IP1

Regional Climate Informatics: A Statistical Perspective

As attention shifts from broad global summaries of climate change to more specific regional impacts there is a need for the data sciences to quantify the uncertainty in regional predictions. This talk will provide an overview on regional climate experiments with an emphasis on the data science problems for interpreting these large and complex simulations. Here a flexible spatial model based on fixed rank Kriging is implemented to handle a large number of spatial locations (LatticeKrig) and also include nonstationary spatial dependence.

Doug Nychka
National Center for Atmospheric Research
nychka@ucar.edu

IP2

Mining Clinical Data to Build Predictive Models

The IOMs envisioned "learning health care system," uses data from each patient to teach about prevention, diagnosis, prognosis, and treatment. Several trends will enable the realization of this goal: (1) universal electronic health records, (2) recording of clinically important details to support "meaningful use" for quality of care, (3) wearable sensors that provide real-time data on every individual, (4) the "$1000 genome" to study relations between clinical and genomic factors, and (5) "big data" techniques for analysis. With enough data, even simple machine learning techniques find strong predictive relationships. I describe our experiences predicting mortality and other important clinical risks and therapeutic opportunities for intensive care patients. Our current efforts abstract more informative features from both coded (tabular) data and narrative descriptions of the patient, biasing those unsupervised processes toward representing existing medical knowledge. I will outline what seem like further promising approaches. (Joint work with R. Joshi, A. Rumshisky, C. Hug, K. Kshetri, M. Ghassemi and T. Naumann.)

Peter Szolovits
Massachusetts Institute of Technology
psz@mit.edu

IP3

Modeling Individual-Level Data in the 21st Century

The collection and analysis of data related to human behavior has changed dramatically over the past 40 years, from the collection of small amounts of relatively static demographic data (such as a person's zipcode and education level), to much more detailed and dynamic transaction data (such as credit card and telephone records). More recently we have seen the rapid advent of individual-level "microdata," including Web search, email, microblogs, online social media, geolocation data, and more. In this talk we will discuss some of the new research challenges and opportunities presented by such data. We will look at common themes across the variety of data sets and research projects in this general area, focusing both on what new types of data analysis techniques are likely to be needed, and what new scientific questions and applications are emerging in areas such as computational social science and public health.

Padhraic Smyth
University of California
Irvine, California
smyth@ics.uci.edu

IP4

Social Networks as Information Filters

Social networks, especially online social networks, are driven by information sharing. But just how much information sharing is influenced by social networks? A large-scale experiment measured the effect of the social network on the quantity and diversity of information being shared within Facebook. While strong ties were found to be individually more influential, collectively it is the strong ties that wield more influence and provide more diverse information exposure. Furthermore, the network not only transmits information, but also often modifies it, allowing it to evolve.

Lada Adamic
University of Michigan, Ann Arbor
School of Information, Center for the Study of Complex Syste
ladamic@umich.edu

CP1

Mining Connection Pathways for Marked Nodes in Large Graphs

Abstract not available at time of publication.

Leman Akoglu
Carnegie Mellon University
leman@cs.stonybrook.edu

Jilles Vreeken
Universiteit Antwerpen
jilles.vreeken@ua.ac.be

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

Polo Chau
Georgia Tech
polo@gatech.edu

Nikolaj Tatti
K.U. Leuven
nikolaj.tatti@cs.kuleuven.be

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

CP1

NetSpot: Spotting Significant Anomalous Regions on Dynamic Networks

How to spot and summarize anomalies in dynamic networks such as road networks, communication networks and social networks? An anomalous event, such as a traffic accident or a denial of service attack, can affect several nearby edges and make them behave abnormally, over several consecutive time-ticks. We focus on spotting and summa-

rizing such significant anomalous regions, spanning space, as well as time. Our first contribution is the problem formulation, namely finding all such Significant Anomalous Regions. The next contribution is the design of novel algorithms: an expensive, exhaustive algorithm, as well as an efficient approximation. Compared to the exhaustive algorithm, our method is up to one order of magnitude faster in real data, while achieving less than 4% average relative error rate. In synthetic datasets, it is more than 30 times faster and solves large problem instances that are otherwise infeasible. The final contribution is the validation on real data.

Misael Mongiovi
University of Catania
Dipartimento di Matematica e Informatica
mongiovi@dmi.unict.it

Petko Bogdanov, Razvan Ranca
UCSB
petko@cs.ucsb.edu, ranca.razvan@gmail.com

Evangelos Papalexakis
CMU
epapalex@cs.cmu.edu

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

Ambuj Singh
UCSB
ambuj@cs.ucsb.edu

CP1

Maximal Deviations of Incomplete U-Statistics with Applications to Empirical Risk Sampling

We show how to extend the ERM paradigm, from a practical perspective, to the situation where a natural estimate of the risk is of the form of a K-sample U-statistics, as it is the case in the K-partite ranking problem for instance. Indeed, the numerical computation of the empirical risk is hardly feasible if not infeasible, even for moderate samples sizes. It involves averaging $O(n^{d_1+\ldots+d_K})$ terms, when considering a U-statistic of degrees $(d_1, \ldots, d_K)$ based on samples of sizes proportional to n. We propose here to consider a drastically simpler Monte-Carlo version of the empirical risk based on $O(n)$ terms solely, which can be viewed as an incomplete generalized U-statistic, and prove that, remarkably, the approximation stage does not damage the ERM procedure and yields a learning rate of order $O_{\mathbf{P}}(1/\sqrt{n})$.

Stephan Clémençon, Sylvain Robbiano
Telecom ParisTech
stephan.clemencon@telecom-paristech.fr,
sylvain.robbiano@telecom-paristech.fr

Jessica Tressou
INRA
jessica.tressou@agroparistech.fr

CP1

Fast Exact Max-Kernel Search

The wide applicability of kernels makes the problem of max-kernel search ubiquitous and more general than the usual similarity search in metric spaces. We focus on solving this problem efficiently. We begin by characterizing the inherent hardness of the max-kernel search problem with a novel notion of *directional concentration*. Following that, we present a method to use an $O(n \log n)$ algorithm to index any set of objects (points in $\mathbf{R^d}$ or abstract objects) *directly in the Hilbert space* without any explicit feature representations of the objects in this space. We present the first provably $O(\log n)$ algorithm for exact max-kernel search using this index. Empirical results for a variety of data sets as well as abstract objects demonstrate up to 4 orders of magnitude speedup in some cases. Extensions for approximate max-kernel search are also presented.

Parikshit Ram
School of Computational Science and Engineering
Georgia Institute of Technology
p.ram@gatech.edu

Ryan Curtin, Alexander Gray
Georgia Institute of Technology
curtinr@gatech.edu, agray@cc.gatech.edu

CP1

Triadic Measures on Graphs: The Power of Wedge Sampling

Counting triangles is a fundamental operation on graphs, but can be computationally expensive for massive graphs. We discuss the method of *wedge sampling*. The versatile technique allows for fast estimation of various clustering coefficients, and has provable approximation guarantees. We perform extensive experimental tests to demonstrate the behavior of this method in practice. Our algorithms are orders of magnitude faster than state-of-the-art while providing nearly the accuracy of full enumeration.

C. Seshadhri, Ali Pinar
Sandia National Labs
scomand@sandia.gov, apinar@sandia.gov

Tamara G. Kolda
Sandia National Laboratories
tgkolda@sandia.gov

CP2

Discriminative Feature Selection for Uncertain Graph Classification

Abstract not available at time of publication.

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

Xue Wang, Ann Ragin
Northwestern University
xue-wang@northwestern.edu, ann-ragin@northwestern.edu

CP2

Mining Probabilistic Representative Frequent Patterns From Uncertain Data

Probabilistic frequent pattern (PFP) mining over uncertain data has received much attention recently. Similar to its counterpart in deterministic databases, however, PFP mining suffers from the same problem of generating an ex-

ponential number of result patterns, which hinders further evaluation and analysis. This paper formally defines the *probabilistic representative frequent pattern (P-RFP) mining* problem, which aims to find the minimal set of patterns with sufficiently high probability to represent all other patterns. The bottleneck turns out to be checking whether a pattern can probabilistically represent another. To address the problem, we propose a novel and efficient dynamic programming-based approach, and devised a set of effective optimization strategies to further improve the computation efficiency. Our proposed approach not only discovers the set of P-RFPs efficiently, but also restores the frequency probability information of patterns with an error guarantee.

Chunyang Liu, Ling Chen, Chengqi Zhang
University of Technology, Sydney
Chunyang.Liu@student.uts.edu.au, ling.chen@uts.edu.au, chengqi.zhang@uts.edu.au

CP2

Missing Or Inapplicable: Treatment of Incomplete Continuous-Valued Features in Supervised Learning

Real-world data are often riddled with data quality problems such as noise, outliers and missing values, which present significant challenges for supervised learning algorithms to effectively classify them. This paper explores the ill-effects of inapplicable features on the performance of supervised learning algorithms. In particular, we highlight the difference between missing and inapplicable feature values. We argue that the current approaches for dealing with missing values, which are mostly based on single or multiple imputation methods, are insufficient to handle inapplicable features, especially those that are continuous valued. We also illustrate how current tree-based and kernel-based classifiers can be adversely affected by the presence of such features if not handled appropriately. Finally, we propose methods to extend existing tree-based and kernel-based classifiers to deal with the inapplicable continuous-valued features.

Pang-Ning Tan, Prakash Mandayam Comar, Lei Liu
Michigan State University
ptan@cse.msu.edu, mandayam@cse.msu.edu, liulei1@cse.msu.edu

Sabyasachi Saha, Antonio Nucci
Narus Inc
ssaha@narus.com, anucci@narus.com

CP2

Collective Kernel Construction in Noisy Environment

Kernels are similarity functions, and play important roles in machine learning. Traditional kernels are built directly from the feature vectors of data instances x_i, x_j. However, data could be noisy, and there are missing values or corrupted values in feature vectors. In this paper, we propose a new approach to build kernel - Collective Kernel, especially from noisy data. We also derive an efficient algorithm to solve the L_1-norm based optimization. Extensive experiments on face data, hand written characters and image scene datasets show improved performance for clustering and semi-supervised classification tasks on our collective kernel comparing with the traditional gaussian kernel.

