KDD Accepted Papers

December 12, 2017

Contents

1. Accelerating Innovation Through Analogy Mining

Tom Hope (Hebrew University of Jerusalem);Joel Chan (Carnegie Mellon University);Aniket Kittur (Carnegie Mellon University);Dafna Shahaf (Hebrew University of Jerusalem)

The availability of large idea repositories (e.g., the U.S. patent database) could significantly accelerate innovation and discovery by providing people with inspiration from solutions to analogous problems. However, finding useful analogies in these large, messy, real-world repositories remains a persistent challenge for either human or automated methods. Previous approaches include costly hand-created databases that have high relational structure (e.g., predicate calculus representations) but are very sparse. Simpler machine-learning/information-retrieval similarity metrics can scale to large, natural-language datasets, but struggle to account for structural similarity, which is central to analogy. In this paper we explore the viability and value of learning simpler structural representations, specifically, "problem schemas", which specify the purpose of a product and the mechanisms by which it achieves that purpose. Our approach combines crowdsourcing and recurrent neural networks to extract purpose and mechanism vector representations from product descriptions. We demonstrate that these learned vectors allow us to find analogies with higher precision and recall than traditional information-retrieval methods. In an ideation experiment, analogies retrieved by our models significantly increased people's likelihood of generating creative ideas compared to analogies retrieved by traditional methods. Our results suggest a promising approach to enabling computational analogy at scale is to learn and leverage weaker structural representations.

http://dl.acm.org/authorize?N33215

2. Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data

David Hallac (Stanford University);Sagar Vare (Stanford University);Stephen Boyd (Stanford University);Jure Leskovec (Stanford University)

Subsequence clustering of multivariate time series is a useful tool for discovering repeated patterns in temporal data. Once these patterns have been discovered, seemingly complicated datasets can be interpreted as a temporal sequence of only a small number of states, or clusters. For example, raw sensor data from a fitness-tracking application can be expressed as a timeline of a select few actions (i.e., walking, sitting, running). However, discovering these patterns is challenging because it requires simultaneous segmentation and clustering of the time series. Furthermore, interpreting the resulting clusters is difficult, especially when the data is high-dimensional. Here we propose a new method of model-based clustering, which we call Toeplitz Inverse Covariance-based Clustering (TICC). Each cluster in the TICC method is defined by a correlation network, or Markov random field (MRF), characterizing the interdependencies between different observations in a typical subsequence of that cluster. Based on this graphical representation, TICC simultaneously segments and clusters the time series data. We solve the TICC problem through an expectation maximization (EM) algorithm. We derive closed-form solutions to efficiently solve both the E and M-steps in a scalable way, through dynamic programming and the alternating direction method of multipliers (ADMM), respectively. We validate our approach by comparing TICC to several state-of-the-art baselines in a series of synthetic experiments, and we then demonstrate on an automobile sensor dataset how TICC can be used to learn interpretable clusters in real-world scenarios.

http://dl.acm.org/authorize?N33203

3. HinDroid: An Intelligent Android Malware Detection System Based on Structured Heterogeneous Information Network

Yanfang Ye (West Virginia University);Shifu Hou (West Virginia University);Yangqiu Song (West Virginia University)

With explosive growth of Android malware and due to the severity of its damages to smart phone users, the detection of Android malware has become an increasingly important topic in cyber security. The increasing sophistication of Android malware calls for new defensive techniques that

are harder to evade, and are capable of protecting users against novel threats. In this paper, to detect Android malware, instead of using Application Programming Interface (API) calls only, we further analyze the different relationships between them and create higher-level semantics which require more efforts for attackers to evade the detection. We represent the Android applications (apps), related APIs, and their rich relationships as a structured heterogeneous information network (HIN). Then we use a meta-path based approach to characterize the semantic relatedness of apps and APIs. We use each meta-path to formulate a similarity measure over Android apps, and aggregate different similarities using multi-kernel learning. Then each meta-path is automatically weighted by the learning algorithm to make predictions. To the best of our knowledge, this is the rest work to use structured HIN for Android malware detection. Comprehensive experiments on real sample collections from Comodo Cloud Security Center are conducted to compare various malware detection approaches. Promising experimental results demonstrate that our developed system HinDroid system outperforms other alternative Android malware detection techniques. HinDroid has already been incorporated into the scanning tool of Comodo Mobile Security product.

