IP1

Title Not Available At Time Of Publication

Abstract Not Available At Time Of Publication

Bernhard Schölkopf
MPI for Intelligent Systems
bernhard.schoelkopf@tuebingen.mpg.de

IP2

Title Not Available At Time Of Publication

Abstract will follow.

Charles Elkan
University of California, San Diego
elkan@cs.ucsd.edu

IP3

Title Not Available At Time Of Publication - Ideker

Abstract will follow.

Trey G. Ideker
University of California, San Diego
tideker@ucsd.edu

IP4

From Robots to Biomolecules: Computing Meets the Physical World

The development of fast and reliable motion planning algorithms has deeply influenced many domains in robotics, such as industrial automation and autonomous exploration, but has also contributed novel methodologies to distant domains such as computational structural biology. This talk will present recent work on the computation of low-level plans from high-level specifications. High-level specifications declare what the robot must do, rather than how this task is to be done. The talk will also discuss robotics-inspired methods for computing the flexibility of proteins and for molecular docking with the ultimate goal of deciphering molecular function and aiding the discovery of new therapeutics.

Lydia Kavraki
Rice University
kavraki@rice.edu

CP1

Mixtures of Block Models for Brain Networks

Block models are a popular method for simplifying a single graph into a set of blocks and interactions between those blocks. A recent innovation is to extend block modeling to a collection of graphs (e.g. RESCAL) to discover *one common block structure* amongst the graphs. However, these approaches are unsuitable in many domains where the collection can comprise items from significantly different groups. Consider the focus of this paper: fMRI analysis on scans of young healthy and Alzheimer's affected individuals. There are implicitly two underlying block structures (one for each group) and some individuals may exhibit the behavior of both. We propose a novel mixtures of block models (MBM) framework that explicitly models each single graph as a linear combination of a small number of block models. Experimental results on synthetic data show that our method is able to recover the ground-truth models. In real-world experiments with fMRI data we show that with proper factorization parameters, MBM (1) outperforms the single block structure models and (2) demonstrates significant structural patterns of brain networks at the cohort level.

Zilong Bai
University of California, Davis
zlbai@ucdavis.edu

Peter Walker
Naval Medical Research Center
peter.b.walker.mil@mail.mil

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

CP1

Image Constrained Blockmodelling: A Constraint Programming Approach

Blockmodelling is an important technique for detecting underlying patterns in graphs. However, existing blockmodelling algorithms do not provide the user with any explicit control to specify which patterns might be of interest. Furthermore, existing algorithms focus on finding standard community structures in graphs, and are likely to overlook informative but more complex patterns, such as hierarchical or ring blockmodel structures. In this paper, we propose a generic constraint programming framework for blockmodelling, which allows a user to specify and search for complex blockmodel patterns in graphs. Our proposed framework can be incorporated into existing iterative blockmodelling algorithms, operating as a hybrid optimization scheme that provides high flexibility and expressiveness. We demonstrate the power of our framework for discovering complex patterns, via experiments over a range of synthetic and real data sets.

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

Jeffrey Chan
RMIT University
jeffrey.chan@rmit.edu.au

Peter. J Stuckey
University of Melbourne
pstuckey@unimelb.edu.au

James Bailey, Christopher Leckie
The University of Melbourne
baileyj@unimelb.edu.au, caleckie@unimelb.edu.au

Kotagiri Ramamohanarao
University of Melbourne
kotagiri@unimelb.edu.au

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

CP1

Unsupervised Neural Categorization for Scientific

Publications

Most conventional document categorization methods require a large number of documents with labeled categories for training. These methods are hard to be applied in scenarios, such as scientific publications, where training data is expensive to obtain and categories could change over years and across domains. In this work, we propose UNEC, an unsupervised representation learning model that directly categories documents without the need of labeled training data. Specifically, we develop a novel cascade embedding approach. We first embed concepts, i.e., significant phrases mined from scientific publications, into continuous vectors, which capture concept semantics. Based on the concept similarity graph built from the concept embedding, we further embed concepts into a hidden category space, where the category information of concepts becomes explicit. Finally we categorize documents by jointly considering the category attribution of their concepts. Our experimental results show that UNEC significantly outperforms several strong baselines on a number of real scientific corpora, under both automatic and manual evaluation.

Keqian Li, Hanwen Zha, Yu Su
University of California Santa Barbara
klee@cs.ucsb.edu, hwzha@cs.ucsb.edu, ysu@cs.ucsb.edu

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

CP1

Many-to-Many Correspondences Between Partitions: Introducing a Cut-Based Approach

Let $\mathcal{P}$ and $\mathcal{P}'$ be finite partitions of the set V. Finding good correspondences between the parts of $\mathcal{P}$ and those of $\mathcal{P}'$ is helpful in classification, pattern recognition, and network analysis. Unlike common similarity measures for partitions that yield only a single value, we provide specifics on how $\mathcal{P}$ and $\mathcal{P}'$ correspond to each other. To this end, we first define natural collections of best correspondences under three constraints. In case of the first one, the best correspondences form a minimum cut basis of a certain bipartite graph, whereas the other two lead to minimum cut bases of $\mathcal{P}$ w.r.t. $\mathcal{P}'$. We also introduce a fourth constraint that tightens the third one; both are useful for finding consensus partitions. We then develop branch-and-bound algorithms for finding minimum P_s-P_t cuts of $\mathcal{P}$ and thus $|\mathcal{P}|-1$ best correspondences under constraints two to four, respectively. In a case study, we use the correspondences to gain insight into a community detection algorithm. The results suggest, among others, that only very minor losses in the quality of the correspondences occur if the branch-and-bound algorithm is restricted to its greedy core. Thus, even for graphs with more than half a million nodes and hundreds of communities, we can find hundreds of best or almost best correspondences in less than a minute.

Roland Glantz
Karlsruhe Institute of Technology (KIT)
glantz@ira.uka.de

Henning Meyerhenke
University of Cologne
h.meyerhenke@uni-koeln.de

CP1

Graph Sketching-Based Space-Efficient Data Clustering

We address the problem of recovering arbitrary-shaped clusters from data in the context of high space constraints, e.g. in real-world applications when analysis algorithms are directly deployed on resources-limited mobile devices collecting the data. We present DBMSTClu a new space-efficient density-based nonparametric method working on a Minimum Spanning Tree (MST) recovered from a limited number of linear measurements, a sketched version of the dissimilarity graph $\mathcal{G}$ between the N objects to cluster. Unlike k-means family algorithms, it does not fail at distinguishing clusters with particular forms thanks to the property of the MST for expressing the underlying structure of a graph. No input parameter is needed contrarily to DBSCAN or Spectral Clustering. An approximate MST is retrieved by following the dynamic semi-streaming model in handling the dissimilarity graph $\mathcal{G}$ as a stream of edge weight updates which is sketched in one pass over the data into a compact structure requiring $O(Npolylog(N))$ space, far better than the theoretical memory cost $O(N^2)$ of $\mathcal{G}$. The recovered approximate MST $\mathcal{T}$ as input, DBMSTClu then successfully detects the right number of nonconvex clusters by performing relevant cuts on $\mathcal{T}$ in a time linear in N. We provide theoretical guarantees on the quality of the clustering partition and also demonstrate its advantage over the existing state-of-the-art on several datasets.

Anne Morvan
CEA and Université Paris-Dauphine
anne.morvan@cea.fr

Krzysztof Choromanski
Google Brain Robotics
kchoro@google.com

Cédric Gouy-Pailler
CEA
cedric.gouy-pailler@cea.fr

Jamal Atif
Université Paris-Dauphine
jamal.atif@lamsade.dauphine.fr

CP1

NLRR++: Scalable Subspace Clustering Via Non-Convex Block Coordinate Descent

Exploring the multiple subspace structures of data such as Low-Rank Representation is effective in subspace clustering. Low-Rank Representation formulated as a constrained convex optimization problem needs SVD in every step, which is challenging regarding time complexity and memory. Recently, its non-convex formulation–NLRR has been proposed for subspace clustering. However, NLRR can't scale to problems with large datasets since it requires the inversion of a large matrix. In this study, we develop a faster method for solving NLRR problem, with time complexity per epoch reduced from $O(n^3)$ to $O(mnd)$ and memory cost from $O(n^2)$ to $O(mn)$(usually $d \ll m \ll n$). The main idea of our method is to reformulate NLRR as a sum of rank-one components and apply a column-wise block coordinate descent to update each component iteratively. Our new method is considerably faster and more accurate in practice for subspace clustering. The proposed method is guaranteed to converge to stationary points. Moreover, considering the high demand in memory and computational time for the final spectral clustering phase, we also propose an efficient clustering approach which further boosts the performance of subspace clustering. Experiments on simulations and real datasets confirm the efficiency of our proposed NLRR++. In particular, we are about 12 times faster than state-of-the-art subspace clustering method, and our method is the only one that can scale to the Imagenet with 120K samples.

Jun Wang
Harbin Institute of Technology
junwangsd16@gmail.com

Cho-Jui Hsieh
University of California, Davis
chohsieh@ucdavis.edu

Daming Shi
Harbin Institute of Technology
Shenzhen University
d.m.shi@hotmail.com

CP2

Causal Inference on Event Sequences

Given two discrete valued time series—that is, event sequences—of length n can we tell whether they are causally related? That is, can we tell whether x^n causes y^n, whether y^n causes x^n? Can we do so without having to make assumptions on the distribution of these time series, or about the lag of the causal effect? And, importantly for practical application, can we do so accurately and efficiently? These are exactly the questions we answer in this paper. We propose a causal inference framework for event sequences based on information theory. We build upon the well-known notion of Granger causality, and define causality in terms of compression. We infer that x^n is likely a cause of y^n if y^n can be (much) better sequentially compressed given the past of both y^n and x^n, than for the other way around. To compress the data we use the notion of sequential normalized maximal likelihood, which means we use minimax optimal codes with respect to a parametric family of distributions. To show this works in practice, we propose `CUTE`, a linear time method for inferring the causal direction between two event sequences. Empirical evaluation shows that `CUTE` works well in practice, is much more robust than transfer entropy, and ably reconstructs the ground truth on river flow and spike train data.

Kailash Budhathoki
Max Planck Institute for Informatics and Saarland University
kbudhath@mpi-inf.mpg.de

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

CP2

Jump: A Fast Deterministic Algorithm to Find the Closest Pair of Subsequences

In this paper we address a classical sequence mining problem, namely, that of finding the Closest Pair of Subsequences. Given a sequence A of length n, the problem is to identify two non-overlapping subsequences of length l each in A, such that their distance is minimum from among all such pairs. This is a fundamental problem that has a wide range of applications such as time series data mining, sequence data pattern matching, data signature identification, biological motif mining, metagenomic clustering, etc. To solve this problem, the state-of-the-art algorithm takes advantage of the overlapping parts of consecutive subsequences. By exploiting these overlaps, researchers have developed an algorithm with a run time of $O(n^2)$, which is independent of the dimension l. In this paper, we propose a deterministic algorithm called JUMP, which further pushes the limit by skipping unnecessary comparisons and multiplication operations, and improves the running time by a large factor. We have performed extensive experiments using standard benchmark datasets, and found that JUMP outperforms existing $O(n^2)$ methods by a factor of up to 100. Our experiments cover different settings of n and l and provide the readers a comprehensive and unbiased comparison under different conditions.

Xingyu Cai, Shanglin Zhou, Sanguthevar Rajasekaran
University of Connecticut
xingyu.cai@uconn.edu, shanglin.zhou@uconn.edu, sanguthevar.rajasekaran@uconn.edu

CP2

Mining Top-K Quantile-Based Cohesive Sequential Patterns

Finding patterns in long event sequences is an important data mining task. Two decades ago research focused on finding all frequent patterns, where the anti-monotonic property of support was used to design efficient algorithms. Recent research focuses on producing a smaller output containing only the most interesting patterns. To achieve this goal, we introduce a new interestingness measure by computing the proportion of the occurrences of a pattern that are cohesive. This measure is robust to outliers, and is applicable to sequential patterns. We implement an efficient algorithm based on constrained prefix-projected pattern growth and pruning based on an upper bound to uncover the set of top-k quantile-based cohesive sequential patterns. We run experiments to compare our method with existing state-of-the-art methods for sequential pattern mining and show that our algorithm is efficient and produces qualitatively interesting patterns on large event sequences.