Miao Zhang
University of Texas, Arlington
miao.zhang@mavs.uta.edu

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

Deguang Kong
University of Texas, Arlington
doogkong@gmail.com

CP2

Patient Risk Prediction Model via Top-k Stability Selection

In this paper, we propose top-k stability selection, which generalizes a powerful sparse learning method for feature selection by overcoming its limitation on parameter selection. In particular, our proposed top-k stability selection includes the original stability selection method as a special case given $k = 1$. Moreover, we show that the top-k stability selection is more robust by utilizing more information from selection probabilities than the original stability selection, and provides stronger theoretical properties. In a large set of real clinical prediction datasets, the top-k stability selection methods outperform many existing feature selection methods including the original stability selection. Through several clinical applications on predicting heart failure related symptoms, we show that top-k stability selection can successfully identify important features that are clinically meaningful.

Jiayu Zhou
Arizona State University
Jiayu.Zhou@asu.edu

Jimeng Sun
IBM T.J. Watson Research Center
jimeng@us.ibm.com

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

Jianying Hu
IBM
jyhu@us.ibm.com

Jieping Ye
Arizona State University
jieping.ye@asu.edu

CP3

Constrained Spectral Clustering Using L1 Regularization

Constrained spectral clustering is a semi-supervised learning problem that aims at incorporating user-defined constraints in spectral clustering. Typically, there are two kinds of constraints: (i) *must-link*, and (ii) *cannot-link*. These constraints represent prior knowledge indicating whether two data objects should be in the same cluster or not; thereby aiding in clustering. In this paper, we propose a novel approach that uses convex subproblems to incor-

porate constraints in spectral clustering and co-clustering. In comparison to the prior state-of-art approaches, our approach presents a more natural way to incorporate constraints in the spectral methods and allows us to make a trade off between the number of satisfied constraints and the quality of partitions on the original graph. We use an L_1 regularizer analogous to LASSO, often used in literature to induce sparsity, in order to control the number of constraints satisfied.

Jaya Kawale
University of Minnesota
kawale@cs.umn.edu

Daniel L. Boley
University of Minnesota
Department of Computer Science
boley@cs.umn.edu

CP3

Efficient Anytime Density-Based Clustering

Many clustering algorithms suffer from scalability problems on massive datasets and do not support any user interaction during runtime. To tackle these problems, anytime clustering algorithms are proposed. They produce a fast approximate result which is continuously refined during the further run. Also, they can be stopped or suspended anytime and provide an answer. In this paper, we propose a novel anytime clustering algorithm based on the density-based clustering paradigm. Our algorithm called A-DBSCAN is applicable to very high dimensional databases such as time series, trajectory, medical data, etc. The general idea of our algorithm is to use a sequence of lower-bounding functions (LBs) of the true similarity measure to produce multiple approximate results of the true density-based clusters. A-DBSCAN operates in multiple levels w.r.t. the LBs and is mainly based on two algorithmic schemes: (1) an efficient distance upgrade scheme which restricts distance calculations to core-objects at each level of the LBs; (2) a local re-clustering scheme which restricts update operations to the relevant objects only. Extensive experiments demonstrate that A-DBSCAN acquires very good clustering results at very early stages of execution thus saves a large amount of computational time. Even if it runs to the end, A-DBSCAN is still orders of magnitude faster than DBSCAN.

Son T. Mai
Institute for Computer Science
University of Munich
mtson@dbs.ifi.lmu.de

Xiao He, Jing Feng, Christian Boehm
University of Munich
he@dbs.ifi.lmu.de, feng@dbs.ifi.lmu.de,
boehm@dbs.ifi.lmu.de

CP3

Determining the Number of Clusters Via Iterative Consensus Clustering

We use a cluster ensemble to determine the number of clusters, k, in a group of data. A consensus similarity matrix is formed from the ensemble using multiple algorithms and several values for k. A random walk is induced on the graph defined by the consensus matrix and the eigenvalues of the associated transition probability matrix are used to determine the number of clusters. For noisy or high-dimensional data, an iterative technique is presented to refine this consensus matrix in way that encourages a block-diagonal form. Results are given for a variety of datasets, with a particular emphasis on text data.

Shaina L. Race
North Carolina State University
slrace@ncsu.edu

CP3

Sparse Subspace Clustering Via Group Sparse Coding

We propose in this paper a novel sparse subspace clustering method that regularizes sparse subspace representation by exploiting the structural sharing between tasks and data points via group sparse coding. We derive simple, provably convergent, and computationally efficient algorithms for solving the proposed group formulations. We demonstrate the advantage of the framework on three challenging benchmark datasets ranging from medical record data to image and text clustering and show that they consistently outperforms rival methods.

Budhaditya Saha
Deakin University
Victoria , Australia
budhaditya.saha@deakin.edu.au

Duc Son Pham
Curtin University
Perth, Western Australia
ducson.pham@curtin.edu.au

Dinh Phung, Svetha Venkatesh
Deakin University, Geelong Waurn Ponds Campus
Victoria, Australia
dinh.phung@deakin.edu.au,
svetha.venkatesh@deakin.edu.au

CP3

Evolutionary Soft Co-Clustering

We consider the mining of block structures from time-varying data using evolutionary co-clustering. Existing methods are based on the spectral framework, thus lacking a probabilistic interpretation. To overcome this limitation, we develop a probabilistic model for evolutionary co-clustering in this paper. The proposed model assumes that the data are generated via a two-step process that depends on the historic co-clusters, thereby capturing the temporal smoothness. We develop an EM algorithm to perform maximum likelihood parameter estimation. An appealing feature of the proposed probabilistic model is that it leads to soft co-clustering assignments naturally. To the best of our knowledge, our work represents the first attempt to perform evolutionary soft co-clustering. We evaluate the proposed method on both synthetic and real data sets. Experimental results show that our method consistently outperforms prior approaches based on spectral method.

Shuiwang Ji
Arizona State University
sji@cs.odu.edu

Wenlu Zhang
Old Dominion University
wzhang@cs.odu.edu

Rui Zhang
City College of New York
ruizhang@ccny.cuny.edu

CP4

What's Your Next Move: User Activity Prediction in Location-Based Social Networks

Location-based social networks have been gaining increasing popularity in recent years. To increase users' engagement with location-based services, it is important to provide attractive features, one of which is geo-targeted ads and coupons. To make ads and coupon delivery more effective, it is essential to predict the location that is most likely to be visited by a user at the next step. In this paper we exploit the check-in category information to model the underlying user movement pattern. We propose a framework which uses a mixed hidden Markov model to predict the category of user activity at the next step and then predict the most likely location given the estimated category distribution. Extensive experimental results show that, with the predicted category distribution, the number of location candidates for prediction is 5.45 times smaller, while the prediction accuracy is 13.21% higher.

Jihang Ye, Zhe Zhu, Hong Cheng
The Chinese University of Hong Kong
yjh010@alumni.ie.cuhk.edu.hk,
zzhu@alumni.ie.cuhk.edu.hk, hcheng@se.cuhk.edu.hk

CP4

DeltaCon: A Principled Massive-Graph Similarity Function

How much did a network change since yesterday? Graph similarity with known node correspondence arises in numerous settings. We formally state the axioms and desired properties of the graph similarity functions, and propose DeltaCon, a principled, intuitive, and scalable algorithm that assesses the similarity between two graphs on the same nodes. Finally, we evaluate when state-of-the-art methods fail to detect crucial connectivity changes in graphs, and apply our method for classification and anomaly detection.

Danai Koutra
Carnegie Mellon University
danai@cs.cmu.edu

Joshua Vogelstein
Duke University
jovo@stat.duke.edu

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

CP4

CoFiSet: Collaborative Filtering Via Learning Pairwise Preferences over Item-Sets

One fundamental challenge of collaborative filtering with implicit feedbacks is the lack of negative feedbacks, because there are only some observed relatively "positive' feedbacks, making it difficult to learn a prediction model. In this paper, we propose a new and relaxed assumption of *pairwise preferences over item-sets*, which defines a user's preference on a set of items (item-set) instead of on a single item. The relaxed assumption can give us more accurate pairwise preference relationships. With this assumption, we further develop a general algorithm called CoFiSet (**co**llaborative **fi**ltering via learning pairwise preferences over item-**set**s). Experimental results show that CoFiSet performs better than several state-of-the-art methods on various ranking-oriented evaluation metrics on two real-world data sets. Furthermore, CoFiSet is very efficient as shown by both the time complexity and CPU time.

Weike Pan, Li Chen
Hong Kong Baptist University
wkpan@comp.hkbu.edu.hk, lichen@comp.hkbu.edu.hk

CP4

Dynamic Community Detection in Weighted Graph Streams

In this paper, we aim to tackle the problem of discovering dynamic communities in weighted graph streams, especially when the underlying social behavior of individuals varies considerably over different graph regions. To tackle this problem, a novel structure termed *Local Weighted-Edge-based Pattern* (LWEP) *Summary* is proposed to describe a local homogeneous region. To efficiently compute LWEPs, some statistics need to be maintained according to the principle of preserving maximum weighted neighbor information with limited memory storage. To this end, the proposed approach is divided into online and offline components. During the online phase, we introduce some statistics, termed top-k neighbor lists and top-k candidate lists, to track. The key is to maintain only the top-k neighbors with the largest link weights for each node. To allow for less active neighbors to transition into top-k neighbors, an auxiliary data structure termed top-k candidate list is used to identify emerging active neighbors. The statistics can be efficiently maintained in the online component. In the offline component, these statistics are used at each snapshot to efficiently compute LWEPs. Clustering is then performed to consolidate LWEPs into high level clusters. Finally, mapping is made between clusters of consecutive snapshots to generate temporally smooth communities. Experimental results are presented to illustrate the effectiveness and efficiency of the proposed approach.