http://dl.acm.org/authorize?N33330

4. DeepSD: Generating High Resolution Climate Change Projections through Single Image Super-Resolution

Thomas Vandal (Northeastern University);Evan Kodra (risQ Inc.);Sangram Ganguly (Bay Area Environmental Research Institute / NASA Ames Research Center);Andrew Michaelis (University Corporation, Monterey Bay);Ramakrishna Nemani (NASA Ames Research Center);Auroop Ganguly (Northeastern University)

The impacts of climate change are felt by most critical systems, such as infrastructure, ecological systems, and power-plants. However, contemporary Earth System Models (ESM) are run at spatial resolutions too coarse for assessing effects this localized. Local scale projections can be obtained using statistical downscaling, a technique which uses historical climate observations to learn a low-resolution to high-resolution mapping. Depending on statistical modeling choices, downscaled projections have been shown to vary significantly terms of accuracy and reliability. The spatio-temporal nature of the climate system motivates the adaptation of super-resolution image processing techniques to statistical downscaling. In our work, we present DeepSD, a generalized stacked super resolution convolutional neural network (SRCNN) framework for statistical downscaling of climate variables. DeepSD augments SRCNN with multi-scale input channels to maximize predictability in statistical downscaling. We provide a comparison with Bias Correction Spatial Disaggregation as well as three Automated-Statistical Downscaling approaches in downscaling daily precipitation from 1 degree (100km) to 1/8 degrees (12.5km) over the Continental United States. Furthermore, a framework using the NASA Earth Exchange (NEX) platform is discussed for downscaling more than 20 ESM models with multiple emission scenarios.

http://dl.acm.org/authorize?N33456

5. An efficient bandit algorithm for realtime multivariate optimization

Daniel Hill (Amazon.com);Houssam Nassif (Amazon.com);Yi Liu (Amazon.com);Anand Iyer (Amazon.com);S. V. N. Vishwanathan (vishy@amazon.com)

Optimization is commonly employed to determine the content of web pages, such as to maximize conversions on landing pages or click-through rates on search engine result pages. Often the layout of these pages can be decoupled into several separate decisions. For example, the composition of a landing page may involve deciding which image to show, which wording to use, what color background to display, etc. Thus, optimization is a combinatorial problem over an exponentially large decision space. Randomized experiments do not scale well to this setting, and therefore, in practice, one is typically limited to optimizing a single aspect of a web page at a time. This represents a missed opportunity in both the speed of experimentation and the exploitation of possible interactions between layout decisions.

Here we focus on multivariate optimization of interactive web pages. We formulate an approach where the possible interactions between different components of the page are modeled explicitly.

We apply bandit methodology to explore the layout space efficiently and use hill-climbing to select optimal content in realtime. Our algorithm also extends to contextualization and personalization of layout selection. Simulation results show the suitability of our approach to large decision spaces with strong interactions between content. We further apply our algorithm to optimize a message that promotes adoption of an Amazon service. After only a single week of online optimization, we saw a 21

http://dl.acm.org/authorize?N33461

6. Randomization or Condensation?: Linear-Cost Matrix Sketching Via Cascaded Compression Sampling

Kai Zhang (Lawrence Berkeley National Laboratories);Chuanren Liu (Drexel University);Jie Zhang (Fudan University);Hui Xiong (Rutgers University);Eric Xing (Carneigie Mellon University);Jieping Ye (University of Michigan)