Len Feremans
University of Antwerp
UA
len.feremans@uantwerpen.be

Boris Cule, Bart Goethals
University of Antwerp
boris.cule@uantwerpen.be, bart.goethals@uantwerpen.be

CP2

Outlier Detection over Distributed Trajectory Streams

The wide deployments of GPS-embedded devices have produced multiple rapid voluminous trajectory streams, which needs to be analyzed to extract abnormal behaviors of moving objects in real-time. To date, outlier detection over distributed trajectory streams has not received enough focuses due to the constraint factors like skewness distribution and evolving nature of trajectory data, and on-the-fly execution requirement with minimal communication cost. In this paper, we present the first scalable decentralized outlier detection framework over distributed trajectory streams, called *ODDTS*. It consists of remote site processing and coordinator processing, with the aim of continuously providing *feature-grouping based* outliers detection over distributed trajectory streams. Extensive experiments over real data demonstrate high detecting validity, less communication cost and linear scalability of *ODDTS* method for online identifying outliers upon distributed trajectory streams.

Jiali Mao, Pengda Sun, Cheqing Jin, Aoying Zhou
East China Normal University
jlmao1231@stu.ecnu.edu.cn,
51174500124@stu.ecnu.edu.cn, cqjin@sei.ecnu.edu.cn,
ayzhou@dase.ecnu.edu.cn

CP2

Staple: Spatio-Temporal Precursor Learning for Event Forecasting

Large-scale societal events such as civil unrest movements occur due to a variety of factors including economics, politics, and security. Societal event detection can be modeled as a system of inter-connected locations, where each location is recording a set of time-dependent observations. In order to detect event occurrence and automatically reconstruct the precursors and signals, it is essential to model relationships between the different locations w.r.t. how events evolve over time. However, existing methods for precursor discovery do not capture or exploit spatial and temporal correlations inherent in event occurrences. The absence of such modeling not only creates shortcomings in the quality of inference but also curtails interpretation by human analysts. Furthermore, forecasting is inhibited when training data is sparse. In this paper, we develop a novel multi-task model with dynamic graph constraints within a multi-instance learning framework. Our model tackles the problem of scarce data distribution and reinforces co-occurring location-specific precursors with augmented representations. Through studies on civil unrest movements in numerous countries, we demonstrate the effectiveness of the proposed method for precursor discovery and event forecasting.

Yue Ning
Virginia Tech
yning@vt.edu

CP2

Streaming Tensor Factorization for Infinite Data Sources

Sparse tensor factorization is a popular tool in multi-way data analysis and is used in applications such as cybersecurity, recommender systems, and social network analysis. In many of these applications, the tensor is not known a priori and instead arrives in a streaming fashion for a potentially unbounded amount of time. Existing approaches for streaming sparse tensors are not practical for unbounded streaming because they rely on maintaining the full factorization of the data, which grows linearly with time. In this work, we present CP-stream, an algorithm for streaming factorization in the model of the canonical polyadic decomposition which does not grow linearly in time or space, and is thus practical for long-term streaming. Additionally, CP-stream incorporates user-specified constraints such as non-negativity which aid in the stability and interpretability of the factorization. An evaluation of CP-stream demonstrates that it converges faster than state-of-the-art streaming algorithms while achieving lower reconstruction error by an order of magnitude. We also evaluate it on real-world sparse datasets and demonstrate its usability in both network traffic analysis and discussion tracking. Our evaluation uses exclusively public datasets and our source code is released to the public as part of SPLATT, an open source high-performance tensor factorization toolkit.

Shaden Smith, Kejun Huang
University of Minnesota
shaden@cs.umn.edu, huang663@umn.edu

Nicholas Sidiropoulos
University of Virginia
nikos@virginia.edu

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

CP3

Robust Road Map Inference Through Network Alignment of Trajectories

In this presentation, we address the challenge of inferring the road network of a city from crowd-sourced GPS traces. While the problem has been addressed before, our solution has the following unique characteristics: (i) we formulate the road network inference problem as a network alignment optimization problem where both the nodes and edges of the network have to be inferred, (ii) we propose both an offline (`Kharita`) and an online (`Kharita`*) algorithms which are intuitive and capture the key aspects of the optimization formulation but are scalable and accurate. `Kharita`* in particular is, to the best of our knowledge, the first known online algorithm for map inference, (iii) we test our approach on two real data sets from the two cities of Doha (Qatar) and Chicago (US) and show the superiority of our formulation compared to state of the art methods for road map inference. Both our code and data sets have been made available for research reproducibility.

Sanjay Chawla
School of IT, the University of Sydney
chawla@it.usyd.edu.au

Sofiane Abbar, Rade Stanojevic, Saravanan Thirumuruganathan
Qatar Computing Research Institute
Hamad Bin Khalifa University
sabbar@hbku.edu.qa, rstanojevic@hbku.edu.qa,
sthirumuruganathan@hbku.edu.qa

Fethi Filali, Ahid Aleimat
Qatar Mobility Innovations Center
filali@qmic.com, ahide@qmic.com

CP3

Near-Optimal Mapping of Network States Using Probes

In many applications, such as the Internet and infrastructure networks, nodes fail or get congested dynamically. We study the problem of inferring all the failed nodes, when only a sample of the failures is known, and there exist correlations between node failures/congestion in networks. We formalize this as the GraphStateInf problem, using the Minimum Description Length (MDL) principle. We propose the GraphMap algorithm for minimizing the MDL cost, and show that it gives an additive approximation, relative to the optimal. We evaluate our methods on synthetic and real datasets, which includes one from WAZE which gives traffic incident reports for the city of Boston. We find that our method gives promising results in recovering the missing failures.

Bijaya Adhikari
Virginia Tech
bijaya@cs.vt.edu

Pavan Rangudu
NDSSL, Biocomplexity Institute, Virginia Tech
rangudu@vt.edu

B. Aditya Prakash
Virginia Tech
badityap@cs.vt.edu

Anil Vullikanti
NDSSL, Biocomplexity Institute, Virginia Tech
Department of Computer Science, Virginia Tech
vsakumar@vt.edu

CP3

Network Inference from Contrastive Groups Using Discriminative Structural Regularization

Gaussian graphical models (GGMs) are a popular tool for exploring conditional dependence among high dimensional data. We consider developing an estimator for GGMs for multiple graph analysis, wherein the graphs are assumed to come from two (or more) contrastive groups, and exhibit not only major global similarity, but also substantial between-group disparity. Under this setting, inferring each group of networks separately ignores the common structure, while simply assuming a global common network structure would mask the critical disparity. We propose a novel approach to pursue simultaneous network inference using discriminative and adaptive structural regularizations. We introduce a heterogeneity ratio parameter to balance the within group similarity and the between group disparity. This formulation for the first time, to our knowledge, generalizes the existing single-group network analysis to multiple-group analysis. By updating a global regularization template, together with a feature screening module specifying relevant dimensions to satisfy the group-level constraints, our approach can recover the conditional independence with greater flexibility and improved accuracy. Theoretically, we show the asymptotic consistency for the proposed method in joint reconstruction of multiple network structures. We demonstrate its superior performance via extensive simulation studies and its practical application to polychromatic flow cytometry data sets for protein interactions.

Ruihua Cheng
New Jersey Institute of Technology, USA
rc298@njit.edu

Zhi Wei
New Jersey Institute of Technology
zhiwei@njit.edu

Kai Zhang
Temple University, USA
kzhang980@gmail.com

CP3

Semi-Supervised Embedding in Attributed Networks with Outliers

In this paper, we propose a novel framework, called Semi-supervised Embedding in Attributed Networks with Outliers (SEANO), to learn a low-dimensional vector representation that systematically captures the topological proximity, attribute affinity and label similarity of vertices in a *partially labeled attributed network* (PLAN). Our method is designed to work in both transductive and inductive settings while explicitly alleviating noise effects from outliers. Experimental results on various datasets drawn from the web, text and image domains demonstrate the advantages of SEANO over the state-of-the-art methods in semi-supervised classification under transductive as well as inductive settings. We also show that a subset of parameters in SEANO are interpretable as outlier scores and can significantly outperform baseline methods when applied for detecting network outliers. Finally, we present the use of SEANO in a challenging real-world setting – flood mapping of satellite images and show that it is able to outperform modern remote sensing algorithms for this task.

Jiongqian Liang, Peter Jacobs, Jiankai Sun, Srinivasan Parthasarathy
The Ohio State University
AlbertLeungPRC@gmail.com, jacobs.269@buckeyemail.osu.edu, sjk2412@gmail.com, srini@cse.ohio-state.edu

CP3

Group Centrality Maximization Via Network Design

Network centrality plays an important role in many applications. Central nodes in social networks can be influential, driving opinions and spreading news or rumors. In hyperlinked environments, such as the Web, where users navigate via clicks, central content receives high traffic, becoming target for advertising campaigns. While there is an extensive amount of work on centrality measures and their efficient computation, controlling nodes' centrality via network updates is a more recent and challenging task. Performing minimal modifications to a network to achieve a desired property falls under the umbrella of network design problems. This paper is focused on improving group (coverage and betweenness) centrality, which is a function of the shortest paths passing through a set of nodes, by adding edges to the network. Several variations of the problem, which are NP-hard as well as APX-hard, are introduced. We present a greedy algorithm, and even faster sampling algorithms, for group centrality maximization with theoretical quality guarantees under realistic constraints. The experimental results show that our sampling algorithms outperform the best baseline solution in terms of centrality by up to 5 times while being 2-3 orders of magnitude faster than our greedy approach.

Sourav Medya, Arlei Silva, Ambuj Singh
UCSB
medya@cs.ucsb.edu, arlei@cs.ucsb.edu,
ambuj@cs.ucsb.edu

Prithwish Basu
Raytheon BBN Technologies
prithwish.basu@raytheon.com

Ananthram Swami
Army Research Laboratory
ananthram.swami.civ@mail.mil

CP3

AspEm: Embedding Learning by Aspects in Heterogeneous Information Networks

Heterogeneous information networks (HINs) are ubiquitous in real-world applications. Due to the heterogeneity in HINs, the typed edges may not fully align with each other. In order to capture the semantic subtlety, we propose the concept of aspects with each aspect being a unit representing one underlying semantic facet. Meanwhile, network embedding has emerged as a powerful method for learning network representation, where the learned embedding can be used as features in various downstream applications. Therefore, we are motivated to propose a novel embedding learning framework—AspEm—to preserve the semantic information in HINs based on multiple aspects. Instead of preserving information of the network in one semantic space, AspEm encapsulates information regarding each aspect individually. In order to select aspects for embedding purpose, we further devise a solution for AspEm based on dataset-wide statistics. To corroborate the efficacy of AspEm, we conducted experiments on two real-words datasets with two types of applications—classification and link prediction. Experiment results demonstrate that AspEm can outperform baseline network embedding learning methods by considering multiple aspects, where the aspects can be selected from the given HIN in an unsupervised manner.

Yu Shi
University of Illinois at Urbana-Champaign
yushi2@illinois.edu

Huan Gui
Facebook Inc.
huangui@fb.com

Qi Zhu
University of Illinois at Urbana-Champaign
qiz3@illinois.edu

Lance Kaplan
U.S. Army Lab
lance.m.kaplan.civ@mail.mil

Jiawei Han
UIUC
hanj@illinois.edu

CP4

Click Versus Share: A Feature-Driven Study of Micro-Video Popularity and Virality in Social Media

Micro-video has recently become an important form of user generated contents in the social media of microblogging. It is propagated by sharing and reaches the other users through being clicked and watched. Besides the traditional popularity metric for a micro-video such as click (or view) count, share count can indicate its virality in social domain. Understanding the differences between clicking and sharing behaviors is fundamental when evaluating the actual influence of micro-videos in social media. Thanks to a massive set of anonymized data from a major operator covering the whole China, we jointly study both clicking and sharing behaviors of over 10,000 micro-videos in Sina Weibo. Having extracted a rich set of features covering micro-video publishers, description texts and those shared users, we are able to identify the most influential features for click and share. From our studies, we observe that publisher-related features (*post* and *followee* counts) as well as the video *duration* have more impact on click, while video-description-related features including topical features and *emoticon* count are more correlated to share. Impacted by different features, the received clicks and shares of a micro-video may differ a lot from each other. Based on above observations, we build a prediction model for existing deviations among these two metrics, which can aid the development of a more effective and attractive micro-video platform.