Chang-Dong Wang, Jian-Huang Lai
Sun Yat-sen University
changdongwang@hotmail.com, stsljh@mail.sysu.edu.cn

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

CP4

On Graph Stream Clustering with Side Information

Recently, many applications generate data in the form of streams. Meanwhile, a large volume of side information is associated with graphs. In this paper, we define a unified distance measure on both link structures and side attributes for clustering. In addition, we propose a novel optimization framework DMO, which dynamically optimize the distance and adapt to the stream. We further introduce $SGS(C)$ which consume constant storage with the progression of streams.

Yuchen Zhao, Philip Yu
University of Illinois at Chicago
yzhao@cs.uic.edu, psyu@uic.edu

CP5

Outlier Detection with Space Transformation and Spectral Analysis

We present an approach that exploits space transformation and uses spectral analysis in the newly transformed space for outlier detection. Unlike most existing techniques, this approach introduces a novel concept based on local quadratic entropy for evaluating the similarity of a data object with its neighbors. This information theoretic quantity is used to regularize the closeness amongst data instances and subsequently benefits the process of mapping data into a usually lower dimensional space. Outliers are then identified by spectral analysis of the eigenspace spanned by the set of leading eigenvectors derived from the mapping procedure. The proposed technique is purely data-driven, making it particularly suitable for identification of outliers from irregular, non-convex shaped distributions and from data with diverse, varying densities.

Xuan-Hong Dang
Department of Computer Science
Aarhus University, Denmark
dang@cs.au.dk

Barbora Micenkova, Ira Assent
Aarhus University, Denmark
Department of Computer Science
barbora@cs.au.dk, ira@cs.au.dk

Raymond T. Ng
University of British Columbia, Canada
Department of Computer Science
rng@cs.ubc.ca

CP5

k-Means--: A Unified Approach to Clustering and Outlier Detection

We present a unified approach for simultaneously clustering and discovering outliers in data. Our approach is formalized as a generalization of the k-means problem. We prove that the problem is NP-hard and then present a practical polynomial time algorithm, which is guaranteed to converge to a local optimum. Furthermore we extend our approach to all distance measures that can be expressed in the form of a Bregman divergence. Experiments on synthetic and real datasets demonstrate the effectiveness of our approach and the utility of carrying out both clustering and outlier detection in a concurrent manner. In particular on the famous KDD cup network-intrusion dataset, we were able to increase the precision of the outlier detection task by nearly 100% compared to the classical nearest-neighbor approach.

Aristides Gionis
Yahoo! Research
Barcelona, Spain
gionis@yahoo-inc.com

Aristides Gionis
University of Sydney
sanjay chawla

CP5

Cost-Sensitive Double Updating Online Learning and Its Application to Online Anomaly Detection

Although both *cost-sensitive classification* and *online learning* have been well studied separately in data mining and machine learning, there was very few comprehensive study of cost-sensitive online classification in literature. In this paper, we formally investigate this problem by directly optimizing cost-sensitive measures for an online classification task. As the first comprehensive study, we propose the Cost-Sensitive Double Updating Online Learning (CSDUOL) algorithms, which explores a recent double updating technique to tackle the online optimization task of cost-sensitive classification by maximizing the weighted sum or minimizing the weighted misclassification cost. We theoretically analyze the cost-sensitive measure bounds of the proposed algorithms, extensively examine their empirical performance for cost-sensitive online classification tasks, and finally demonstrate the application of our technique to solve online anomaly detection tasks.

Peilin Zhao, Steven C.H. Hoi
Nanyang Technological University
peilinzhao@ntu.edu.sg, chhoi@ntu.edu.sg

CP5

CMI: An Information-Theoretic Contrast Measure for Enhancing Subspace Cluster and Outlier Detection

In multi-dimensional data, the knowledge is likely hidden in subspaces. It is an open research issue to select meaningful subspaces without any prior knowledge about such hidden patterns. We focus on finding subspaces with strong mutual dependency in the selected dimensions. To do this, we propose a novel contrast score that quantifies mutual correlations in subspaces by considering their cumulative distributions. Chosen subspaces provide a high discrepancy between clusters and outliers and enhance their detection.

Hoang Vu Nguyen, Emmanuel Müller
Karlsruhe Institute of Technology
hoang.nguyen@kit.edu, emmanuel.mueller@kit.edu

Jilles Vreeken
Universiteit Antwerpen
jilles.vreeken@ua.ac.be

Fabian Keller, Klemens Böhm
Karlsruhe Institute of Technology
fabian.keller@kit.edu, klemens.boehm@kit.edu

CP5

Efficient Selection of Globally Optimal Rules on Large Imbalanced Data Based on Rule Coverage Relationship Analysis

Rule-based anomaly and fraud detection systems often suffer from massive false alerts against a huge number of enterprise transactions. In this paper, we analyze the interactions and relationships between rules and their coverage on transactions, and propose a novel metric, Max Coverage Gain. An effective algorithm, MCGminer, is then designed with a series of built-in mechanisms and pruning strategies to handle complex rule interactions and reduce computational complexity towards identifying the globally optimal rule set. Substantial experiments on 13 UCI data sets and a real time online banking transactional database demonstrate that MCGminer achieves significant improvement on both accuracy, scalability, stability and efficiency on large imbalanced data compared to several state-of-the-art rule

selection techniques.

Jinjiu Li
University of Technology Sydney
joe20150101@gmail.com

Can Wang, Longbing Cao
University of Technolgoy Sydney
canwang613@gmail.com, longbing.cao@gmail.com

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

CP6

Multi-Objective Multi-View Spectral Clustering Via Pareto Optimization

Traditionally, spectral clustering is limited to a single objective: finding the normalized min-cut of a single graph. However, many real-world datasets are generated from multiple heterogeneous sources. How to optimally combine knowledge from multiple sources to improve spectral clustering remains a developing area. Previous work on multi-view clustering formulated the problem as a single objective function to optimize, typically by combining the views under a compatibility assumption and requiring the users to decide the importance of each view *a priori*. In this work, we propose a multi-objective formulation and show how to solve it using Pareto optimization. The Pareto frontier captures all possible good cuts without requiring the users to set the "correct' parameter. The effectiveness of our approach is justified by both theoretical analysis and empirical results. We also demonstrate a novel application of our approach: resting-state fMRI analysis.

Xiang Wang, Buyue Qian
UC Davis
xiang@ucdavis.edu, byqian@ucdavis.edu

Ian Davidson
University of California, Davis
davidson@cs.ucdavis.edu

Jieping Ye
Arizona State University
jieping.ye@asu.edu

CP6

On Handling Negative Transfer and Imbalanced Distributions in Multiple Source Transfer Learning

In this paper, we propose a novel two-phase framework to effectively transfer knowledge from multiple sources even when there exist irrelevant sources and imbalanced class distributions. First, an effective Supervised Local Weight (SLW) scheme is proposed to assign a proper weight to each source domain. The second phase then learns a classifier for the target domain by solving an optimization problem. Extensive experiments demonstrate the significant improvement in classification performance over existing baseline approaches.

Liang Ge
The Sate University of New York at Buffalo
liangge@buffalo.edu

Jing Gao
University at Buffalo
jing@buffalo.edu

Hung Ngo, Kang Li
The State University of New York at Buffalo
hungngo@buffalo.edu, kli22@buffalo.edu

Aidong Zhang
Department of Computer Science
State University of New York at Buffalo
azhang@buffalo.edu

CP6

Multi-View Clustering Via Joint Nonnegative Matrix Factorization

To integrate information from multiple views in the unsupervised setting, multi-view clustering algorithms have been developed to cluster multiple views simultaneously to derive a solution which uncovers the common latent structure shared by multiple views. In this paper, we propose a novel NMF-based multi-view clustering algorithm by searching for a factorization that gives compatible clustering solutions across multiple views. The key idea is to formulate a joint matrix factorization process with the constraint that pushes clustering solution of each view towards a common consensus instead of fixing it directly. The main challenge is how to keep clustering solutions across different views meaningful and comparable. To tackle this challenge, we design a novel and effective normalization strategy inspired by the connection between NMF and PLSA. Experimental results on synthetic and several real datasets demonstrate the effectiveness of our approach.

Jialu Liu, Chi Wang
University of Illinois at Urbana-Champaign
jliu64@illinois.edu, chiwang1@illinois.edu

Jing Gao
University at Buffalo
jing@buffalo.edu

Jiawei Han
UIUC
hanj@illinois.edu

CP6

Multi-Transfer: Transfer Learning with Multiple Views and Multiple Sources

In many real-world applications, auxiliary data are described from multiple views and carried by multiple sources. For example, to help classify videos on Youtube, which include three views: image, voice and subtitles, one may borrow auxiliary data from Flickr, Last.FM and Google News. Although any single instance in these domains can only cover a part of the views on Youtube, actually the piece of information carried by them may compensate with each other. In this paper, we define this problem as Transfer Learning with Multiple Views and Multiple Sources. As different sources may have different probability distributions, merging all data in a simplistic manner will not give optimal result. Thus, we propose a novel algorithm to leverage knowledge from different views and sources collaboratively, by letting different views from different sources complement each other through a co-training style framework, while revise the distribution differences in

different domains.

