture deep LSTM model. Mixture deep LSTM simultaneously models the dynamic behavior of normal condition traffic patterns and automatically learns the representation of accidents. We evaluate our model on a real-world large-scale traffic dataset in Los Angeles to forecast traffic at both peak-hours and post-accidents. When trained end-to-end with suitable regularizations, our approach achieves the state-of-the-art performance, with 30-50% improvement over baselines. We also demonstrate a novel model inspection technique and obtain interesting observations from the trained neural network.

Rose Yu, Yaguang Li, Cyrus Shahabi, Ugur Demiryurek, Yan Liu
University of Southern California
qiyu@usc.edu, yaguang@usc.edu, shahabi@usc.edu, demiryur@usc.edu, yanliu.cs@usc.edu

CP17

H-Fuse: Efficient Fusion of Aggregated Historical Data

In this paper, we address the challenge of recovering a time sequence of counts from aggregated historical data. For example, given a mixture of the monthly and weekly sums, how can we find the daily counts of people infected with flu? In general, what is the best way to recover historical counts from aggregated, possibly overlapping historical reports, in the presence of missing values? Equally importantly, how much should we trust this reconstruction? We propose H-Fuse, a novel method that solves above problems by allowing injection of domain knowledge in a principled way, and turning the task into a well-defined optimization problem. H-Fuse has the following desirable properties: (a) *Effectiveness*, recovering historical data from aggregated reports with high accuracy; (b) *Self-awareness*, providing an assessment of when the recovery is not reliable; (c) *Scalability*, computationally linear on the size of the input data. demonstrates that H-Fuse reconstructs the original data $30-81\%$ better than the least squares method.

Zongge Liu, Hyun Ah Song
Carnegie Mellon University
zonggel@andrew.cmu.edu, hyunahs@cs.cmu.edu

Vladimir Zadorozhny
University of Pittsburgh
vladimir@sis.pitt.edu

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

Nicholas Sidiropoulos
University of Minnesota
nikos@umn.edu

CP17

Error Metrics for Learning Reliable Manifolds from Streaming Data

Spectral dimensionality reduction is frequently used to identify low-dimensional structure in high-dimensional data. However, learning manifolds, especially from the streaming data, is computationally and memory expensive. In this paper, we argue that a stable manifold can be learned using only a fraction of the stream, and the remaining stream can be mapped to the manifold in a significantly less costly manner. Identifying the transition point at which the manifold is stable is the key step. We present error metrics that allow us to identify the transition point for a given stream by quantitatively assessing the quality

of a manifold learned using Isomap. We further propose an efficient mapping algorithm, called S-Isomap, that can be used to map new samples onto the stable manifold. We describe experiments on a variety of data sets that show that the proposed approach is computationally efficient without sacrificing accuracy.

Frank Schoeneman
University at Buffalo
Department of Computer Science & Engineering
fvschoen@buffalo.edu

Suchismit Mahapatra, Varun Chandola, Nils Napp, Jaroslaw Zola
University at Buffalo
suchismi@buffalo.edu, chandola@buffalo.edu, nnapp@buffalo.edu, jzola@buffalo.edu

CP18

Efficiently Summarising Event Sequences with Rich Interleaving Patterns

Discovering the key structure of a database is one of the main goals of data mining. In pattern set mining we do so by discovering a small set of patterns that together describe the data well. The richer the class of patterns we consider, and the more powerful our description language, the better we will be able to summarise the data. In this paper we propose SQUISH, a novel greedy MDL-based method for summarising sequential data using rich patterns that are allowed to interleave. Experiments show SQUISH is orders of magnitude faster than the state of the art, results in better models, as well as discovers meaningful semantics in the form patterns that identify multiple choices of values.

Apratim Bhattacharyya
Max Planck Institute for Informatics and Saarland University
abhattac@mpi-inf.mpg.de

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

CP18

Uncovering the Spatiotemporal Patterns of Collective Social Activity

Social media users and microbloggers post about a wide variety of (off-line) collective social activities as they participate in them, ranging from concerts and sporting events to political rallies and civil protests. In this context, people who take part in the same collective social activity often post closely related content from nearby locations at similar times, resulting in distinctive spatiotemporal patterns. Can we automatically detect these patterns and thus provide insights into the associated activities? In this paper, we propose a modeling framework for clustering streaming spatiotemporal data, the Spatial Dirichlet Hawkes Process (SDHP), which allows us to automatically uncover a wide variety of spatiotemporal patterns of collective social activity from geolocated online traces. Moreover, we develop an efficient, online inference algorithm based on Sequential Monte Carlo that scales to millions of geolocated posts. Experiments on synthetic data and real data gathered from Twitter show that our framework can recover a wide variety of meaningful social activity patterns in terms of both content and spatiotemporal dynamics, that it yields interesting insights about these patterns, and that it can be used to estimate the location from where a tweet was posted.

Martin Jankowiak
NYU CUSP
jankowiak@gmail.com

Manuel Gomez-Rodriguez
Max Planck Institute for Software Systems
manuelgr@mpi-sws.org

CP18

Discovery of Causal Time Intervals

Abstract not available at time of publication.

Zhenhui Li, Guanjie Zheng, Amal Agarwal, Lingzhou Xue
Pennsylvania State University
jessieli@ist.psu.edu, gjz5038@ist.psu.edu, aua257@psu.edu, lingzhou@psu.edu

Thomas Lauvaux
tul5@meteo.psu.edu
the pennsylvania state university

CP18

Robust Map Matching for Heterogeneous Data Via Dominance Decompositions

For a given sequence of location measurements, the goal of the geometric map matching problem is to compute a sequence of movements along edges of a spatially embedded graph which provides a 'good explanation' for the measurements. The problem gets challenging as real world data, like traces or graphs from the OpenStreetMap project, does not exhibit homogeneous data quality. Graph details and errors vary in areas and each trace has changing noise and precisions. Hence formalizing what a 'good explanation' is, becomes quite difficult. We propose a novel map matching approach which locally adapts to the data quality by constructing what we call *dominance decompositions*. While our approach is computationally more expensive than previous approaches, our experiments show that it allows for high quality map matching even in presence of highly variable data quality without parameter tuning.

Martin P. Seybold
Algorithms Group - FMI
University of Stuttgart
seybold@fmi.uni-stuttgart.de

CP18

Discovering Bursts Revisited: Guaranteed Optimization of the Model Parameters

One of the classic data mining tasks is to discover bursts, time intervals, where events occur at abnormally high rate. In this paper we revisit Kleinberg's seminal work, where bursts are discovered by using exponential distribution with a varying rate parameter: the regions where it is more advantageous to set the rate higher are deemed bursty. The model depends on two parameters, the initial rate and the change rate. The initial rate, that is, the rate that is used when there are no burstiness was set to the average rate over the whole sequence. The change rate is provided by the user. We argue that these choices are suboptimal: it leads to worse likelihood, and may lead to missing some existing bursts. We propose an alternative problem setting,

where the model parameters are selected by optimizing the likelihood of the model. While this tweak is trivial from the problem definition point of view, this changes the optimization problem greatly. To solve the problem in practice, we propose efficient $(1 + \epsilon)$ approximation schemes. Finally, we demonstrate empirically that with this setting we are able to discover bursts that would have otherwise be undetected.

Nikolaj Tatti
HIIT, Aalto University
nikolaj.tatti@aalto.fi