Jingtao Ding, Yanghao Li, Yong Li, Depeng Jin
Tsinghua University
dingjt15@mails.tsinghua.edu.cn, liyanghao14@mails.tsinghua.edu.cn, liyong07@tsinghua.edu.cn, jindp@tsinghua.edu.cn

CP4

Modeling the Interaction Coupling of Multi-View Spatiotemporal Contexts for Destination Prediction

Bike-Sharing Systems (BSSs) are being introduced to more and more cities recently, and therefore they have generated huge amounts of data. Mobike is a station- less BSS which is suffering from the chaotic parking problem. To solve this problem, it is necessary to predict where the bikes are going. Traditional works deal- ing with destination prediction mainly focus on station- based BSSs, and they merely leverages context-aware information technically. Thus it is naturally promising to investigate how to improve the destination prediction of station-less bikes by context information. To that end, in this paper, we develop a multi-view ma- chine (MVM) method, by incorporating the context information from Point of Interest (POI) data and human mobility data into destination prediction. Specifically, we first describe three different views, namely start position, start time and destination by features extracted from POI data and human mobility data. Then, we capture the relationship between these three views interactions and the trips possibility by a multi-view ma- chine. Finally, since multi-view machine contains too many parameters to be optimized, we leverage tensor factorization (TF) to reduce the quantity of calculation. The experiment results show that the model can effectively capture the potential relationship of three views with trips possibility and the approach is thus much more effective than traditional prediction methods for destination.

Kunpeng Liu, Pengyang Wang
Missouri University of Science and Technology
kl6ph@mst.edu, pwqt3@mst.edu

Jiawei Zhang
Florida State University, USA
jzhang@cs.fsu.edu;

Guannan Liu
BeiHang University, China
liugn@buaa.edu.cn

Yanjie Fu
Missouri University of Science and Technology, USA

fuyan@mst.edu

Sajal Das
Missouri Univeristy of Science and Technology
sdas@mst.edu

CP4

A Probabilistic Hough Transform for Opportunistic Crowd-Sensing of Moving Traffic Obstacles

Traffic congestion in developing cities like Nairobi, Kenya can be significantly impacted by the presence of Moving Traffic Obstacles (MTOs). These MTOs are events that temporarily exist on the road, moving with or against the direction of traffic at slower speeds. They include two-wheelers, pushcarts, animals, and pedestrians, which have quite different influence on traffic compared with static obstacles, such as potholes and speed bumps. As smartphones and supporting 3G infrastructures are wide spread even in developing countries, recent studies enabled frugal traffic obstacle data collection from smartphones in probe cars. Assuming the opportunistic, unevenly-distributed, sparse and errorful observation of traffic obstacles, we propose an MTO detection algorithm extending an image analysis technique called Probabilistic Hough Transform for collective observations as input. Based on our experiences with a small set of real-world data collected in a smartphone-based probe car project with Nairobi City County, we conducted experiments with simulated observation data to see the effectiveness of the algorithm.

Michiaki Tatsubori
IBM Research AI, Tokyo
mich@jp.ibm.com

Aisha Walcott-Bryant, Reginald Bryant
IBM Research - Kenya
awalcott@ke.ibm.com, bryantre@ke.ibm.com

John Wamburu
University of Massachusetts, Amherst
jwamburu@cs.umass.edu

CP4

You Are How You Move: Linking Multiple User Identities From Massive Mobility Traces

Understanding the linkability of online user identifiers (IDs) is critical to both service providers (for business intelligence) and individual users (for assessing privacy risks). Existing methods are designed to match IDs across *two services*, but face key challenges of matching multiple services in practice, particularly when users have multiple IDs per service. In this paper, we propose a novel system to link IDs across multiple services by exploring the spatial-temporal locality of user activities. The core idea is that the same user's online IDs are more likely to repeatedly appear at the same location. Specifically, we first utilize a *contact graph* to capture the "co-location' of all IDs across multiple services. Based on this graph, we propose a set-wise matching algorithm to discover candidate ID sets, and use Bayesian inference to generate confidence scores for candidate ranking, which is proved to be optimal. We evaluate our system using two real-world ground-truth datasets from an ISP (4 services, 815K IDs) and Twitter-Foursquare (2 services, 770 IDs). Extensive results show that our system significantly outperforms the state-of-the-art algorithms in accuracy (AUC is higher by 0.1-0.2), and it is highly robust against matching order and number of services.

Huandong Wang
Tsinghua University
Tsinghua University
whd14@mails.tsinghua.edu.cn

Yong Li
Tsinghua University
liyong07@tsinghua.edu.cn

Gang Wang
Virginia Tech.
gangwang@vt.edu

Depeng Jin
Tsinghua University
jindp@tsinghua.edu.cn

CP4

Who Will Attend This Event Together? Event Attendance Prediction Via Deep Lstm Networks

Event-based social network (EBSN) services have emerged as a new platform on which users can choose events of interest to attend in the physical world. Over years, there are growing research interests in predicting whether certain actors will participate in an event together. In this work, we refer to this task as the event attendance prediction problem and aim to address the predictability of individuals' event attendance. In real-world settings, the factors that influence an individual's attendance may change over time, leading to the dynamic nature of individuals' behavior. However, existing event attendance prediction methods cannot deal with such dynamic scenarios. To address this issue, we propose an end-to-end Deep Event Attendance Prediction (DEAP) framework—a three-level hierarchical LSTM architecture—to explicitly model users' multi-dimensional and evolving preferences. Extensive experiments on three real-world datasets demonstrate that DEAP significantly outperforms the state-of-the-art techniques across various settings.

Xian Wu
university of notre dame
xwu9@nd.edu

Yuxiao Dong, Baoxu SHI
University of Notre Dame
ydong1@nd.edu, bshi@nd.edu

Ananthram Swami
Army Research Laboratory
ananthram.swami.civ@mail.mil

Nitesh Chawla
University of Notre Dame
nchawla@nd.edu

CP4

Online Truth Discovery on Time Series Data

Truth discovery, with the goal of inferring true information from massive data through aggregating the information from multiple data sources, has attracted significant attention in recent years. It has demonstrated great advantages in real applications since it can automatically learn the reliability degrees of the data sources without supervision and in turn helps to find more reliable information. In many applications, however, the data may arrive in a stream and present various temporal patterns. Unfortunately, there is no existing truth discovery work that can handle such time series data. To tackle this challenge, we propose a novel online truth discovery framework that incorporates the predictions on the time series data into the truth estimation process. By jointly considering the multisource information and the temporal patterns of the time series data, the proposed framework can improve the accuracy of the truth discovery results as well as the time series prediction. The effectiveness of the proposed framework is validated on both synthetic and real-world datasets.

Liuyi Yao, Lu Su
SUNY Buffalo
liuyiyao@buffalo.edu, lusu@buffalo.edu

Qi Li
University of Illinois at Urbana-Champaign
qili5@illinois.edu

Yaliang Li
Baidu Research Big Data Lab
yaliangli@baidu.com

Fenglong Ma
SUNY Buffalo
fenglong@buffalo.edu

Jing Gao
University at Buffalo
jing@buffalo.edu

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

CP5

Interpretable Categorization of Heterogeneous Time Series Data

Understanding heterogeneous multivariate time series data is important in many applications ranging from smart homes to aviation. Learning models of heterogeneous multivariate time series that are also human-interpretable is challenging and not adequately addressed by the existing literature. We propose grammar-based decision trees (GBDTs) and an algorithm for learning them. GBDTs extend decision trees with a grammar framework. Logical expressions derived from a context-free grammar are used for branching in place of simple thresholds on attributes. The added expressivity enables support for a wide range of data types while retaining the interpretability of decision trees. In particular, when a grammar based on temporal logic is used, we show that GBDTs can be used for the interpretable classification of high-dimensional and heterogeneous time series data. Furthermore, we show how GBDTs can also be used for categorization, which is a combination of clustering and generating interpretable explanations for each cluster. We apply GBDTs to analyze the classic Australian Sign Language dataset as well as data on near mid-air collisions (NMACs). The NMAC data comes from aircraft simulations used in the development of the next-generation Airborne Collision Avoidance System (ACAS X).

Ritchie Lee
Carnegie Mellon University Silicon Valley
ritchie.lee@sv.cmu.edu

Mykel Kochenderfer
Stanford University
mykel@stanford.edu

Ole Mengshoel
Carnegie Mellon University Silicon Valley
ole.mengshoel@sv.cmu.edu

Joshua Silbermann
Johns Hopkins University Applied Physics Laboratory
joshua.silbermann@jhuapl.edu

CP5

Evolving Separating References for Time Series Classification

The mining of time series data has attracted much attention in the past two decades due to the ubiquity of time series in our daily lives. In particular, classification is perhaps one of the most well-studied topics for time series data. Many state-of-the-art classification techniques work by identifying and extracting patterns or characteristics from the training data, and then applying these patterns or characteristics to classify unlabeled time series. This talk presents a novel finding that sequences of values that are very different from the patterns in the labeled time series can be used as references to classify time series effectively. We propose an evolution process to generate these sequences of values, which we call separating references, from the training data. The proposed method is robust to over-fitting and is especially suitable for the situation where little labeled data is available. We demonstrate that the proposed approach is highly competitive on the well-known UCR time series classification benchmarks.

Xiaosheng Li, Jessica Lin
George Mason University
xli22@gmu.edu, jessica@gmu.edu

CP5

Classifying Multivariate Time Series by Learning Sequence-Level Discriminative Patterns

Time series classification algorithms designed to use local context do not work on landcover classification problems where the instances of the two classes may often exhibit similar feature values due to the large natural variations in other land covers across the year and unrelated phenomena that they undergo. In this paper, we propose to learn discriminative patterns from the entire length of the time series, and use them as predictive features to identify the class of interest. We propose a novel neural network algorithm to learn the key signature of the class of interest as a function of the feature values together with the discriminative pattern made from that signature through the entire time series in a joint framework. We demonstrate the utility of this technique on the landcover classification application of burned area mapping that is of considerable societal importance.

Guruprasad Nayak
University of Minnesota
nayak013@umn.edu

Varun Mithal
LinkedIn
vamithal@linkedin.com

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

Vipin Kumar
University of Minnesota
kumar001@umn.edu

CP5

Accelerating Time Series Searching with Large Uniform Scaling

Similarity search is arguably the most important primitive in time series data mining. It is useful in its own right as an exploratory tool, and a subroutine in almost all higher level algorithms. Because of this, and the prevalence of time series data, the last decade has seen fast algorithms for time series similarity search under Dynamic Time Warping (DTW) and Uniform Scaling (US) distance measures. However, current state-of-the-art algorithms for US have only been demonstrated for the modest amounts of rescaling in datasets produced by human behaviors and physiological measurements. As we shall show, in many industrial and commercial contexts we may encounter much greater amounts of rescaling, rendering current solutions little better than brute force search. To mitigate this problem we introduce novel lower bounds, LBnew, which, for the first time allows efficient search even in domains that exhibit more than a factor-of-two variability in scale. We demonstrate the utility of our ideas with both theoretical guarantees and comprehensive experiments on real data from commercial important domains, including power consumption monitoring and ECG monitoring. The results show the application of our lower bounds significantly outperforms state-of-the-art approaches for accelerating similarity searching of time series with more than a factor-of-two variability in scale as well as high level time series mining tasks.

Yilin Shen, Yanping Chen
Samsung Research America
yilin.shen@samsung.com, yanping.c@samsung.com

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

Hongxia Jin
Samsung Research America
hongxia.jin@samsung.com

CP5

Efficient Search of the Best Warping Window for Dynamic Time Warping

Time series classification maps time series to labels. The nearest neighbor algorithm (NN) using the Dynamic Time Warping (DTW) similarity measure is a leading algorithm for this task and a component of the current best ensemble classifiers for time series. However, NN-DTW is only a winning combination when its meta-parameter – its warping window – is learned from the training data. The warping window (WW) intuitively controls the amount of distortion allowed when comparing a pair of time series. With a training database of N time series of lengths L, a naive approach to learning the WW requires $\Theta(N^2 \cdot L^3)$ operations. This often results in NN-DTW requiring days for training on datasets containing a few thousand time series only. In this paper, we introduce FastWWSearch an *efficient* and *exact* method to learn WW. We show on 86 datasets that our method is always faster than the state of the art, with at least one order of magnitude and up to 1000x speed-up.