Ben Tan
Hong Kong University of Science and Technology,
Clear Water Bay, Hong Kong
btan@cse.ust.hk

Erheng Zhong
Hong Kong University of Science and Technology
Department of Computer Science and Engineering
ezhong@cse.ust.hk

Evan Wei Xiang
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
wxiang@cse.ust.hk

Qiang Yang
Hong Kong University of Science and Technology
Department of Computer Science and Engineering
qyang@cse.ust.hk

CP6

Unsupervised Feature Selection for Multi-View Data in Social Media

The explosive popularity of social media produces mountains of high-dimensional data and the nature of social media also determines that its data is often unlabelled, noisy and partial, presenting challenges to feature selection. Social media data can be represented by heterogeneous feature spaces in the form of multiple views. In general, multiple views can be complementary and, when used together, can help handle noisy and partial data for any single-view feature selection. These unique challenges and properties motivate us to develop a novel feature selection framework to handle multi-view social media data. In this paper, we investigate how to exploit relations among views to help each other select relevant features, and propose a novel unsupervised feature selection framework MVFS for multi-view social media data. We systematically evaluate the proposed framework in multi-view datasets from social media websites and the results demonstrate the effectiveness and potential of MVFS.

Jiliang Tang
Arizona State University
ARIZONA STATE UNIVERISTY
Jiliang.Tang@asu.edu

Xia Hu, Huiji Gao, Huan Liu
Arizona State University
xia.hu@asu.edu, huiji.gao@asu.edu, huan.liu@asu.edu

CP7

Probabilistic Combination of Classifier and Cluster Ensembles for Non-Transductive Learning

Unsupervised models can provide supplementary soft constraints to help classify new target data. Such models can also help detect possible divergences between training and target distributions, which is useful in applications where concept drift may take place. This paper describes a Bayesian framework that takes as input class labels from existing classifiers (designed based on labeled data from the source domain), as well as cluster labels from a cluster ensemble operating solely on the target data to be classified, and yields a consensus labeling of the target data. This framework is particularly useful when the statistics of the target data drift or change from those of the training data. We also show that the proposed framework is privacy-aware and allows performing distributed learning when data/models have sharing restrictions. Experiments show that our framework can yield superior results to those provided by applying classifier ensembles only.

Ayan Acharya
University of Texas at Austin
Department of ECE
aacharya@utexas.edu

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

Joydeep Ghosh
University of Texas at Austin
ghosh@ece.utexas.edu

Badrul Sarwar, Jean-David Ruvini
eBay Research Lab, San Jose.
eBay Inc.
bsarwar@ebay.com, jruvini@ebay.com

CP7

Active Class Discovery and Learning for Networked Data

With the recent explosion of social network applications, active learning has increasingly become an important paradigm for classifying networked data. For most social network applications, the dynamic change of users and their evolving relationships, along with the emergence of new social events, often result in new classes that need to be immediately discovered and labeled for classification. This paper proposes a novel approach called ADLNET for active class discovery and learning with networked data. Our proposed method uses the Dirichlet process defined over class distributions to enable active discovery of new classes, and explicitly models label correlations in the utility function of active learning. Experimental results on two real-world networked data sets demonstrate that our proposed approach outperforms other state-of-the-art methods.

Meng Fang
University of Technology Sydney
Meng.Fang@student.uts.edu.au

Jie Yin
CSIRO
jie.yin@csiro.au

Xingquan Zhu
University of Technology Sydney
xingquan.zhu@uts.edu.au

Chengqi Zhang
University of Technology, Sydney
chengqi.zhang@uts.edu.au

CP7

ActNeT: Active Learning for Networked Texts in

Microblogging

In order to reduce the labeling cost in supervised learning, active learning is an effective way to select representative and informative instances to query for labels for improving the learned model. Inspired by social correlation theories, we investigate whether social relations can help perform effective active learning on networked data. In this paper, we propose a novel Active learning framework for the classification of Networked Texts in microblogging (**ActNeT**). In particular, we study how to incorporate network information into text content modeling, and design strategies to select the most representative and informative instances from microblogging for labeling by taking advantage of social network structure. Experimental results show that the proposed framework significantly outperforms existing state-of-the-art methods.

Xia Hu
Arizona State University
xia.hu@asu.edu

Jiliang Tang
Arizona State University
ARIZONA STATE UNIVERISTY
Jiliang.Tang@asu.edu

Huiji Gao, Huan Liu
Arizona State University
huiji.gao@asu.edu, huan.liu@asu.edu

CP7

Active Learning to Rank Using Pairwise Supervision

We investigate learning a ranking function using pairwise constraints in the context of human-machine interaction. Our active learning to rank is performed by querying domain experts of pairwise orderings, which are selected by considering both *local* and *global* uncertainty. We evaluate our approach on three real data sets and compare with representative methods. The promising experimental results demonstrate the effectiveness of actively using pairwise orderings to improve ranking performance.

Buyue Qian
UC Davis
byqian@ucdavis.edu

Hongfei Li
IBM Research
liho@us.ibm.com

Jun Wang
IBM Thomas J. Watson Research Center
Business Analytics and Mathematical Sciences Department
wangjun@us.ibm.com

Xiang Wang, Ian Davidson
UC Davis
xiang@ucdavis.edu, indavidson@ucdavis.edu

CP7

Smart: Semi-Supervised Music Emotion Recognition with Social Tagging

Music emotion recognition (MER) aims to recognize the affective content of music, which is important for music recommendation, etc. MER is commonly formulated as a supervised learning problem. In practice, there is little labeled data in most genres except for Pop music, and emotion is genre specific in music. Thus, labeled data of Pop music cannot be used for other genres. In this paper, we aim to solve the genre-specific MER problem by effectively exploiting *unlabeled songs* and *social tags*, which is a non-trivial task, e.g., tags are too noisy to be treated as fully trustworthy. To build an accurate model, we present **SMART**: Semi-Supervised Music Affective Emotion Recognition with Social Tagging, combining a graph-based semi-supervised learning algorithm with a novel tag refinement method. Experiments on the Million Song Dataset show that our approach, trained with only 10 labeled songs, is as accurate as Support Vector Regression trained with 750 labeled songs.

Bin Wu, Erheng Zhong, Derek Hao Hu, Andrew Horner, Qiang Yang
Hong Kong University of Science and Technology
Department of Computer Science and Engineering
bwuaa@cse.ust.hk, ezhong@cse.ust.hk, derekhh@cse.ust.hk, horner@cse.ust.hk, qyang@cse.ust.hk

CP8

A Distribution Regularized Regression Framework for Climate Modeling

Regression-based approaches are widely used in climate modeling to capture the relationship between a climate variable of interest and a set of predictor variables. However, some climate modeling applications emphasize fitting the distribution properties of the observed data. In this paper, we show the limitations of current regression-based approaches in terms of preserving the distribution of observed climate data and present a multi-objective regression framework that simultaneously fits the distribution properties and minimizes the prediction error. The framework is highly flexible and can be applied to linear, nonlinear, and conditional quantile models.

Zubin Abraham, Pang-Ning Tan, Perdinan Perdinan, Julie Winkler, Shiyuan Zhong
Michigan State University
abraha84@msu.edu, ptan@cse.msu.edu, perdinan@msu.edu, winkler@msu.edu, zhongs@msu.edu

Malgozata Liszewska
University of Warsaw
m.liszewska@icm.edu.pl

CP8

Dynamic Shaker Detection from Evolving Entities

Abstract not available at time of publication.

Xiaoxiao Shi
Computer Department, University of Illinois at Chicago
xiao.x.shi@gmail.com

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

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

CP8

Monitoring and Mining Gps Traces in Transit Space

Users of mass transit systems such as those of buses and trains normally rely on accurate route maps, stop locations, and service schedules when traveling. If the route map, service schedule, or stop location has errors it can reduce the transit agencys ridership. In this paper, the problem of deriving transit systems by mining raw GPS data is studied. Specifically, we propose and evaluate novel classification features with spatial and temporal clustering techniques that derive bus stop locations, route geometries, and service schedules from GPS data. Subsequently, manual and expensive field visits to record and annotate the initial or updated route geometries, transit stop locations, or service schedules is no longer required by transit agencies. This facilitates a massive reduction in cost for transit agencies. The effectiveness of the proposed algorithms is validated on the third largest public transit system in the United States.

Leon O. Stenneth
University of Illinois Chicago
lstenneth@gmail.com

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

CP8

Climate Multi-Model Regression Using Spatial Smoothing

In this paper, we address the problem of combining multiple Global Climate Model (GCM) outputs with spatial smoothing as a desired criterion. The problem formulation takes the form of multiple least squares regression for each geographic location with graph Laplacian based smoothing amongst the neighboring locations. We discuss a few approaches to solve the problem, and establish the superiority of our approach in terms of model accuracy and smoothing.

Karthik Subbian, Arindam Banerjee
University of Minnesota
karthik@umn.edu, banerjee@cs.umn.edu

CP8

Sparse Representation for Hiv-1 Protease Drug Resistance Prediction

Abstract not available at time of publication.

Xiaxia Yu, Irene Weber, Robert Harrison
Georgia State University
xyu3@student.gsu.edu, iweber@gsu.edu, rharrison@cs.gsu.edu

CP9

Opinion Maximization in Social Networks

Abstract not available at time of publication.

Aristides Gionis
Yahoo! Research
Barcelona, Spain
gionis@yahoo-inc.com

Evamaria Terzi
Boston University
evimaria@cs.bu.edu

Panayiotis Tsaparas
University of Ioannina
tsap@cs.uoi.gr

CP9

Point-of-Interest Recommendation in Location Based Social Networks with Topic and Location Awareness

Unlike traditional recommendation tasks, Point-of-Interest recommendation is personalized, location-aware, and context depended. In light of this difference, this paper proposes a *topic and location* aware POI recommender system by exploiting associated textual and context information. Specifically, we first exploit an aggregated latent Dirichlet allocation model to learn the interest topics of users and to infer the interest POIs by mining textual information associated with POIs. Then, a *Topic and Location-aware* probabilistic matrix factorization (TL-PMF) method is proposed for POI recommendation. A unique perspective of TL-PMF is to consider both the extent to which a user interest matches the POI in terms of topic distribution and the word-of-mouth opinions of the POIs. Finally, experiments on real-world LBSNs data show that the proposed recommendation method outperforms state-of-the-art probabilistic latent factor models with a significant margin.