Chang Wei Tan, Matthieu Herrmann
Faculty of Information Technology, Monash University
chang.tan@monash.edu, matthieu.herrmann@monash.edu

Germain Forestier
MIPS, University of Haute Alsace
germain.forestier@uha.fr

Geoff Webb
Monash University
geoff.webb@monash.edu

Francois Petitjean
Faculty of Information Technology
Monash University
francois.petitjean@monash.edu

CP6

Eeg-Based Motion Intention Recognition Via Multi-Task Rnns

Recognition of human intention based on Electroencephalography (EEG) signals attracts strong research interest in pattern recognition because of its promising applications that enable non-muscular communications and controls. Over the past few years, most EEG-based recognition works make significant efforts to learn extracted features to explore specific patterns between a segment of EEG signals and the corresponding activities. Unfortunately, vectorization-based feature representations, either vector-like or matrix-like ones, suffer from massive signal noise and difficulties of exploiting signal correlations between adjacent sensors of EEG signals. Most importantly, EEG signals are represented by one unique frequency and then fed into the subsequent learning model. Neglecting different frequencies of EEG signals can be detrimental to activity recognition because a particular frequency of EEG signals is more helpful to recognize some activities. Inspired by this idea, we propose to extract EEG signals with different frequencies and introduce a novel Multi-task deep learning model to learn the human intentions. We have conducted extensive experiments on a publicly available EEG benchmark dataset and compared our method with many state-of-the-art algorithms. The experimental results demonstrate that the proposed Multi-task deep recurrent neural network outperforms all the compared methods in a multi-class scenario.

Weitong Chen
The University of Queensland
uqwche12@uq.edu.au

Sen Wang
Griffith University
sen.wang@griffith.edu.au

Xiang Zhang
The University of New South Wales, Australia
xiang.zhang3@student.unsw.edu.au

Lina Yao
The University of New South Wales
lina.yao@unsw.edu.au

Lin Yue
Northeast Normal University
yuel563@163.com

Buyue Qian
Xian Jiaotong University
qianbuyue@xjtu.edu.cn

Xue Li
School of Information Technology and Electrical Engineering,
The University of Queensland, Brisbane, Queensland 4072
xueli@itee.uq.edu.au

CP6

Deep Attention Model for Triage of Emergency Department Patients

Optimization of patient throughput and wait time in emergency departments (ED) is an important task for hospital systems. For that reason, Emergency Severity Index (ESI) system for patient triage was introduced to help guide manual estimation of acuity levels, which is used by nurses to rank the patients and organize hospital resources. However, despite improvements that it brought to managing medical resources, such triage system greatly depends on nurse's subjective judgment and is thus prone to human errors. Here, we propose a novel deep model based on the word attention mechanism designed for predicting a number of resources an ED patient would need. Our approach incorporates routinely available continuous and nominal data with medical text data, including patient's chief complaint, past medical history, medication list, and nurse assessment collected for 338,500 ED visits over three years in a large urban hospital. Using both structured and unstructured data, the proposed approach achieves the AUC of 88% for the task of identifying resource intensive patients, and the accuracy of 44% for predicting exact category of number of resources, giving an estimated lift over nurses' performance by 16% in accuracy. Furthermore, the attention mechanism of the proposed model provides interpretability by assigning attention scores for nurses' notes which is crucial for decision making and implementation of such approaches in the real systems working on human health.

Djordje Gligorijevic, Jelena Stojanovic
Temple University
gligorijevic@temple.edu, jelena.stojanovic@temple.edu

Wayne Satz, Ivan Stojkovic, Kraftin Schreyer, Daniel Del Portal
Temple University, USA
wayne.satz@tuhs.temple.edu, ivan.stojkovic@temple.edu,
kraftin.schreyer@tuhs.temple.edu,
daniel.delportal@tuhs.temple.edu

Zoran Obradovic
Temple University
zoran@ist.temple.edu

CP6

Uncorrelated Patient Similarity Learning

Patient similarity learning aims to derive a clinically meaningful similarity metric to measure the similarity between a pair of patients according to their historical clinical information, which could help to predict the clinical outcomes of the patient of interest. However, the patient clinical data are usually complex, and contain much irrelevant and redundant information, which makes it difficult to learn the similarity metric with high accuracy. Although some methods have been proposed to address the complex nature of patient data, they overemphasize sparsity-based relevant feature selection and fail to take into consideration the redundant features that are highly correlated with each other, and this heavily degrades the accuracy of the learned results. To address the above challenges, we propose a novel uncorrelated patient similarity learning approach, which can not only select the most relevant features for the learning task, but also guarantee that the selected features have low correlations with each other. Additionally, to address the scenarios where the patient data are distributed across different sites, we extend the proposed approach and design a distributed mechanism, based on which the similarity metric can be accurately learned without directly accessing the raw patient data at each site. The desirable performance of the proposed methods are verified through extensive experiments conducted on both real-world and synthetic datasets.

Mengdi Huai, Chenglin Miao, Qiuling Suo
State University of New York at Buffalo
mengdihu@buffalo.edu, cmiao@buffalo.edu,
qiulings@buffalo.edu

Yaliang Li
Baidu Research Big Data Lab
yaliangli@baidu.com

Jing Gao
University at Buffalo
jing@buffalo.edu

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

CP6

Multi-Task Learning Based Survival Analysis for Predicting Alzheimer's Disease Progression with Multi-Source Block-Wise Missing Data

There are many major diseases that remain incurable, e.g., Alzheimer's disease (AD). Hence, the prevention for these diseases has more impact than diagnosis and treatment. Survival analysis aims at predicting the time of occurrence of specific events of interest. It can be used to identify patients of high risk, which helps healthcare system to effectively allocate limited medical resources. In many real-world applications, such as healthcare analysis, a lot of datasets are collected from multiple data sources and exhibit a block-wise missing pattern, i.e., each patient takes different types of tests and receives various treatments, and each test/treatment associates with a corresponding set of features. However, all the existing survival analy-

sis methods are designed for fully observed datasets. The proposed work in this paper aims at addressing aforementioned research challenges. Specifically, we employ a partition method that decomposes the multi-source block-wise missing data into the multiple completed sub-matrix; thus, transforms the original problem into a series of related multi-source survival analysis problems. To deal with these problems, we propose a two-layer multi-task learning model that achieves both feature-level and source-level analysis. We apply the proposed method in a real-world AD dataset to study the stage conversion of AD patients. Our experimental results show that the proposed method outperforms the other state-of-the-art methods.

Yan Li
University of Michigan
yanliwl@umich.edu

Tao Yang
Arizona State University
t.yang@asu.edu

Jiayu Zhou
MSU
jiayuz@msu.edu

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

CP6

Health-Atm: A Deep Architecture for Multifaceted Patient Health Record Representation and Risk Prediction

Leveraging massive electronic health records (EHR) brings tremendous promises to advance clinical and precision medicine informatics research. However, it is very challenging to directly work with multifaceted patient information encoded in their EHR data. Deriving effective representations of patient EHRs is a crucial step to bridge raw EHR information and the endpoint analytical tasks, such as risk prediction or disease subtyping. In this paper, we propose Health-ATM, a novel and integrated deep architecture to uncover patients' comprehensive health information from their noisy, longitudinal, heterogeneous and irregular EHR data. Health-ATM extracts comprehensive multifaceted patient information patterns with attentive and time-aware modulars (ATM) and a hybrid network structure composed of both Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN). The learned features are finally fed into a prediction layer to conduct the risk prediction task. We evaluated the Health-ATM on both artificial and real world EHR corpus and demonstrated its promising utility and efficacy on representation learning and disease onset predictions.

Tengfei Ma
IBM T.J. Watson Research Center
Tengfei.Ma1@ibm.com

Cao Xiao
IBM Research
cxiao@us.ibm.com

Fei Wang
Cornell University
few2001@med.cornell.edu

CP7

On2Vec: Embedding-Based Relation Prediction for Ontology Population

Populating ontology graphs represents a long-standing problem for the Semantic Web community. Recent advances in translation-based graph embedding methods for populating instance-level knowledge graphs lead to promising new approaching for the ontology population problem. However, unlike instance-level graphs, the majority of relation facts in ontology graphs come with comprehensive semantic relations, which often include the properties of transitivity and symmetry, as well as hierarchical relations. These comprehensive relations are often too complex for existing graph embedding methods, and direct application of such methods is not feasible. Hence, we propose On2Vec, a novel translation-based graph embedding method for ontology population. On2Vec integrates two model components that effectively characterize comprehensive relation facts in ontology graphs. The first is the Component-specific Model that encodes concepts and relations into low-dimensional embedding spaces without a loss of relational properties; the second is the Hierarchy Model that performs focused learning of hierarchical relation facts. Experiments on several well-known ontology graphs demonstrate the promising capabilities of On2Vec in predicting and verifying new relation facts. These promising results also make possible significant improvements in related methods.

Muhao Chen
University of California Los Angeles
muhaochen@ucla.edu

CP7

Ensemble-Spotting: Ranking Urban Vibrancy Via Poi Embedding with Multi-View Spatial Graphs

Vibrant residential communities are de?ned as places with permeability, vitality, variety, accessibility, identity and legibility. Developing vibrant communities can help boost commercial activities, enhance public security, foster social interaction, and thus yield livable, sustainable, and viable environments. However, it is challenging to understand the underlying drivers of vibrant communities to make them traceable and predictable. Toward this goal, we study the problem of ranking vibrant communities using human mobility data and point-of-interests (POIs) data. We analyze large-scale urban and mobile data related to residential communities and ?nd that in order to e?ectively identify vibrant communities, we should not just consider community contents such as buildings, facilities, and transportation, but also take into account the spatial structure. The spatial structure of a community refers to how the geographical items (POIs, road networks, public transits, etc.) of a community are spatially arranged and interact with one another. Along this line, we ?rst develop a geographical learning method to ?nd proper representations of communities. In addition, we propose a novel geographic ensemble ranking strategy, which aggregates a variety of weak rankers to e?ectively spot vibrant communities. Finally, we conduct a comprehensive evaluation with real-world residential community data. The experimental results demonstrate the e?ectiveness of the proposed method.

Pengyang Wang
Missouri University of Science and Technology

pwqt3@mst.edu

Jiawei Zhang
Florida State University, USA
jzhang@cs.fsu.edu;

Guannan Liu
BeiHang University, China
liugn@buaa.edu.cn

Yanjie Fu
Missouri University of Science and Technology, USA
fuyan@mst.edu

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

CP7

A Family of Tractable Graph Distances

Important data mining problems such as nearest-neighbor search and clustering admit theoretical guarantees when restricted to objects embedded in a metric space. Graphs are ubiquitous, and clustering and classification over graphs arise in diverse areas, including, e.g., image processing and social networks. Unfortunately, popular distance scores used in these applications, that scale over large graphs, are not metrics and thus come with no guarantees. Classic graph distances such as, e.g., the chemical and the CKS distance are arguably natural and intuitive, and are indeed also metrics, but they are intractable: as such, their computation does not scale to large graphs. We define a broad family of graph distances, that includes both the chemical and the CKS distance, and prove that these are all metrics. Crucially, we show that our family includes metrics that are tractable. Moreover, we extend these distances by incorporating auxiliary node attributes, which is important in practice, while maintaining both the metric property and tractability.

Stratis Ioannidis
Northeastern University
ioannidis@ece.neu.edu

Jose Bento
Boston College
jose.bento@bc.edu

CP7

On Spectral Graph Embedding: A Non-Backtracking Perspective and Graph Approximation

In this paper, we present a novel spectral graph embedding framework called NOn-Backtracking Embedding (NOBE), which offers a new perspective that organizes graph data at a deep level by tracking the flow traversing on the edges with backtracking prohibited. Further, by analyzing the non-backtracking process, a technique called graph approximation is devised, which provides a channel to transform the spectral decomposition on an edge-to-edge matrix to that on a node-to-node matrix. Theoretical guarantees are provided by bounding the difference between the corresponding eigenvalues of the original graph and its graph approximation.

Fei Jiang
Peking University
allen.feijiang@gmail.com

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

Yi Zheng
Peking Univerisity
zhengyi@pku.edu.cn

Enqiang Zhu
Guangzhou University
zhuenqiang@gzhu.edu.cn

Jin Xu
Peking University
jxu@pku.edu.cn

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

CP7

Learning Graph Representation Via Frequent Subgraphs

We propose a novel approach to learn distributed representation for graph data. Our idea is to combine a recently introduced neural document embedding model with a traditional pattern mining technique, by treating a graph as a document and frequent subgraphs as atomic units for the embedding process. Compared to the latest graph embedding methods, our proposed method offers three key advantages: fully unsupervised learning, entire-graph embedding, and edge label leveraging. We demonstrate our method on several datasets in comparison with a comprehensive list of up-to-date state-of-the-art baselines where we show its advantages for both classification and clustering tasks.

Dang Nguyen, Wei Luo, Tu Nguyen
Center for Pattern Recognition and Data Analytics
Deakin University, Geelong, Australia
ngdang@deakin.edu.au, wei.luo@deakin.edu.au,
tu.nguyen@deakin.edu.au

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

CP7

Fast Flow-Based Random Walk with Restart in a Multi-Query Setting

In this work, we focus on solving Random Walk with Restart fast and accurately under large queries. We introduce a new, intuitive two-step divide-and-conquer formulation and a parallelizable method, FlowR, for solving RWR with two goals: (i) fast and accurate computation under large queries; (ii) one-time message exchange between subproblems. We further speed up the method by extending our formulation to carefully designed overlapping subproblems (FlowR-OV) and by leveraging the strengths of

iterative methods (FlowR-Hyb).

Yujun Yan, Mark Heimann, Di Jin
University of Michigan
yujunyan@umich.edu, mheimann@umich.edu,
dijin@umich.edu

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

CP8
SamBaTen: Sampling-Based Batch Incremental Tensor Decomposition

Tensor decompositions are invaluable tools in analyzing multimodal datasets. In many real-world scenarios,such datasets are far from being static, to the contrary they tend to grow over time. For instance, in an online social network setting, as we observe new interactions over time, our dataset gets updated in its “time” mode. How can we maintain a valid and accurate tensor decomposition of such a dynamically evolving multimodal dataset, without having to re-compute the entire decomposition after every single update? In this paper, we introduce SamBaTen, a Sampling-based Batch Incremental Tensor Decomposition algorithm, which incrementally maintains the decomposition given new updates to the tensor dataset. SamBaTen is able to scale to datasets that the state-of-the-art in incremental tensor decomposition is unable to operate on, due to its ability to effectively summarize the existing tensor and the incoming updates, and perform all computations in the reduced summary space. We extensively evaluate SamBaTen using synthetic and real datasets. Indicatively, SamBaTen achieves comparable accuracy to state-of-the-art incremental and non-incremental techniques, while being up to 25-30 times faster. Furthermore, SamBaTen scales to very large sparse and dense dynamically evolving tensors of dimensions up to 100K x 100K x 100K where state-of-the-art incremental approaches were not able to operate.

Ekta Gujral, Ravdeep Pasricha
University of California, Riverside
Computer Science and Engineering
egujr001@ucr.edu, rpasr001@ucr.edu

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

CP8
The Trustworthy Pal: Controlling the False Discovery Rate in Boolean Matrix Factorization

Boolean matrix factorization (BMF) is a popular and powerful technique for inferring knowledge from data. The mining result is the Boolean product of two matrices, approximating the input dataset. The Boolean product is a disjunction of rank-1 binary matrices, each describing a feature-relation, called pattern, for a group of samples. Yet, there are no guarantees that any of the returned patterns do not actually arise from noise, i.e., are false discoveries. In this paper, we propose and discuss the usage of the false discovery rate in the unsupervised BMF setting. We prove two bounds on the probability that a found pattern is constituted of random Bernoulli-distributed noise. Each bound exploits a specific property of the factorization which minimizes the approximation error—yielding new insights on the minimizers of Boolean matrix factorization. This leads to improved BMF algorithms by replacing heuristic rank selection techniques with a theoretically well-based approach. Our empirical demonstration shows that both bounds deliver excellent results in various practical settings.

Sibylle Hess, Nico Piatkowski, Katharina Morik
TU Dortmund
sibylle.hess@tu-dortmund.de, nico.piatkowski@tu-dortmund.de, katharina.morik@tu-dortmund.de

CP8
Latitude: A Model for Mixed LinearTropical Matrix Factorization

NMF is one of the most widespread models in data analysis and is known for its ‘parts of whole’ interpretation. Recently subtropical matrix factorization (SMF) was introduced, that has equally useful ‘winner takes it all’ interpretation. We propose a new mixed linear–tropical model that smoothly transitions between NMF and SMF, using latent parameters for controlling their mixture. We also present an algorithm for this model that reveals more latent structure than either NMF or SMF.

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

James Hook
University of Bath
j.l.hook@bath.ac.uk

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

CP8
Discovering Hidden Topical Hubs and Authorities in Online Social Networks

Finding influential users in online social networks is an important problem with many possible useful applications. HITS and other link analysis methods, in particular, have been often used to identify hub and authority users in web graphs and online social networks. These works, however, have not considered topical aspect of links in their analysis. A straightforward approach to overcome this limitation is to first apply topic models to learn the user topics before applying the HITS algorithm. In this paper, we instead propose a novel topic model known as Hub and Authority Topic (HAT) model to combines the two process so as to jointly learn the hub, authority and topical interests. We evaluate HAT against several existing state-of-the-art methods in two aspects: (i) modeling of topics, and (ii) link recommendation. We conduct experiments on two real-world datasets from Twitter and Instagram. Our experiment results show that HAT is comparable to state-of-the-art topic models in learning topics and it outperforms the state-of-the-art in link recommendation task.

Roy Ka-Wei Lee
Singapore Management University
Singapore Management University
roylee.2013@smu.edu.sg

Tuan-Anh Hoang

L3S Research Center
hoang@l3s.de

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

CP8

Topic Modeling Based on Keywords and Context

Current topic models often suffer from discovering topics not matching human intuition, unnatural switching of topics within documents and high computational demands. We address these shortcomings by proposing a topic model and an inference algorithm based on automatically identifying characteristic keywords for topics. Keywords influence the topic assignments of nearby words. Our algorithm learns (key)word-topic scores and self-regulates the number of topics. The inference is simple and easily parallelizable. A qualitative analysis yields comparable results to those of state-of-the-art models, but with different strengths and weaknesses. Quantitative analysis using eight datasets shows gains regarding classification accuracy, PMI score, computational performance, and consistency of topic assignments within documents, while most often using fewer topics.

Johannes Schneider
Zurich insurances
johannes.schneider@uni.li

CP8

ParaSketch: Parallel Tensor Factorization Via Sketching

Tensor factorization methods have gained increased popularity in the data mining community. A key feature that renders tensors attractive is the essential uniqueness (identifiability) of their decomposition into latent factors: this is crucial for *explanatory* data analysis – model uniqueness makes interpretations well grounded. In this work, we propose ParaSketch, a distributed tensor factorization algorithm that enables massive parallelism, to deal with large tensors. The idea is to compress/sketch the large tensor into multiple small tensors, decompose each small tensor, and combine the results to reconstruct the desired latent factors. Prior art in this direction entails potentially very high complexity in the (Gaussian) compression and final combining stages. Utilizing sketching matrices for compression, the proposed method greatly reduces compression complexity, and features much simpler combining. Moreover, theoretical analysis shows that the compressed tensors inherit latent identifiability under mild conditions, hence establishing correctness of the overall approach. Our approach to establish identifiability for the sketched tensor is original, and of interest in its own right.

Bo Yang, Ahmed Zamzam
University of Minnesota
yang4173@umn.edu, ahmedz@umn.edu

Nicholas Sidiropoulos
University of Virginia
nikos@virginia.edu

CP9

Markov Chain Monitoring

In networking applications, one often wishes to obtain estimates about the number of objects at different parts of the network (e.g., the number of cars at an intersection of a road network or the number of packages that are expected to reach a node in a computer network) by monitoring the traffic in a small number of network nodes or edges. To formalize the task of choosing what to monitor, we introduce the Markov Chain Monitoring problem. Given an initial distribution of items over the nodes of a Markov chain, we wish to estimate the distribution of items at subsequent times. In deriving these estimates, we issue queries to retrieve partial information on the distribution of items (e.g., ask how many items transitioned to a specific node or over a specific edge at a particular time). We consider different types of queries each defining a different variant of the problem. For each variant, we design efficient algorithms for picking the right queries that make our estimates as accurate as possible. In our experiments with synthetic and real datasets, we demonstrate the efficiency and the efficacy of our algorithms in a variety of settings.

Harshal Chaudhari
Boston University
harshal@bu.edu

Michael Mathioudakis
University of Helsinki
michael.mathioudakis@helsinki.fi

Evimaria Terzi
Boston University
evimaria@cs.bu.edu

CP9

Exploiting Structure for Fast Kernel Learning

We propose two methods for exact Gaussian process (GP) inference and learning on massive image, video, spatial-temporal, or multi-output datasets with missing values (or ”gaps”) in the observed responses. The first method ignores the gaps using sparse selection matrices and a highly effective low-rank preconditioner is introduced to accelerate computations. The second method introduces a novel approach to GP training whereby response values are inferred on the gaps before explicitly training the model. We find this second approach to be greatly advantageous for the class of problems considered. Both of these novel approaches make extensive use of Kronecker matrix algebra to design massively scalable algorithms which have low memory requirements. We demonstrate exact GP inference for a spatial-temporal climate modelling problem with 3.7 million training points as well as a video reconstruction problem with 1 billion points.

Trefor W. Evans, Prasanth B. Nair
University of Toronto
trefor.evans@mail.utoronto.ca, pbn@utias.utoronto.ca

CP9

Co-Regularized Monotone Retargeting for Semi-Supervised Letor

This work proposes a new model for listwise Learning to

Rank (LeTOR) in an inductive semisupervised setting. We pose the task as that of ranking in a multiview setting, encountered quite commonly in practice. We formulate a novel and efficient co-regularization mechanism that efficiently enforces agreement between views on the rank order of unlabeled samples. This formulation is based on leveraging the convex structures of isotonic vectors and to the best of our knowledge, the first of such coregularization based frameworks for semisupervised ranking. We demonstrate the utility of the method when labels are scarce even in settings where supervision is available only as pairwise preferences as well as in comparison to transductive semisupervised baselines.

Shalmali Joshi
The University of Texas at Austin
shalmali@utexas.edu

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

CP9

Global Nonlinear Metric Learning by Gluing Local Linear Metrics

We address the nonlinear metric learning by constructing a smooth nonlinear metric from the data. First, we locally define an initial linear metric on each cluster by principal component analysis. Second, we glue such local linear metrics to form a smooth nonlinear metric by a partition of unity on the sample space, and further learn the global nonlinear metric. Third, we conduct the intrinsic steepest descent algorithm on matrix manifolds for implementation. Finally, we compare our approach with several state-of-the-art methods on a variety of datasets. The results validate that the robustness and accuracy of classification are both improved under our nonlinear metric. The novelty of our global smooth nonlinear metric learning model lies in that it has completely overcome drawbacks of local metric learning methods: the partition coefficients obtained by the partition of unity is smooth, while the metric at any point on the manifold can be directly defined.

Yaxin Peng, Lingfang Hu, Shihui Ying
Shanghai University
yaxin.peng@shu.edu.cn, lingfanghu2016@gmail.com, shying@shu.edu.cn

Chaomin Shen
East China Normal University
cmshen@cs.ecnu.edu.cn

CP9

Multi-view Weak-label Learning Based on Matrix Completion

Weak-label learning is an important branch of multi-label learning; it deals with samples annotated with incomplete (weak) labels. Previous work on weak-label learning mainly considers data represented by a single view. An intuitive way to leverage multiple features obtained from different views is to concatenate the features into a single vector. However, this process is not only prone to overfitting and often results in very high time-complexity, but also ignores the potentially useful complementary information spread across the different views. In this paper, we propose an approach based on Matrix Completion for multi-view Weak-label Learning (McWL). Matrix completion (MC) has sound theoretical properties and is robust to missing values in both feature and label spaces. Our method enforces the optimization of multiple view integration and of MC-based classification within a unified objective function. Specifically, a kernel target alignment technique and the loss function of an MC-based classifier are used to jointly and iteratively adjust the weights assigned to individual views, and to optimize the classifier. McWL can selectively integrate views and is able to assign small weights to views of low quality. Extensive experiments on a broad range of datasets validate the effectiveness of our approach against competitive algorithms.

Qiaoyu Tan, Guoxian Yu
Southwest University
tqy1995119@email.swu.edu.cn, gxyu@swu.edu.cn

Carlotta Domeniconi
George Mason University
cdomenic@gmu.edu

Jun Wang, Zili Zhang
Southwest University
kingjun@swu.edu.cn, zhangzl@swu.edu.cn

CP9

Efficient and Effective Accelerated Hierarchical Higher-Order Logistic Regression for Large Data Quantities

Machine learning researchers are facing a data deluge – quantities of training data have been increasing at a rapid rate. However, most of machine learning algorithms were proposed in the context of learning from relatively smaller quantities of data. We argue that a big data classifier should have superior feature engineering capability, minimal tuning parameters and should be able to learn decision boundaries in fewer passes through the data. In this paper, we have proposed an (computationally) efficient yet (classification-wise) effective family of learning algorithms that fulfils these properties. The proposed family of learning algorithms is based on recently proposed accelerated higher-order logistic regression algorithm: ALR^n. The contributions of this work are three-fold. First, we have added the functionality of out-of-core learning in ALR^n, resulting in a limited pass learning algorithm. Second, superior feature engineering capabilities are built and third, a far more efficient (memory-wise) implementation has been proposed. We demonstrate the competitiveness of our proposed algorithm by comparing its performance not only with state-of-the-art classifier in out-of-core learning such as Selective KDB but also with state-of-the-art in in-core learning such as Random Forest.