Bin Liu
Rutgers University
benbin.liu@rutgers.edu

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

CP9

Community Detection with Prior Knowledge

The problem of community detection is challenging due to the presence of hubs and noisy links, which tend to create highly imbalanced graph clusters. With the growing availability of network information, there is significant amount of prior knowledge available about the communities. We explore the use of such prior knowledge for finding balanced and high quality communities. We propose and evaluate an adaptive density-based clustering with prior knowledge to produce both overlapping and non-overlapping clusters.

Karthik Subbian
University of Minnesota
karthik@umn.edu

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

Jaideep Srivastava
University of Minnesota
srivasta@cs.umn.edu

Philip Yu
University of Illinois Chicago
psyu@cs.uiuc.edu

CP9

Exploring and Inferring User-User Pseudo-Friendship for Sentiment Analysis with Heterogeneous Networks

In this paper, we propose a novel information network-based framework which can infer hidden similarity and dissimilarity between users by exploring similar and opposite opinions, so as to improve post- level and user-level sentiment classification in the same time. More specifically, we develop a new meta path-based measure for inferring pseudo-friendship as well as dissimilarity between users, and propose a semi-supervised refining model by encoding similarity and dissimilarity from both user-level and post-level relations. We extensively evaluate the proposed approach and compare with several state-of-the-art techniques on two real-world forum datasets. Experimental results show that our proposed model with 10.5% labeled samples can achieve better performance than a traditional supervised model trained on 61.7% data samples.

Hongbo Deng, Jiawei Han
University of Illinois at Urbana-Champaign
hbdeng@illinois.edu, hanj@cs.uiuc.edu

Hao Li, Heng Ji
City University of New York
haoli.qc@gmail.com, hengjicuny@gmail.com

Hongning Wang
University of Illinois at Urbana-Champaign
wang296@illinois.edu

Yue Lu
Twitter Inc.
yuelu@twitter.com

Chi Wang
University of Illinois at Urbana-Champaign
chiwang1@illinois.edu

CP9

Exploiting Synchronicity Networks for Finding Valuables in Heterogeneous Networks

Successful enterprises depend on high performing teams consisting of productive individuals, who can effective find valuable information. Predictive methods are highly desired to identify these inter-related, multi-typed entities that can potentially help to improve enterprise performance based on observation of their dynamic behavior in the organizational social networks In this paper, we propose a novel approach to analyze and rank heterogeneous objects in multi-level networks by their value for improving productivity. Compared to existing approaches, our work offers two unique contributions. First, we propose a novel multi-level synchronicity network representation which allows us to exploit the structural characteristics of various entities' dynamic behavior. Furthermore, based on the synchronicity networks, we propose a novel HMR algorithm, to simultaneously rank inter-related heterogeneous entities (e.g., topics, individuals and teams) by their value.

Zhen Wen
IBM T. J. Watson Research Center
zhenwen@us.ibm.com

Ching-Yung Lin
IBM Reserach
chingyung@us.ibm.com

CP10

Regularization of Latent Variable Models to Obtain Sparsity

We present a *pseudo-observed* variable based regularization technique for latent variable mixed-membership models that provides a mechanism to impose preferences on the characteristics of aggregate functions of latent and observed variables. The regularization framework is used to regularize topic models where documents and words often exhibit only a *slight* degree of mixed-membership behavior. The regularization introduced in the paper is used to control the degree of polysemy of words permitted and to prefer sparsity in topic distributions of documents with flexibility. The utility of the regularization in exploiting sentiment-indicative features is evaluated using document perplexity and by using the models to predict star counts in movie and product reviews. Results of our experiments show that using the regularization to finely control the behavior of topic models leads to better perplexity and lower mean squared error rates in the star-prediction task.

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

CP10

Sparse Max-Margin Multiclass and Multi-Label Classifier Design for Fast Inference

We address the problems of sparse multiclass and multi-label classifier design and devise new algorithms using margin based ideas. Many online applications such as image classification or text categorization demand fast inference. State-of-the-art classifiers such as Support Vector Machines (SVM) are not preferred in such applications because of slow inference, which is mainly due to the large number of support vectors required to form the SVM classifier. We propose algorithms which solve primal problems directly by greedily adding required number of basis functions into the classifier model. Experiments on various real-world data sets demonstrate that the proposed algorithms output significantly smaller number of basis functions, while achieving nearly the same generalization performance as that given by SVM and other state-of-the-art sparse classifiers. Thus, the proposed algorithms provide powerful alternatives to the existing algorithms for faster classification inference.

Tanuja Ganu
IBM Research, India
tanuja.ganu@in.ibm.com

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

S Sudararajan
Microsoft Research, India
ssrajan@microsoft.com

CP10

A New Perspective on Convex Relaxations of

Sparse Svm

This paper proposes a convex relaxation of a sparse support vector machine (SVM) based on the perspective relaxation of mixed-integer nonlinear programs. We seek to minimize the zero-norm of the hyperplane normal vector with a standard SVM hinge-loss penalty and extend our approach to a zero-one loss penalty. The relaxation that we propose is a second-order cone formulation that can be efficiently solved by standard conic optimization solvers. We compare the optimization properties and classification performance of the second-order cone formulation with previous sparse SVM formulations suggested in the literature.

Noam Goldberg
Carnegie Mellon University
noam.goldberg@gmail.com

CP10

Reduced Set Kpca for Improving the Training and Execution Speed of Kernel Machines

We present a practical, and theoretically well- founded, approach to improve the speed of kernel manifold learning algorithms relying on spectral decomposition. Utilizing recent insights in kernel smoothing and learning with integral operators, we propose Reduced Set KPCA (RSKPCA), which also suggests an easy-to-implement method to remove or replace samples with minimal effffect on the empirical operator. A simple data point selection procedure is given to generate a substitute density for the data, with accuracy that is governed by a user-tunable parameter ℓ. The effffect of the approximation on the quality of the KPCA solution, in terms of spectral and operator errors, can be shown directly in terms of the density estimate error and as a function of the parameter ℓ. We show in experiments that RSKPCA can improve both training and evaluation time of KPCA by up to an order of magnitude, and compares favorably to the widely-used Nystrom and density-weighted Nystrom methods.

Hassan A. Kingravi
Georgia Institute of Technology
kingravi@gatech.edu

CP10

An Empirical Study of the Suitability of Class Decomposition for Linear Models: When Does It Work Well?

The presence of sub-classes within a data sample suggests a class decomposition approach to classification, where each subclass is treated as a new class. Class decomposition can be effected using multiple linear classifiers in an attempt to outperform a single global linear classifier; the goal is to gain in model complexity while keeping error variance low. We describe a study aimed at understanding the conditions behind the success or failure of class decomposition when combined with linear classifiers. We identify two relevant data properties as indicators of the suitability of class decomposition: 1) linear separability; and 2) class overlap. We use well-known data complexity measures to evaluate the presence of these properties in a data sample. Our methodology indicates when to avoid performing class decomposition based on such data properties. In addition we conduct a similar analysis at a more granular level for data samples marked as suitable for class decomposition. This extra analysis shows how to improve in efficiency during class decomposition. From an empirical standpoint, we test our technique on several real-world classification problems; results validate our methodology.

Francisco Ocegueda-Hernandez, Ricardo Vilalta
Department of Computer Science, University of Houston
ocegueda@cs.uh.edu, vilalta@cs.uh.edu

PP1

Time-Sensitive Classification of Behavioral Data

This paper addresses a classification task under a practical setting, where the amount of observations made before the prediction is considered a cost. This setting reflects rewards that depends on the response time, e.g., in surveillance and diagnostic systems. However, there generally exists a trade-off between such cost and the accuracy and the reliability. We formalize the task as the classification of subsequences in a time series, aimed at predicting both the label of events from subsequent observations and when to commit to a response considering the trade-off. We propose a training algorithm for an ensemble of classifiers that makes predictions from subsequences of different lengths, respectively. The ensemble returns the earliest confident prediction among the individual classifiers, which are trained jointly considering their temporal dependence. We compare the proposed algorithm against conventional approaches over a collection of behavior trajectory datasets.

Shin Ando
Guma University
shin.ando@acm.org

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

PP1

An Examination of Practical Granger Causality Inference

Granger causality is one of the most popular techniques in uncovering the temporal dependencies among time series; however it faces two main challenges: (i) the spurious effect of unobserved time series and (ii) the computational challenges in high dimensional settings. In this paper, we utilize the confounder path delays to find a subset of time series that via conditioning on them we are able to cancel out the spurious confounder effects. After study of consistency of different Granger causality techniques, we propose Copula-Granger and show that while it is consistent in high dimensions, it can efficiently capture non-linearity in the data.

Mohammad Taha Bahadori, Yan Liu
University of Southern California
mohammab@usc.edu, yanliu.cs@usc.edu

PP1

Bregman Divergence and Triangle Inequality

While Bregman divergences have been used for clustering and embedding problems in recent years, the facts that they are asymmetric and do not satisfy triangle inequality have been a major concern. In this paper, we investigate the relationship between two families of symmetrized Bregman divergences and metrics that satisfy the triangle inequality. The first family can be derived from any

well-behaved convex function. The second family generalizes the Jensen-Shannon divergence, and can only be derived from convex functions with certain conditional positive definiteness structure. We interpret the required structure in terms of cumulants of infinitely divisible distributions, and related results in harmonic analysis. We investigate kmeans-type clustering problems using both families of symmetrized divergences, and give efficient algorithms for the same.