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

Francois Petitjean
Faculty of Information Technology
Monash University
francois.petitjean@monash.edu

Geoffrey Webb
Monash University

geoff.webb@monash.edu

CP10

Discriminative Prototype Set Learning for Nearest Neighbor Classification

The nearest neighbor rule is a classic but essential classification model, particularly in problems where the supervising information is given by pairwise dissimilarities and the embedding function are not easily obtained. Prototype selection provides means of generalization and improving efficiency of the nearest neighbor model, but many existing methods assume and rely on the analyses of the input vector space. In this paper, we explore a dissimilarity-based, parametrized model of the nearest neighbor rule. In the proposed model, the selection of the nearest prototypes is influenced by the parameters of the respective prototypes. It provides a formulation for minimizing the violation of the extended nearest neighbor rule over the training set in a tractable form to exploit numerical techniques. We show that the minimization problem reduces to a large-margin principled learning and demonstrate its advantage by empirical comparisons with other prototype selection methods.

Shin Ando
Tokyo University of Science
Dept. of Management
ando@rs.tus.ac.jp

CP10

Limited-Memory Common-Directions Method for Distributed L1-Regularized Linear Classification

For distributed linear classification, L1 regularization is useful because of a smaller model size. However, with the non-differentiability, it is more difficult to develop efficient optimization algorithms. In the past decade, OWLQN has emerged as the major method for distributed training of L1 problems. In this work, we point out issues in OWLQN's search directions. Then we extend the recently developed limited-memory common-directions method for L2-regularized problems to L1 scenarios. Through a unified interpretation of batch methods for L1 problems, we explain why OWLQN has been a popular method and why our method is superior in distributed environments. Experiments confirm that the proposed method is faster than OWLQN in most situations.

Wei-Lin Chiang, Yu-Sheng Li
National Taiwan University
b02902056@ntu.edu.tw, b03902086@ntu.edu.tw

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

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

CP10

A Salient Ensemble of Trees Using Cascaded Linear Classifiers with Feature-Cost Constraints

In many applications the classification model needs to utilize limited resources properly while predicting an instance, e.g. the limited response time for a real-time search engine. In order to satisfy the resource constraint, many researchers try to simplify the model structure or shrink the feature subset size. Because the informative features may take too much cost for the model, a common way is to build a model by considering the trade-off between performance and cost. However, most previous works assume that the cost of a feature is independent of the cost of another feature, which is not practical in reality. In the paper, we consider two categories of the feature cost, individual cost and group cost. The former is independent of the cost of any other feature whereas the latter regards the cost dependency between the other features in the corresponding group. We propose a two-stage framework that integrates the cost-sensitive feature selection and learning a model with a cost budget constraint. First, we propose the group-cost-sensitive random forest (GOAT) model to consider these two costs to select a proper feature subset. Second, we propose a salient ensemble of trees each of which uses cascaded linear classifiers (ETIC) with the satisfaction of the feature-cost constraints using the derived features from the GOAT model.

Chien-Wen Huang, Chung-Kuang Chou
National Taiwan University
cwhuang1021@arbor.ee.ntu.edu.tw,
ckchou@arbor.ee.ntu.edu.tw

Ming-Syan Chen
Dept. of Electrical Engineering National Taiwan University
mschen@ntu.edu.tw

CP10

ALE: Additive Latent Effect Models for Grade Prediction

The past decade has seen a growth in the development and deployment of educational technologies for assisting college-going students in choosing majors, selecting courses and acquiring feedback based on past academic performance. Grade prediction methods seek to estimate a grade that a student may achieve in a course that she may take in the future (e.g., next term). Accurate and timely prediction of students academic grades is important for developing effective degree planners and early warning systems, and ultimately improving educational outcomes. In this paper, we propose additive latent effect models that incorporate several factors to predict the student next-term grades. The proposed models take into account four factors: (i) students academic level, (ii) course instructors, (iii) student global latent factor, and (iv) latent knowledge factors. We compared the new models with several state-of-the-art methods on students of various characteristics (e.g., whether a student transferred in or not). The experimental results demonstrate that the proposed methods significantly outperform the baselines on grade prediction problem. Moreover, we perform a thorough analysis on the importance of different factors and how these factors can practically assist students in course selection, and finally improve their academic performance.

Zhiyun Ren
George Mason University
jessica_dl2008@hotmail.com

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

Huzefa Rangwala
George Mason University
indiana university - purdue university indianapoli

CP10

An Lstm Approach to Patent Classification Based on Fixed Hierarchy Vectors

Recently, innovative techniques for text processing like Latent Dirichlet Allocation (LDA) and embedding algorithms like Paragraph Vectors (PV) allowed for improved text classification and retrieval methods. Even though these methods can be adjusted to handle different text collections, they do not take advantage of the fixed document structure that is mandatory in many application areas. In this paper, we focus on patent data which mandates a fixed structure. We propose a new classification method which represents documents as Fixed Hierarchy Vectors (FHV), reflecting the document's structure. FHVs represent a document on multiple levels where each level represents the complete document but with a different local context. Furthermore, we sequentialize this representation and classify documents using LSTM-based architectures. Our experiments show that FHVs provide a richer document representation and that sequential classification improves classification performance when classifying patents into the International Patent Classification (IPC) taxonomy.

Matthias Schubert
Ludwig-Maximilians University Munich
schubert@dbs.ifi.lmu.de

Marawan Shalaby
Technical University of Munich
shalaby@in.tum.de

Jan Stutzki
LMU Munich
stutzki@dbs.ifi.lmu.de

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

CP10

A Practitioners' Guide to Transfer Learning for Text Classification Using Convolutional Neural Networks

Transfer Learning (TL) plays a crucial role when a given dataset has insufficient labeled examples to train an accurate model. In such scenarios, the knowledge accumulated within a model pre-trained on a source dataset can be transferred to a target dataset, resulting in the improvement of the target model. Though TL is found to be successful in the realm of image-based applications, its impact and practical use in Natural Language Processing (NLP) applications is still a subject of research. Due to their hierarchical architecture, Deep Neural Networks (DNN) provide flexibility and customization in adjusting their parameters and depth of layers, thereby forming an apt area for exploiting the use of TL. In this paper, we report the results and conclusions obtained from extensive empirical experiments using a Convolutional Neural Network (CNN) and try to uncover thumb rules to ensure a successful *positive* transfer. In addition, we also highlight the flawed means that could lead to a *negative* transfer. We explore the transferability of various layers and describe the effect of varying hyper-parameters on the transfer performance. Also, we present a comparison of accuracy value and model size against state-of-the-art methods. Finally, we derive inferences from the empirical results and provide best practices to achieve a successful positive transfer.

Tushar Semwal
Indian Institute of Technology Guwahati
semwaltushar@gmail.com

Gaurav Mathur, Promod Yenigalla
Samsung R & D Institute-Bangalore
gaurav.m4@samsung.com, promod.y@samsung.com

Shivashankar Nair
Indian Institute of Technology Guwahati
sbnair@iitg.ac.in

CP11

Framework for Inferring Leadership Dynamics of Complex Movement from Time Series

Leadership plays a key role in social animals, including humans, decision-making and coalescence in coordination such as hunting, migration, sport etc. Leadership is a process that organizes interactions among members to make a group achieve collective goals. Understanding initiation of coordination allows scientists to gain more insight into social species behaviors. However, by using only time series of activities data, inferring leadership as manifested by the initiation of coordination faces many challenging issues. First, coordination is dynamic and are changing over time. Second, several different coordination events might occur simultaneously among subgroups. Third, there is no fundamental concept to describe these activities computationally. In this paper, we formalize Faction Initiator Inference Problem and propose a leadership inference framework as a solution of this problem. The framework makes no assumption about the characteristics of a leader or the parameters of the coordination process. The framework performs better than our non-trivial baseline in both simulated and biological datasets (schools of fish). We also illustrate the application of our framework as a tool to study group merging and splitting dynamics on trajectories of wild baboons. In addition, our problem formalization and framework enable opportunities for scientists to analyze coordination and generate hypotheses about collective behaviors that can be tested statistically and in the field.

Chainarong Amornbunchornvej
Department of Computer Science
University of Illinois at Chicago
camorn2@uic.edu

Tanya Y. Berger-Wolf
University of Illinois at Chicago
tanyabw@uic.edu

CP11

Sparse Decomposition for Time Series Forecasting and Anomaly Detection

Anomaly detection and forecasting are two fundamental problems in time series analysis. Although these problems have been investigated in the literature previously, the assumptions therein are too restrictive for autonomous analysis. Common examples of limiting assumptions include perfect knowledge about the time series seasonality and/or presence of anomaly free time windows. Current practice is to manually input this knowledge into anomaly detection and forecasting systems which negate any possibility of autonomous analysis. This paper relaxes these assumptions by jointly estimating the latent components (viz. seasonality, level changes, and spikes) in the observed time series without assuming the availability of anomaly-free time windows. The novel and flexible two stage approach proposed herein is based on (a) sparse modeling of the different latent components of the time series and (b) ARMA modeling for fitting the error. The approach leads to a solution for anomaly detection with control over type-I errors. Further, by design, the method is robust against anomalies in the observation window when it is used to solve the forecasting problem by extrapolation. Experiments are conducted with both synthetic and real datasets to demonstrate the efficacy of the proposed method. We compare our approach to various popular baselines. The presented approach outperforms baseline algorithms for anomaly detection in all our experiments and performs favorably for the forecasting task.

Sunav Choudhary
Adobe Research
schoudha@adobe.com

Gaurush Hiranandani
UIUC
gaurush2@illinois.edu

Shiv Saini
Adobe Research
shsaini@adobe.com

CP11

Exact Mean Computation in Dynamic Time Warping Spaces

Dynamic time warping constitutes a major tool for analyzing time series. In particular, computing a mean series of a given sample of series in dynamic time warping spaces (by minimizing the Frchet function) is a challenging computational problem, so far solved by several heuristic, inexact strategies. We spot several inaccuracies in the literatue on exact mean computation in dynamic time warping spaces. Our contributions comprise an exact dynamic program computing a mean (useful for benchmarking and evaluating known heuristics). Empirical evaluations reveal significant deficits of the state-of-the-art heuristics in terms of their output quality. Finally, we give an exact polynomial-time algorithm for the special case of binary time series.

Markus Brill
Technische Universität Berlin, Germany
brill@tu-berlin.de

Till Fluschnik, Vincent Froese, Brijnesh Jain, Rolf Niedermeier, David Schultz
TU Berlin, Germany
till.fluschnik@tu-berlin.de, vincent.froese@tu-berlin.de, johannes.jain@dai-labor.de, rolf.niedermeier@tu-berlin.edu, david.schultz@dailabor.de

CP11

StreamCast: Fast and Online Mining of Power Grid Time Sequences

How can we efficiently forecast the power consumption of a location for the next few days? More challengingly, how can we forecast the power consumption if the temperature increases by 10 degrees C, the number of appliances in the grid increase by 20%, and voltage levels increase by 5%? Such 'what-if scenarios' are crucial for future planning, to ensure that the grid remains reliable even under extreme conditions. Our contributions are as follows: 1) Domain knowledge infusion: we propose a novel Temporal BIG model that extends the physics-based BIG model, allowing it to capture changes over time, trends, and seasonality, and temperature effects. 2) Forecasting: our method algorithm forecasts multiple steps ahead and outperforms baselines in accuracy. Our algorithm is online, requiring constant update time per new data point and bounded memory. 3) What-if scenarios and anomaly detection: our approach can handle scenarios in which the voltage levels, temperature, or number of appliances change. It also spots anomalies in real data, and provides confidence intervals for its forecasts, to assist in planning for various scenarios. Experimental results show that method has 27% lower forecasting error than baselines on real data, scales linearly, and runs in 4 minutes on a time sequence of 40 million points.

Bryan Hooi, Hyun Ah Song, Amritanshu Pandey, Marko Jereminov, Larry Pileggi, Christos Faloutsos
Carnegie Mellon University
bhooi@andrew.cmu.edu, hyunahs@cs.cmu.edu, amritanp@andrew.cmu.edu, mjeremin@andrew.cmu.edu, pileggi@andrew.cmu.edu, christos@cs.cmu.edu

CP11

Brain EEG Time Series Selection: A Novel Graph-Based Approach for Classification

Brain Electroencephalography (EEG) classification is widely applied to analyze cerebral diseases in recent years. Unfortunately, invalid/noisy EEGs degrade the diagnosis performance and most previously developed methods ignore the necessity of EEG selection for classification. To this end, this paper proposes a novel maximum weight clique-based EEG selection approach, named mwcEEGs, to map EEG selection to searching maximum similarity-weighted cliques from an improved Frechet distance-weighted undirected EEG graph simultaneously considering edge weights and vertex weights. Our mwcEEGs improves the classification performance by selecting intra-clique pairwise similar and inter-clique discriminative EEGs with similarity threshold. Experimental results demonstrate the algorithm effectiveness compared with the state-of-the-art time series selection algorithms on real-world EEG datasets.

Chenglong Dai
Nanjing University of Aeronautics and Astronautics
chenglongdai@nuaa.edu.cn

Jia Wu
Macquarie University
wujiawb@gmail.com

Dechang Pi, Lin Cui
Nanjing University of Aeronautics and Astronautics
dc.pi@nuaa.edu.cn, jsjxcuilin@nuaa.edu.cn

CP12

Avoidance Region Discovery: A Summary of Re-

sults

Given a set of GPS trajectories, avoidance region discovery (ARD) finds regions that are avoided by drivers. ARD is important for applications such as sociology, city/transportation planning and crime mitigation, where it can help domain users understand the driver behavior under different concerns (e.g. rush hour, congestion, dangerous neighborhood, etc.). ARD is challenging because of the large number of trajectories with thousands of GPS points, large number of candidate avoidance regions, and the cost of evaluating those. Related work is focused on finding evasive trajectories for a given set of avoidance regions. Distinct from the related work, we propose an Avoidance Region Miner (ARM) approach that can detect both the avoidance regions and evasive trajectories just by using the trajectories in hand without the need of an additional input. A case study on real trajectory data confirms that ARM discovers such regions for further investigation by domain users. Experiments show that ARM yields substantial computational savings compared to a baseline approach.

Emre Eftelioglu, Shashi Shekhar, Xun Tang
University of Minnesota
emre@cs.umn.edu, shekhar@umn.edu, xuntang@cs.umn.edu

CP12

A Rare and Critical Condition Search Technique and its Application to Telescope Stray Light Analysis

Many systems, including space satellites, cannot be upgraded or repaired easily during their missions. Simulation-based design techniques are often used to check conditions that can induce critical malfunctions in them, to ensure sufficient credibility and reliability during operation. However, critical conditions with a very low probability of occurring (e.g., 10^{-8} per trial) rarely appear within a tractable number of simulations. We propose herein a multicanonical Markov Chain Monte Carlo (MCMC) technique extended for the efficient search of rare but critical conditions, to significantly enhance simulation efficiency. Furthermore, we demonstrate an application of our proposed technique to an efficient search of stray light in a space telescope satellite.

Keiichi Kisamori
AIST
NEC Corporation
k-kisamori@aist.go.jp

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

Yoshio Kameda
AIST
NEC Corporation
y.kameda@aist.go.jp

Ryohei Fujimaki
NEC Corporation
rfujimaki@nec-labs.com

CP12

Black-Box Expectation Propagation for Bayesian Models

In this paper, we develop a generic black-box expectation propagation (BBEP) algorithm that can be directly applied to Bayesian models without model-specific derivations. BBEP is built on the spirit of using Monte Carlo estimates, where the moment matching step in EP is replaced with Monte Carlo approximations. To avoid high variance, we employ importance sampling for variance reduction and analyze how to find an optimal proposal distribution. We compare BBEP against the state-of-the-art black-box algorithms on both synthetic and real-world data sets. The experimental results indicate that BBEP can reach better predictive performance than baseline algorithms, and even can be a par with analytical solutions in some settings.

Ximing Li
Jilin university
liximing86@gmail.com

Changchun Li, Jinjin Chi
Jilin University, China
changchunli93@gmail.com, chijinjin616@gmail.com

Jihong Ouyang
Jilin university
ouyj@jlu.edu.cn

Wenting Wang
Jilin University, China
jlu@163.com

CP12

Revenue Maximization on the Multi-Grade Product

Revenue maximization with utilization of social influences aims at earning the highest revenue by properly pricing the product and/or seeding customers. This paper considers the marketing of multi-grade product, where competitive and promotional relationships exist simultaneously among different grades of a product. To tackle the problem, we propose a new influence diffusion model, Multi-Grade IC, and a novel algorithm, Pricing-Seeding. Simulations using real customer reviews show the Pricing-Seeding can promote high-revenue grades of a product.

Ya-Wen Teng
Department of Electrical Engineering
National Taiwan University
ywteng@arbor.ee.ntu.edu.tw

Chih-Hua Tai
Department of Computer Science and Information Engineering
National Taipei University
hanatai@mail.ntpu.edu.tw

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

Ming-Syan Chen
Dept. of Electrical Engineering National Taiwan University

mschen@ntu.edu.tw

CP12

Learning Convolutional Text Representations for Visual Question Answering

Visual question answering (VQA) is a recently proposed artificial intelligence task that requires a deep understanding of both images and texts. In deep learning, images are typically modeled through convolutional neural networks (CNNs) while texts are typically modeled through recurrent neural networks (RNNs). In this work, we perform a detailed analysis on the natural language questions in VQA, which raises a different need for text representations as compared to other natural language processing tasks. Based on the analysis, we propose to rely on CNNs for learning text representations. By exploring various properties of CNNs specialized for text data, we present our "CNN Inception + Gate' model for text feature extraction in VQA. The experimental results show that simply replacing RNNs with our CNN-based model improves question representations and thus the overall accuracy of VQA models. In addition, our model has much fewer parameters and the computation is much faster. We also prove that the text representation requirement in VQA is more complicated and comprehensive than that in conventional natural language processing tasks. Shallow models like the fastText model, which can obtain comparable results with deep learning models in simple tasks like text classification, have poor performances in VQA.

Zhengyang Wang
Washington State University
zwang6@eecs.wsu.edu

Shuiwang Ji
Washington State Univeristy
sji@eecs.wsu.edu

CP13

Learning to Interact with Users: A Collaborative-Bandit Approach

Learning to interact with users and discover their preferences is central in most web applications, with recommender systems being a notable example. From such a perspective, merging interactive learning algorithms with recommendation models is natural. While recent literature has explored the idea of combining collaborative filtering approaches with bandit techniques, there exist two limitations: they usually consider Gaussian rewards, which are not suitable for implicit feedback data powering most recommender systems, and they are restricted to one-item recommendation while typically a list of recommendations is given. To address these limitations, apart from Gaussian rewards we also consider Bernoulli rewards, the latter being suitable for dyadic data. Also, we use two user click models: the one-item click/no-click model, and the cascade click model which is suitable for top-K recommendations. For these settings, we propose novel machine learning algorithms that learn to interact with users by learning the underlying parameters collaboratively across users and items. We provide an extensive empirical study, which is the first to illustrate all pairwise empirical comparisons across different interactive learning recommendation algorithms. Our experiments show that when the number of users and items is large, propagating the feedback across users and items while learning latent features is the most effective approach for systems to learn to interact with the users.

Konstantina Christakopoulou, Arindam Banerjee
University of Minnesota
christa@cs.umn.edu, banerjee@cs.umn.edu

CP13

Modeling Item-Specific Effects for Video Click

Prediction is widely employed to improve the number of video clicks and views, which are the key important indicators (KPIs) due to their contribution to revenue. The available predictive features, however, are generally limited as compared to the expected prediction capability from the algorithm side. Inspired by the intrinsic dependence among multiple clicks for the same video, we hypothesize that there exist some consistent effects involved in grouped click records. We then propose to recover such effects from the associated hidden features, which are likely to alleviate the insufficiency of features. The simulation studies are performed to elucidate how the derived grouped effects empower a model with additional discriminating capacity compared with the original one. The proposed methodology is further examined on the repository of PPTV (a leading video service provider in China) click records comprehensively. The results confirm the existence of the hypothesized effects and demonstrate their critical role in the performance improvement of video click prediction.

Fei Tan, Kuang Du, Zhi Wei, Haoran Liu
New Jersey Institute of Technology
ft54@njit.edu, kd226@njit.edu, zhiwei@njit.edu,
hl425@njit.edu

Chenguang Qin, Ran Zhu
PPLive Inc
exceptionqin@pptv.com, ranzhu@pptv.com

CP13

One-Class Recommendation with Asymmetric Textual Feedback

Personalized ranking with implicit feedback (e.g. purchases, views, check-ins) is an important paradigm in recommender systems. Such feedback sometimes comes with textual information (e.g. reviews, comments, tips), which could be a useful signal to reveal item properties, identify users' tastes and interpret their behavior. Although incorporating such information is common in *explicit* feedback settings (such as rating prediction), it is less common when dealing with implicit feedback, as it is often not available for negative instances (e.g. there is no review associated with the item the user *didn't* buy). Thus our goal in this study is to propose a ranking method (**PRAST**) to incorporate such personalized, asymmetric textual signals in implicit feedback settings. We evaluate our model on two real-world datasets. Quantitative and qualitative results indicate that the proposed approach significantly outperforms standard recommendation baselines, alleviates 'cold start' issues, and is able to provide potential textual interpretations for latent feedback dimensions.

Mengting Wan, Julian McAuley
University of California, San Diego
m5wan@ucsd.edu, jmcauley@ucsd.edu

CP13

Online It Ticket Automation Recommendation Us-

ing Hierarchical Multi-Armed Bandit Algorithms

The increasing complexity of IT environments urgently requires the use of analytical approaches and automated problem resolution for more efficient delivery of IT services. In this paper, we model the automation recommendation procedure of IT automation services as a contextual bandit problem with dependent arms, where the arms are in the form of hierarchies. Intuitively, different automations in IT automation services, designed to automatically solve the corresponding ticket problems, can be organized into a hierarchy by domain experts according to the types of ticket problems. We introduce a novel hierarchical multi-armed bandit algorithms leveraging the hierarchies, which can match the coarse-to-fine feature space of arms. Empirical experiments on a real large-scale ticket dataset have demonstrated substantial improvements over the conventional bandit algorithms. In addition, a case study of dealing with the cold-start problem is conducted to clearly show the merits of our proposed algorithms.

Qing Wang
Florida International University, USA
qwang028@fiu.edu

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

S.S. Iyengar
Florida International University, USA
iyengar@cis.fiu.edu

Larisa Shwartz
IBM T.J. Watson Research Center, USA
lschwart@us.ibm.com

Genady Grabarnik
St. John's University, USA
grabarng@stjohns.edu

CP13

Robust Cost-Sensitive Learning for Recommendation with Implicit Feedback

This paper aims at improvement on the effectiveness of matrix decomposition (MD) methods for implicit feedback. We highlight two critical limitations of existing works: imbalance and outlier. We address the above two issues by learning a robust asymmetric learning model. Particularly, a novel log-determinant function is employed to refine the nuclear norm with respect to the low-rank approximation. We show the promising theoretical and experimental results of our algorithm.

Peng Yang
Institute for Infocomm Research, Singapore
peng.yang.2@kaust.edu.sa

Peilin Zhao
South China University of Technology, China
peilinzhao@hotmail.com

Yong Liu
Link Analytics Centre, NTUC Link, Singapore
liuysc@acm.org

Xin Gao
KAUST, Saudi Arabia
xin.gao@kaust.edu.sa

CP13

Investigating Deep Reinforcement Learning Techniques in Personalized Dialogue Generation

In this paper, we propose a personalized dialogue generation system, which combines reinforcement learning techniques with an attention-based hierarchical recurrent encoder-decoder model. Firstly, we incorporate user-specific information into the decoder to capture user's background information and speaking style. Secondly, we employ reinforcement learning techniques to maximize future reward in dialogue, which enables our system to generate topic-coherent, informative and grammatical responses. Moreover, we propose three types of rewards to characterize good conversations. Finally, we compare the performance of the following reinforcement learning methods in dialogue generation: policy gradient, Q-learning, and actor-critic algorithms. We conduct experiments to verify the effectiveness of the proposed model on two dialogue datasets. Experimental results demonstrate that our model can generate better-personalized dialogues for different users. Quantitatively, our method achieves better performance than the state-of-the-art dialogue systems in terms of BLEU score, perplexity, and human evaluation.