Sreangsu Acharyya
University of Texas Asutin
sreangsu@gmail.com

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

Daniel L. Boley
University of Minnesota
Department of Computer Science
boley@cs.umn.edu

PP1

Joint Segmentation and Clustering in Text Corpuses

To allow for more effective information extraction from digital corpuses, we propose combining two common document processing tasks, (i) clustering and (ii) segmentation, into one process to simultaneously segment documents within a corpus and assign each segment to a category. We have developed a generative probabilistic model to accomplish this task and show by experiments that it can accurately partition documents and assign meaningful categories to each partition.

Samuel J. Blasiak
George Mason University
sblasiak@gmu.edu

Sithu Sudarsan
FDA
sdsudarsan@ualr.edu

Huzefa Rangwala
George Mason University
rangwala@cs.gmu.edu

PP1

Automatic Detection and Correction of Multi-Class Classification Errors Using System Whole-Part Relationships

Real-world dynamic systems often exhibit a hierarchical system-subsystem structure. In this paper, we propose DETECTOR, a hierarchical method for detecting and correcting forecast errors by employing the *whole-part* relationships. Experimental results show that DETECTOR can successfully detect and correct forecasting errors made by state-of-art classifier ensemble techniques and traditional single classifier methods at an average rate of 22% in seasonal forecasting of hurricanes and landfalling hurricanes in North Atlantic and North African rainfall.

Zhengzhang Chen
Northwestern University
zhengzhangchen@northwestern.edu

John Jenkins
North Carolina State University
jpjenki2@ncsu.edu

Alok Choudhary
Dept. of Electrical Engineering and Computer Science
Northwestern University, Evanston, USA
choudhar@eecs.northwestern.edu

Jinfeng Rao
Zhejiang University
North Carolina State University
raojinfeng@gmail.com

Fredrick Semazzi, Anatoli Melechko
North Carolina State University
fred_semazzi@ncsu.edu, tolik@sciencedom.com

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

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

PP1

Contextual Time Series Change Detection

Time series data are common in a variety of fields ranging from economics to medicine and manufacturing. As a result, time series analysis and modeling has become an active research area in statistics and data mining. In this paper, we focus on a type of change we call *contextual time series change* (CTC) and propose a novel two-stage algorithm to address it. In contrast to traditional change detection methods, which consider each time series separately, CTC is defined as a change relative to the behavior of a group of related time series. As a result, our proposed method is able to identify novel types of changes not found by other algorithms. We demonstrate the unique capabilities of our approach with several case studies on real-world datasets from the financial and Earth science domains.

XI Chen
University of Minnesota
chen@cs.umn.edu

Karsten Steinhaeuser
Department of Computer Science and Engineering
University of Minnesota
ksteinha@cs.umn.edu

Shyam Boriah
Department of Computer Science
University of Minnesota
sboriah@cs.umn.edu

Snigdhansu Chatterjee
School of Statistics
University of Minnesota, Twin Cities
chatterjee@stat.umn.edu

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

PP1

Very Fast Similarity Queries on Semi-Structured Data from the Web

In this paper, we propose a single low-dimensional representation for entities found in different datasets on the web. Our proposed PIC-D embeddings can represent large D-partite graphs using small number of dimensions enabling fast similarity queries. Our experiments show that this representation can be constructed in small amount of time (linear in number of dimensions). We demonstrate how it can be used for variety of similarity queries like set expansion, automatic set instance acquisition, and column classification. Our approach results in comparable precision with respect to task specific baselines and up to two orders of magnitude improvement in terms of query response time.

Bhavana Dalvi, William Cohen
Carnegie Mellon University
bbd@cs.cmu.edu, wcohen@cs.cmu.edu

PP1

Topic Models For Feature Selection in Document Clustering

We investigate the idea of using a topic model such as the popular Latent Dirichlet Allocation model as a feature selection step for unsupervised document clustering, where documents are clustered using the proportion of the various topics that are present in each document. One concern with using "vanilla' LDA as a feature selection method for input to a clustering algorithm is that the Dirichlet prior on the topic mixing proportions is too smooth and well-behaved. It does not encourage a "bumpy' distribution of topic mixing proportion vectors, which is what one would desire as input to a clustering algorithm. As such, we propose two variant topic models that are designed to do a better job of producing topic mixing proportions that have a good clustering structure.

Anna Drummond
Rice University, Computer Science
ag37@rice.edu

Zografoula Vagena
LogicBlox Inc
zografoula.vagena@logicbloc.com

Chris Jermaine
Computer Science Department, Rice University
cmj4@rice.edu

PP1

A Nonparametric Mixture Model for Topic Modeling over Time

A single, stationary topic model such as latent Dirichlet allocation is inappropriate for modeling corpora that span long time periods, as the popularity of topics is likely to change over time. A number of models that incorporate time have been proposed, but in general they either exhibit limited forms of temporal variation, or require computationally expensive inference methods. In this paper we propose nonparametric Topics over Time (npTOT), a model for time-varying topics that allows an unbounded number of topics and ?exible distribution over the temporal variations in those topics popularity. We develop a collapsed Gibbs sampler for the proposed model and compare against existing models on synthetic and real document sets.

Ahmed Hefny, Avinava Dubey, Sinead Williamson
Carnegie Mellon University
ahefny@cs.cmu.edu, akdubey@cs.cmu.edu,
sinead@cs.cmu.edu

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

PP1

Discriminative Transfer Learning on Manifold

We impose a discriminative regression model over the latent factors to enhance the capability of label prediction in transfer learning. Moreover, we propose to minimize the Maximum Mean Discrepancy in the latent manifold subspace, as opposed to typically in the original data space, to bridge the gap between different domains. We formulate these objectives into a joint optimization framework simultaneously. An iterative algorithm is developed and empirical study validates the superiority in transfer learning classification.

Zheng Fang
Dept. of Information Science and Electronic Engineering, Zhe
fangzheng354@zju.edu.cn

Zhongfei Zhang
Dept. of Information Science and Electronic Engineering
Zhejiang University,China
zhongfei@zju.edu.cn

PP1

SemInf: A Burst-Based Semantic Influence Model for Biomedical Topic Influence

In this paper we consider the problem of mining influence in a network of topics, where we seek to model direct influences between topics over time, in the form of bursts of topic occurrence and this influence propagates among topics, leading to frequent occurrences of the topics. We propose a novel model: **SemInf**. As topics can recur, influence in our model is not constant or single-timestamp, but is instead multi-timestamp, with periods of influence that can span multiple time intervals. A topic hierarchy is used to provide a distance measure among topics and characterize their semantic relatedness. Experiments on biomedical topics give some surprising results, showing both that our model is successful at identifying topics with high impact, and that it can be potentially used as an alternative model of impact in the scientific literature (which can be useful when citation information is not available).

Dan He
IBM T.J. Watson
IBM
dhe@us.ibm.com

Douglas Parker
UCLA
stott@cs.ucla.edu

PP1

Pinch Ratio Clustering from a Topologically Intrinsic Lexicographic Ordering

This paper introduces an algorithm for determining data clusters called TILO/PRC (Topologically Intrinsic Lexicographic Ordering/Pinch Ratio Clustering). The theoretical foundation for this algorithm uses ideas from topology (particularly knot theory) suggesting that it should be very flexible and robust with respect to noise. The TILO portion of the algorithm progressively improves a linear ordering of the points in a data set until the ordering satisfies a topological condition called strongly irreducible. The PRC algorithm then divides the data set based on this ordering and a heuristic metric called the pinch ratio. We demonstrate the effectiveness of TILO/PRC for finding clusters in a wide variety of real and synthetic data sets and compare the results to existing clustering methods. These results verify that both the theoretical foundations of TILO and the heuristic notion of pinch ratio are reasonable.

Douglas R. Heisterkamp, Jesse Johnson
Oklahoma State University
doug@cs.okstate.edu, jjohnson@math.okstate.edu

PP1

Retweeting: An Act of Viral Users, Susceptible Users, Or Viral Topics?

When a user retweets, there are three behavioral factors that cause the actions. They are the topic virality, user virality and user susceptibility. Topic virality captures the degree to which a topic attracts retweets by users. For each topic, user virality and susceptibility refer to the likelihood that a user attracts retweets and performs retweeting respectively. To model a set of observed retweet data as a result of these three topic specific factors, we first represent the retweets as a three-dimensional tensor of the tweet authors, their followers, and the tweets themselves. We then propose the V2S model, a tensor factorization model, to simultaneously derive the three sets of behavioral factors. Our experiments on a real Twitter data set show that the V2S model can effectively mine the behavioral factors of users and tweet topics during an election event. We also demonstrate that the V2S model outperforms the other topic based models in retweet prediction.

Tuan-Anh Hoang
Living Analytics Research Centre
Singapore Management University
tahoang.2011@phdis.smu.edu.sg

Ee-Peng Lim
Singapore Management University
eplim@smu.edu.sg

PP1

Time Series Classification under More Realistic Assumptions

Abstract not available at time of publication.

Bing Hu
University of California, Riverside
University of California, Riverside
bhu002@ucr.edu

Yanping Chen, Eamonn Keogh
University of California, Riverside
ychen053@ucr.edu, eamonn@cs.ucr.edu

PP1

Finding Affordable and Collaborative Teams from a Network of Experts

Given an expert network, we tackle the problem of finding a team of experts that covers a set of required skills and also minimizes the communication cost as well as the personnel cost of the team. First, we propose several algorithms that receive a budget on one objective and minimizes the other objective within the budget with guaranteed performance bounds. Then, an approximation algorithm is proposed to find a set of Pareto-optimal teams.

Aijun An, Mehdi Kargar, Morteza Zihayat
York University
aan@cse.yorku.ca, kargar@cse.yorku.ca, zihayatm@cse.yorku.ca

PP1

IBSM: Interval-Based Sequence Matching

Sequences of event intervals appear in several application domains including sign language, medicine, motion databases, and sensor networks. Such sequences comprise events that occur at time intervals. In this paper, we propose a new method, called IBSM, for comparing such sequences, which performs full sequence matching using a vector-based representation of the original sequences. Experiments on eight real datasets show that IBSM outperforms existing state-of-the-art methods in terms of nearest neighbor classification accuracy and runtime.