Min Yang
Chinese Academy of Sciences
min.yang1129@gmail.com

Qiang Qu
Shenzhen Institutes of Advanced Technology
Chinese Academy of Sciences
qiang@siat.ac.cn

Kai Lei
School of Electronics and Computer Engineering
Peking University Shenzhen Graduate School
leik@pkusz.edu.cn

Jia Zhu
School of Computer Science
South China Normal University
jzhu@m.scnu.edu.cn

Zhou Zhao
School of Computer Science
Zhejiang University
zhaozhou@zju.edu.cn

Xiaojun Chen, Joshua Zhexue Huang
School of Computing Science
Shenzhen University
xjchen@szu.edu.cn, zx.huang@szu.edu.cn

CP14

SMACD: Semi-Supervised Multi-Aspect Community Detection

Community detection in real-world graphs has been shown to benefit from using multi-aspect information, e.g., in the form of means of communication" between nodes in the network. An orthogonal line of work, broadly construed as semi-supervised learning, approaches the problem by introducing a small percentage of node assignments to communities and propagates that knowledge throughout the graph. In this paper we introduce SMACD, a novel semi-supervised multi-aspect community detection method along with an automated parameter tuning algorithm which essentially renders SMACD parameter-free. To the best of our knowledge, SMACD is the first approach to incorporate multi-aspect graph information and semi-supervision, while being able to discover overlapping and non-overlapping communities. We extensively evaluate SMACD's performance in comparison to state-of-the-art approaches across eight real and two synthetic datasets and demonstrate that SMACD, through combining semi-supervision and multi-aspect edge information, outperforms the baselines.

Ekta Gujral
University of California, Riverside
Computer Science and Engineering
egujr001@ucr.edu

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

CP14

Maximizing the Effect of Information Adoption: A General Framework

With the development of social networking services, social influence analyses, as well as the influence maximization tasks, have attracted wide attention in both academia and industry. Traditional studies mainly focus on simulating process of influence spread. However, two basic functions of social spread, i.e., information propagation and information adoption have not been clearly distinguished. Usually, as information adoption could be even more significant for information publishers in application scenarios, more comprehensive analysis for effect of adoption is urgently required. To that end, in this paper, we propose a novel framework to generally describe social spread, in which information adoption process is separately formulated as random events. Along this line, when we apply this framework to the information adoption maximization task, with proving that the adoption maximization problem is NP-hard and submodular, we further design a polling-based algorithm to achieve an effective approximation. Extensive experiments on four real-world data sets demonstrate the effectiveness and efficiency of proposed algorithms, which validates that our approach could better summarize the complete social spread process, and further support the necessity of distinguishing information adoption from information propagation.

Tianyuan Jin
University of Science and Technology of China
tongxu@mail.u
jty123@mail.ustc.edu.cn

Tong Xu, Hui Zhong, Enhong Chen, Zhefeng Wang, Qi Liu
University of Science and Technology of China
tongxu@ustc.edu.cn, zhuiwin@mail.ustc.edu.cn, cheneh@ustc.edu.cn, zhfwang@mail.ustc.edu.cn, qiliuql@ustc.edu.cn

CP14

Multi-Layered Network Embedding

Network embedding has gained more attentions in recent years. It has been shown that the learned low-dimensional node vector representations could advance a myriad of graph mining tasks such as node classification, community detection, and link prediction. A vast majority of the existing efforts are overwhelmingly devoted to single-layered networks or homogeneous networks with a single type of nodes and node interactions. However, in many real-world applications, a variety of networks could be abstracted and presented in a multi-layered fashion. Typical multi-layered networks include critical infrastructure systems, collaboration platforms, social recommender systems, to name a few. Despite the widespread use of multi-layered networks, it remains a daunting task to learn vector representations of different types of nodes due to the bewildering combination of both within-layer connections and cross-layer network dependencies. In this paper, we study a novel problem of multi-layered network embedding. In particular, we propose a principled framework - MANE to model both within-layer connections and cross-layer network dependencies simultaneously in a unified optimization framework for embedding representation learning. Experiments on real-world multi-layered networks corroborate the effectiveness of the proposed framework.

Jundong Li, Chen Chen, Hanghang Tong, Huan Liu
Arizona State University
jundongl@asu.edu, chen_chen@asu.edu, hanghang.tong@asu.edu, huan.liu@asu.edu

CP14

Toward Relational Learning with Misinformation

Relational learning has been proposed to cope with the interdependency among linked instances in a network, and it is a fundamental tool to categorize social network users for various tasks. However, the emerging widespread of misinformation in social networks, information that is inaccurate or false, poses novel challenges to utilizing social media data. Malicious users may actively manipulate their content and characteristics, which easily lead to a noisy dataset. Hence, it is intricate for traditional relational learning approaches to deliver an accurate predictive model in the presence of misinformation. In this work, we precisely focus on the problem by proposing a joint framework that simultaneously constructs a relational learning model and mitigates the effect of misinformation by restraining anomalous points. Empirical results on real-world social media data prove the superiority of the proposed approach, Relational Learning with Misinformation (RLM), over traditional approaches on modeling social network users.

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

Fred Morstatter
University of Southern California
morstatt@usc.edu

Huan Liu
Arizona State University
huanliu@asu.edu

CP14

Reconstructing a Cascade from Temporal Observations

Given a subset of active nodes in a network can we reconstruct the cascade that has generated these observations? This is a problem that has been studied in the literature, but here we focus in the case that temporal information is available about the active nodes. In particular, we assume that in addition to the subset of active nodes we also know their activation time. We formulate this cascade-reconstruction problem as a variant of a Steiner-tree problem: we ask to find a tree that spans all reported active nodes while satisfying temporal-consistency constraints. We present three approximation algorithms. The best algorithm in terms of quality achieves a $\mathcal{O}(\sqrt{k})$-approximation guarantee, where k is the number of active nodes, while the most efficient algorithm has linearithmic running time, making it scalable to very large graphs. We evaluate our algorithms on real-world networks with both simulated and real cascades. Our results indicate that utilizing the available temporal information allows for more accurate cascade reconstruction. Furthermore, our objective leads to finding the “backbone’ of the cascade and it gives solutions of high precision.

Han Xiao, Polina Rozenshtein
Aalto University
han.xiao@aalto.fi, polina.rozenshtein@aalto.fi

Nikolaj Tatti
F-Secure
nikolaj.tatti@f-secure.fi

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

CP14

Modeling Co-Evolution Across Multiple Networks

Multiple and co-evolving networks are common in many real settings such as social networks, communication networks and other information networks. Most of the work in the field of network evolution has focused on a single evolving network or specific network pairs, lacking generality in the analysis of multiple networks and ignoring the co-evolutionary dynamics between networks. In practice, a significant amount of information is encoded in the evolution of multiple networks with respect to one another. In this paper, we show how to use a shared temporal matrix factorization framework to model co-evolution across multiple networks, and we refer to this framework as CoEvol. Specifically, the proposed framework decomposes the adjacency matrix of each co-evolving network into a product of network-independent shared factor and a set of network-specific temporal factors, and impose a non-negativity constraint on the factors for greater interpretability. Our approach has the potential to predict multiple changes in co-evolving networks over time, because of its ability to explicitly represent co-evolving networks as a function of time. The CoEvol framework also has the advantage of generality in addressing various temporal tasks across multiple networks. We show the benefits of this approach in predicting co-evolution across multiple networks on the tasks including cross-network link prediction, lag correlation detection and community detection.

Wenchao Yu
University of California, Los Angeles
yuwenchao@ucla.edu

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

Wei Wang
University of California Los Angeles
weiwang@cs.ucla.edu

CP15

Strongly Hierarchical Factorization Machines and Anova Kernel Regression

High-order parametric models that include terms for feature interactions are applied to various data mining tasks, where ground truth depends on interactions of features. However, with sparse data, the high-dimensional parameters for feature interactions often face three issues: expensive computation, difficulty in parameter estimation and lack of structure. Previous work has proposed approaches which can partially resolve the three issues. In particular, models with factorized parameters (e.g. Factorization Machines) and sparse learning algorithms (e.g. FTRL-Proximal) can tackle the first two issues but fail to address the third. Regarding to unstructured parameters, constraints or complicated regularization terms are applied such that hierarchical structures can be imposed. However, these methods make the optimization problem more challenging. In this work, we propose Strongly Hierarchical Factorization Machines and ANOVA kernel regression where all the three issues can be addressed without making the optimization problem more difficult. Experimental results show the proposed models significantly outperform the state-of-the-art in two data mining tasks: cold-start user response time prediction and stock volatility prediction.

Ruocheng Guo, Hamidreza Alvari, Paulo Shakarian
Arizona State University
rguo12@asu.edu, halvari@asu.edu, shak@asu.edu

CP15

Personalized Ranking on Poisson Factorization

Matrix factorization (MF) has earned great success on recommender systems. However, the commonly-used regression-based MF is not only sensitive to outliers but also unable to guarantee that the predicted values are in line with the user preference orders, which is the basis of common measures in recommender systems, e.g., nDCG. To overcome this drawback, we propose personalized ranking on Poisson factorization (PRPF), which utilizes the posteriori based on pair-wise learning to rank instead of the classical regression-based ones. Since the posteriori that combines learning to rank and Poisson factorization does not follow the conjugate prior relationship, we estimate variational parameters approximately and propose two optimization approaches based on variational interference. Due to the combination, PRPF not only preserves user preference but also performs well on a sparse matrix. In the experiment, we show that PRPF outperforms the state-of-the-art methods and achieves promising results for recommendation tasks.

Li-Yen Kuo
Dept. of Electrical Engineering National Taiwan
University
lykuo@arbor.ee.ntu.edu.tw

Chung-Kuang Chou
National Taiwan University
ckchou@arbor.ee.ntu.edu.tw

Ming-Syan Chen

Dept. of Electrical Engineering National Taiwan University
mschen@ntu.edu.tw

CP15

Making Kernel Density Estimation Robust Towards Missing Values in Highly Incomplete Multivariate Data Without Imputation

Density estimation is one of the most frequently used data analytics techniques. A major challenge is that real-world datasets often contain missing values, e.g. due to sampling errors or data loss. The recovery of these missing values is often impossible or too expensive. Missing values are not necessarily limited to a few features or few data objects. This renders many methods based on auxiliary variables unsuitable. In this paper we explore three alternative models that are based on the new concept of virtual objects. Additionally we present an approximation that is computationally efficient. Experiments with incomplete datasets show that our method is superior to established imputation methods.

Richard Leibrandt
Technische Universität München, Germany
r.leibrandt@tum.de

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

CP15

A Novel Genetic Algorithm for Feature Selection in Hierarchical Feature Spaces

Feature selection methods have been widely adopted to prepare high-dimensional feature spaces for the classification task of data mining. However, in many real-world datasets, the feature space is formed by binary features related via generalization-specialization relationships, also known as hierarchical feature spaces. Although there are many methods for the traditional feature selection problem, methods which properly consider hierarchical features are still very underexplored. In this work, we propose a novel genetic algorithm (GA) for hierarchical feature selection. The proposed GA has two novel hierarchical mutation operators tailored to deal with redundant features in hierarchical feature spaces. The computational experiments show that our proposed approach exhibited better predictive performance than two state-of-the-art hierarchical feature selection methods (SHSEL and HIP) and also than two traditional feature selection methods (ReliefF and CFS).

Pablo Silva, Alexandre Plastino
Universidade Federal Fluminense
psilva@ic.uff.br, plastino@ic.uff.br

Alex A. Freitas
University of Kent
a.a.freitas@kent.ac.uk

CP15

Dense Neighborhood Pattern Sampling in Numerical Data

Pattern mining in numerical data remains a challenging task due to the pattern search space that becomes potentially infinite with continuous dimensions. Most approaches reluctantly reduced the expressiveness of mined patterns to make possible extraction. Despite this expressiveness loss, they do not provide results within a short response time of a few seconds. This paper addresses the instant discovery of patterns in numerical data based on sampling techniques. Instead of splitting each dimension into intervals, we use a metric to introduce the density as new interestingness measure, and to define neighborhood patterns. The language of neighborhood patterns is semantically rich but in return, its size is infinite. We then present a new exact and non-enumerative random procedure to sample this infinite language according to density. An experimental study demonstrates the good compromise between precision and diversity of neighborhood patterns. Finally, in the context of associative classification, we show that a sample of neighborhood patterns is as accurate as traditional methods that traverses the entire search space.

Arnaud Giacometti
University Francois Rabelais of Tours, France
arnaud.giacometti@univ-tours.fr

Arnaud Soulet
University François Rabelais of Tours, France
arnaud.soulet@univ-tours.fr