Alexios Kotsifakos
University of Texas at Arlington
alexios.kotsifakos@mavs.uta.edu

Panagiotis Papapetrou
Birkbeck, University of London
panos@dcs.bbk.ac.uk

Vassilis Athitsos
University of Texas at Arlington
athitsos@uta.edu

PP1

Modeling the Diffusion of Preferences on Social Networks

Information diffusion on social networks has been studied for decades. Most models consider the propagated information as single values. Representing media as single values however would not be proper for certain cases such as voter preferences toward candidates in elections. The paper aims at the diffusion of preferences on social networks, which is a novel problem to solve in this direction. We propose a preference propagation model to handle the diffusion of vector-type information instead of single values. We further prove the convergence of diffusion, and that a consensus among strongly connected nodes can eventually be reached. We extract relevant information from a public bibliography datasets to evaluate our model. Lastly, we exploit the extracted data to demonstrate the usefulness of our model and compare it with other well-known diffusion models such as independent cascade, linear threshold, and diffusion rank. We find that our model consistently

outperforms other models.

J Lou
Department of Electrical Engineering
National Taiwan University
kaeaura@gmail.com

Fu-Min Wang, Chin-Hua Tsai, San-Chuan Hung,
Perng-Hwa Kung, Shou-De Lin
Department of Computer Science and Information Engineering
National Taiwan University
r98723077@ntu.edu.tw, zhichin@gmail.com, c2016.tw@gmail.com, answerseeker95@gmail.com, sdlin@csie.ntu.edu.tw

PP1

Mining Labelled Tensors by Discovering Both Their Common and Discriminative Subspaces

Practical data are usually generated from different time periods or by different class labels, which are represented by a sequence of multiple tensors associated with different labels. This raises the problem that when one needs to analyze and compare multiple tensors, existing non-negative tensor factorization (NTF) is unsuitable for discovering all potentially useful patterns. To tackle this problem, we design a novel factorization algorithm called `CSNTF` (common subspace non-negative tensor factorization), which takes both features and class labels into account in the factorization process. Experiment results on solving graph classification problems demonstrate the power and the effectiveness of the subspaces discovered by our method.

Wei Liu
School of IT, the University of Sydney
weiliu.au@gmail.com

Jeffrey Chan, James Bailey, Christopher Leckie,
Ramamohanarao Kotagiri
The University of Melbourne
jeffrey.chan@unimelb.edu.au, baileyj@unimelb.edu.au, caleckie@unimelb.edu.au, rao@csse.unimelb.edu.au

PP1

Modeling Clinical Time Series Using Gaussian Process Sequences

Abstract not available at time of publication.

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

PP1

Integrity Verification of K-Means Clustering Outsourced to Infrastructure As a Service (IaaS) Providers

The Cloud-based infrastructure-as-a-service ($IaaS$) paradigm enables a client to outsource her dataset and data mining tasks to the Cloud. However, as the Cloud may not be fully trusted, it raises serious concerns about the *integrity* of the mining results returned by the Cloud. To this end, in this paper, we provide a focused study about how to perform integrity verification of the k-means clustering task outsourced to an $IaaS$ provider. We consider the untrusted *sloppy* $IaaS$ service provider that intends to return wrong clustering results by terminating the iterations early to save computational cost. We develop both probabilistic and deterministic verification methods to catch the incorrect clustering result by the service provider. Our experimental results show that our verification methods can effectively and efficiently capture the sloppy service provider.

Wendy Hui Wang
Stevens Institute of Technology
hwang4@stevens.edu

Ruilin Liu, Philippos Mordohai
Department of Computer Science
Stevens Institute of Technology
rliu3@stevens.edu, philippos.mordohai@stevens.edu

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

PP1

Selective Transfer Learning for Cross Domain Recommendation

Collaborative filtering (CF) aims to predict users' ratings on items according to historical user-item preference data. In many real-world applications, preference data are usually sparse, which would make models overfit and fail to give accurate predictions. Recently, several research works show that by transferring knowledge from some manually selected source domains, the data sparseness problem could be mitigated. However for most cases, parts of source domain data are not consistent with the observations in the target domain, which may misguide the target domain model building. In this paper, we propose a novel criterion based on empirical prediction error and its variance to better capture the consistency across domains in CF settings. Consequently, we embed this criterion into a boosting framework to perform selective knowledge transfer.

Zhongqi Lu
Hong Kong University of Science & Technology
zluab@cse.ust.hk

Erheng Zhong
Hong Kong University of Science and Technology
Department of Computer Science and Engineering
ezhong@cse.ust.hk

Lili Zhao, Wei Xiang, Weike Pan
Hong Kong University of Science & Technology
skyezhao@cse.ust.hk, wxiang@cse.ust.hk, weikep@cse.ust.hk

Qiang Yang
Hong Kong University of Science and Technology
Department of Computer Science and Engineering
qyang@cse.ust.hk

PP1

Change Detection from Temporal Sequences of Class Labels: Application to Land Cover Change Mapping

Mapping land cover change is an important problem for the scientific community as well as policy makers. Traditionally, bi-temporal classification of satellite data is used

to identify areas of land cover change. However, these classification products often have errors due to classifier inaccuracy or poor data, which poses significant issues when using them for land cover change detection. In this paper, we propose a generative model for land cover label sequences and use it to reassign a more accurate sequence of land cover labels to every pixel.

Varun Mithal, Ankush Khandelwal
University of Minnesota- Twin Cities
mithal@cs.umn.edu, ankush@cs.umn.edu

Shyam Boriah
Department of Computer Science
University of Minnesota
sboriah@cs.umn.edu

Karsten Steinhaeuser
Department of Computer Science and Engineering
University of Minnesota, Twin Cities
ksteinha@umn.edu

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

PP1
Fractional Immunization in Networks

Preventing contagion in networks by targeting nodes is an important problem in public health and other domains. However, the assumption that selected nodes can be rendered completely immune does not hold for infections for which there is no vaccination or effective treatment. Instead, one can confer fractional immunity to some nodes by allocating variable amounts of infection-prevention resource to them. We formulate the problem to distribute a fixed amount of resource across nodes in a network such that the infection rate is minimized, prove that it is NP-complete and derive a highly effective and efficient linear-time algorithm. We demonstrate the efficiency and accuracy of our algorithm real-world networks including US-MEDICARE and state-level interhospital patient transfer data. We find that concentrating resources using our algorithm is up to *6 times* more effective than distributing them uniformly (as is current practice) or using network-based heuristics.

B. Aditya Prakash
CMU
badityap@cs.vt.edu

Lada Adamic
University of Michigan, Ann Arbor
School of Information, Center for the Study of Complex Syste
ladamic@umich.edu

Theodore Iwashyna
Univ. of Michigan-Ann Arbor
tiwashyn@umich.edu

Hanghang Tong
City Collge, CUNY
tong@cs.ccny.cuny.edu

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

PP1
Fast Shapelets: A Scalable Algorithm for Discovering Time Series Shapelets

Abstract not available at time of publication.

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

PP1
Mc-MinH: Metagenome Clustering Using Minwise Based Hashing

Abstract not available at time of publication.

Zeehasham Rasheed, Huzefa Rangwala
George Mason University
zrasheed@gmu.edu, rangwala@cs.gmu.edu

PP1
Shattering and Compressing Networks for Betweenness Centrality

The betweenness metric has always been intriguing and used in many analyses. Yet, it is one of the most computationally expensive kernels in graph mining. For that reason, making betweenness centrality computations faster is an important and well-studied problem. In this work, we propose the framework, BADIOS, which compresses a network and shatters it into pieces so that the centrality computation can be handled independently for each piece. Although BADIOS is designed and tuned for betweenness centrality, it can easily be adapted for other centrality metrics. Experimental results show that the proposed techniques can be a great arsenal to reduce the centrality computation time for various types and sizes of networks. In particular, it reduces the computation time of a 4.6 million edges graph from more than 5 days to less than 16 hours.

A. Erdem Sariyuce
The Ohio State University
sariyuce.1@osu.edu

Erik Saule, Kamer Kaya, Umit V. Catalyurek
The Ohio State University
Department of Biomedical Informatics
esaule@bmi.osu.edu, kamer@bmi.osu.edu, umit@bmi.osu.edu

PP1
CoSelect: Feature Selection with Instance Selection for Social Media Data

Feature selection is widely used in preparing high-dimensional data for effective data mining. Attribute-value data in traditional feature selection differs from social media data, although both can be large-scale. Social media data is inherently not independent and identically distributed (*i.i.d.*), but linked. Furthermore, there is a lot of noise. The quality of social media data can vary drastically. These unique properties present challenges as well as opportunities for feature selection. Motivated by these differences, we propose a novel feature selection framework, CoSelect, for social media data. In particular, CoSelect can exploit link information by applying social correlation theories, incorporate instance selection with feature selection, and select relevant instances and features simultaneously.

Experimental results on real-world social media datasets demonstrate the effectiveness of our proposed framework and its potential in mining social media data.

Jiliang Tang
Arizona State University
ARIZONA STATE UNIVERISTY
Jiliang.Tang@asu.edu

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

PP1

A Hierarchical Probabilistic Model for Low Sample Rate Home-Use Energy Disaggregation

Energy crisis and climate change have caused a global concern and motivated efforts to reduce energy consumption. Studies have shown that providing appliance-level consumption information can help users conserve a significant amount of energy. Existing methods focus on learning parallel signal signatures, but the inherent relationships between the signatures have not been well explored. This paper presents a Hierarchical Probabilistic Model for Energy Disaggregation (HPMED) to bridge the discriminative features, working states, and aggregated consumption.

Bingsheng Wang, Haili Dong
Computer Science, Virginia Tech
claren89@vt.edu, hailid@vt.edu

Arnold Boedihardjo
U. S. Army Corps of Engineers
arnold.p.boedihardjo@erdc.dren.mil

Feng Chen, Chang-Tien Lu
Computer Science, Virginia Tech
chenf@vt.edu, ctlu@vt.edu

PP1

On the Detectability of Node Grouping in Networks

In typical studies of node grouping detection, the grouping is presumed to have a certain type of correlation with the network structure, which is quantified by different fitness measures such as modularity and conductance. We study a fundamental problem: Given a particular grouping in a network, whether and to what extent it can be discovered with a given fitness measure. We propose two approaches of testing the detectability, namely ranking-based and correlation-based randomization tests, which can effectively predict the detectability of groupings of various types.

Chi Wang, Hongning Wang, Jialu Liu, Ming Ji, Lu Su, Yuguo Chen
University of Illinois at Urbana-Champaign
chiwang1@illinois.edu, wang296@illinois.edu, jliu64@illinois.edu, mingji1@illinois.edu, lusu2@illinois.edu, yuguo@illinois.edu

Jiawei Han
UIUC
hanj@illinois.edu

PP1

Robust Textual Data Streams Mining Based on Continuous Transfer Learning

In the data stream environment, since concept drift can occur at any point of the streams, it will certainly occur within chunks, which is called random concept drift. The paper proposed an approach, which is called chunk level-based concept drift method (CLCD), that can overcome this chunking problem by continuously monitoring chunk characteristics to revise the classifier based on transfer learning in positive and unlabeled (PU) textual data stream environment.

Yanshan Xiao, Bo Liu
UTS
xiaoyanshan@gmail.com, csbliu@gmail.com

Yanshan Xiao
Guangdong University of Technology
xiaoyanshan@gmail.com

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

Longbing Cao
University of Technology, Sydney
longbing.cao@uts.edu.au

Zhifeng Hao
Faculty of Computer, Guangdong University of Technology
mazfhao@scut.edu.cn

PP1

Mods: Multiple One-Class Data Streams Learning from Homogeneous Data

This paper presents a novel approach, called MODS, to build an accurate time evolving classifier from multiple one-class data streams learning time evolving classifier. We first construct local one-class classifiers on the labeled positive examples from each sub-data stream respectively and collect the informative examples around each local classifier. We then construct a global one-class classifier on the collected informative examples. Experiments showed our MODS approach can achieve high performance and efficiency.

Yanshan Xiao, Bo Liu
UTS
xiaoyanshan@gmail.com, csbliu@gmail.com

Yanshan Xiao
Guangdong University of Technology
xiaoyanshan@gmail.com

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

Zhifeng Hao
Faculty of Computer, Guangdong University of Technology
mazfhao@scut.edu.cn

PP1

Graphical Modeling of Macro Behavioral Targeting in Social Networks

We investigate a class of emerging online marketing challenges in social networks; macro behavioral targeting (MBT) is introduced as non-personalized broadcasting efforts to massive populations. We propose a new probabilistic graphical model for MBT. Further, a linear-time approximation method is proposed to circumvent an intractable parametric representation of user behaviors. We compare the proposed model with the existing state-of-the-art method on real datasets from social networks. Our model outperforms in all categories by comfortable margins.

Yusheng Xie, Zhengzhang Chen, Kunpeng Zhang, Md. Mostofa Ali Patwary, Yu Cheng, Haotian Liu, Ankit Agrawal
Northwestern University
yushengxie2011@u.northwestern.edu,
zhengzhangchen@northwestern.edu,
kzh980@eecs.northwestern.edu,
mpatwary@eecs.northwestern.edu,
ych130@eecs.northwestern.edu,
haotian@u.northwestern.edu,
ankitag@eecs.northwestern.edu

Alok Choudhary
Dept. of Electrical Engineering and Computer Science
Northwestern University, Evanston, USA
choudhar@eecs.northwestern.edu

PP1

Learning Topics in Short Texts by Non-Negative Matrix Factorization on Term Correlation Matrix

The severe sparsity of short texts hinders existing topic models to learn reliable topics.To tackle this problem, we formulated the topic learning problem as symmetric non-negative matrix factorization on the term correlation matrix. After learning the topics, we can easily infer the topics of documents. Experimental results on three data sets show that our method provides substantially better performance than the baseline methods.

Xiaohui Yan
Institute of Computing Technology of the Chinese Academy of
l0he1g@gmail.com

Jiafeng Guo
nstitute of Computing Technology of the Chinese Academy of S
guojiafeng@ict.ac.cn

Shenghua Liu
Institute of Computing Technology of the Chinese Academy of
liushenghua@ict.ac.cn

Xueqi Cheng
Institute of Computing Technology of the Chinese Academy
cxq@ict.ac.cn

Yanfeng Wang
Sogou Inc
wangyanfeng@sogou-inc.com

PP1

Set Coverage Problems in a One-Pass Data Stream

The Max-k-Cover and the Partial-Cover problems are important combinatorial optimization problems, and have various applications. Dealing with large-scale dataset or in an online environment, we need an one-pass algorithm other than the in-memory standard greedy solution. Previous one-pass algorithms for the Max-k-Cover problem cannot be extended to the Partial-Cover problem. In this paper, we propose a novel one-pass streaming algorithm producing a prefix-optimal ordering of sets which can easily be used to solve both problems. Our algorithm consumes space linear to the size of the universe of elements. The processing time for a set is linear to the size of this set. We also show with the aid of computer simulation that the approximation ratio of the Max-k-Cover problem is around 0.3. We conduct experiments on extensive datasets to study the performance of our algorithm and demonstrate its efficiency and quality.

Huiwen Yu, Dayu Yuan
Department of Computer Science and Engineering
The Pennsylvania State University
hwyu@cse.psu.edu, duy113@cse.psu.edu

PP1

Sentiment Topic Model with Decomposed Prior

Abstract not available at time of publication.

Chengtao Li
Tsinghua University
lichengtao2010@gmail.com

Jianwen Zhang, Jian-Tao Sun, Zheng Chen
Microsoft Research Asia
jianwenzh@gmail.com, jtsun@microsoft.com, zhengc@microsoft.com

PP1

Butterfly Mixing: Accelerating Incremental-Update Algorithms on Clusters

Incremental model-update strategies are widely used in machine learning and data mining. By "incremental update' we refer to models that are updated many times using small subsets of the training data. Two well-known examples are stochastic gradient and MCMC. Both provide fast sequential performance and have generated many of the best-performing methods for particular problems. But these methods are difficult to adapt to parallel because of the overhead of distributing model updates through the network. Updates can be locally batched to reduce communication overhead, but convergence typically suffers as the batch size increases. In this paper we introduce butterfly mixing, an approach which interleaves communication with computation. We evaluate butterfly mixing on stochastic gradient algorithms for logistic regression and SVM. Results show that butterfly mix steps are fast and failure-tolerant, and overall we achieved a 3.3x speed-up over full mix on an Amazon EC2 cluster.

Huasha Zhao
EECS
UC Berkeley
hzhao@eecs.berkeley.edu

John Canny

University of California, Berkeley
jfc@cs.berkeley.edu

PP1

Topic-Level Expert Modeling in Community Question Answering

As more knowledge is shared in Community Question Answerings, how to use the repository for solving new questions has become a crucial problem. In this paper, we tackle the problem by finding experts from the question answering history first and then recommending the appropriate experts to answer the new questions. We develop the Topic-level Expert Learning (TEL) model to find experts on topic level in CQA. The main difference between TEL and other generative models is that TEL can automatically adjust and update the sampling parameters during iterations in order to better model the experts on topic level. The experimental results show that our method can effectively find experts to answer new questions and can better predict best responders for new questions.

Tong Zhao
Tsinghua University
zt882001@Hotmail.com

Naiwen Bian, Chunping Li, Mengya Li
Tsinghua University
Beijing, China
bnwivy@gmail.com, cli@tsinghua.edu.cn, imyli1024@gmail.com

PP1

It Is Not Just What We Say, But How We Say Them: Lda-Based Behavior-Topic Model

Textual information exchanged among users on online social network platforms provides deep understanding into users' interest and behavioral patterns. However, unlike traditional text-dominant settings such as offline publishing, one distinct feature for online social network is users' rich interactions with the textual content, which, unfortunately, has not yet been well incorporated in the existing topic modeling frameworks. In this paper, we propose an LDA-based behavior-topic model (B-LDA) which jointly models user topic interests and behavioral patterns. We focus the study of the model on online social network settings such as microblogs like Twitter where the textual content is relatively short but user interactions on them are rich. Empirical evaluations show the topics obtained by our model are both informative and insightful, and our method can help the task of Twitter followee recommendation.

Minghui Qiu, Feida Zhu, Jing Jiang
Singapore Management University
minghui.qiu.2010@smu.edu.sg, fdzhu@smu.edu.sg, jingjiang@smu.edu.sg

PP1

Feature Selection by Joint Graph Sparse Coding

This paper takes manifold learning and regression simultaneously into account to perform unsupervised spectral feature selection. We first extract the bases of the data, and then represent the data sparsely using the extracted bases by proposing a novel joint graph sparse coding model, JGSC for short. We design a new algorithm TOSC to compute the resulting objective function of JGSC. We repeat the extraction and the TOSC calculation until the value of the objective function of JGSC satisfies pre-defined conditions. Eventually the derived new representation of the data may only have a few non-zero rows, and we delete the zero rows (a.k.a. zero-valued features) to conduct feature selection on the new representation of the data. Our empirical studies demonstrate that the proposed method outperforms several state-of-the-art algorithms on real datasets in term of the kNN classification performance.

Xiaofeng Zhu
The university of Queensland
seanzhuxf@gmail.com

Xindong Wu
The University of Vermont, USA
xwu@uvm.edu

Wei Ding
The University of Massachusetts Boston, USA
ding@cs.umb.edu

Shichao Zhang
Guangxi Normal University, China
zhangsc@mailbox.gxnu.edu.cn