The problem of matrix decomposition is to learn compact representations of a matrix while simultaneously preserving most of its properties, which is a fundamental building block in modern scientific computing and big data applications. Currently, even state-of-the-art solutions still require the use of the entire input matrix in generating desired factorizations, causing a major computational and memory bottleneck. In this paper, we uncover an interesting theoretic connection between matrix low-rank decomposition and *lossy data compression*, based on which a cascaded compression sampling framework is devised to approximate an $m \times n$ matrix in only $\mathcal{O}(m+n)$ time and space. Indeed, the proposed method accesses only a small number of matrix rows and columns, which significantly improves the memory footprint. Meanwhile, by sequentially teaming two rounds of approximation procedures and upgrading the sampling strategy from a uniform probability to more sophisticated, encoding-orientated sampling, significant algorithmic boosting is achieved to uncover more granular structures in the data. Empirical results on a wide spectrum of real-world, large-scale matrices show that by taking only linear time and space, the accuracy of our method rivals those state-of-the-art randomized algorithms consuming a quadratic, $\mathcal{O}(mn)$, amount of resources.

http:////dl.acm.org/authorize?N33243

7. Scalable and Sustainable Deep Learning via Randomized Hashing

Ryan Spring (Rice University);Anshumali Shrivastava (Rice University)

Current deep learning architectures are growing larger in order to learn from complex datasets. These architectures require giant matrix multiplication operations to train millions of parameters. Conversely, there is another growing trend to bring deep learning to low-power, embedded devices. The matrix operations, associated with both training and testing of deep networks, are very expensive from a computational and energy standpoint. We present a novel hashing-based technique to drastically reduce the amount of computation needed to train and test deep networks. Our approach combines two recent ideas, Adaptive Dropout and Randomized Hashing for Maximum Inner Product Search (MIPS), to select the nodes with the highest activation efficiently. Our new algorithm for deep learning reduces the overall computational cost of the forward and backward propagation steps by operating on significantly fewer (sparse) nodes. As a consequence, our algorithm uses only 5

http://dl.acm.org/authorize?N33236

8. TrioVecEvent: Embedding-Based Online Local Event Detection in Geo-Tagged Tweet Streams

Chao Zhang (University of Illinois at Urbana-Champaign);Liyuan Liu (University of Illinois at Urbana-Champaign);Dongming Lei (University of Illinois at Urbana-Champaign);Quan Yuan (University of Illinois at Urbana-Champaign);Honglei Zhuang (University of Illinois at Urbana-Champaign);Tim Hanratty (U.S. Army Research Lab);Jiawei Han (University of Illinois at Urbana-Champaign)

Detecting local events (e.g., protest, disaster) at their onsets is an important task for a wide spectrum of applications, ranging from disaster control to crime monitoring and place recommendation. Recent years have witnessed growing interest in leveraging geo-tagged tweet streams for online local event detection. Nevertheless, the accuracies of existing methods still remain unsatisfactory for building reliable local event detection systems. We propose TrioVecEvent, a method that leverages multimodal embeddings to achieve accurate online local event detection. The effectiveness of TrioVecEvent is underpinned by its two-step detection scheme. First, it ensures a high coverage of the underlying local events by dividing the tweets in the query window into coherent geo-topic clusters. To generate quality geo-topic clusters, we capture short-text semantics by learning multimodal embeddings of the location, time, and text, and then perform online clustering with a novel Bayesian mixture model. Second, TrioVecEvent considers the geo-topic clusters as candidate events and extracts a set of features for classifying the candidates. Leveraging the multimodal embeddings as background knowledge, we introduce discriminative features that can well characterize local events, which enables pinpointing true local events from the candidate pool with a small amount of training data. We have used crowdsourcing to evaluate TrioVecEvent, and found that it improves the detection precision of the state-of-the-art method from 36.8

http://dl.acm.org/authorize?N33241

9. Clustering Individual Transactional Data for Masses of Users

Riccardo Guidotti (University of Pisa);Anna Monreale (University of Pisa);Mirco Nanni (KDD-Lab ISTI-CNR Pisa);Fosca Giannotti (ISTI-CNR);Dino Pedreschi (University of Pisa)

Mining a large number of datasets recording human activities for making sense of individual data is the key enabler of a new wave of personalized knowledge-based services. In this paper we focus on the problem of clustering individual transactional data for a large mass of users. Transactional data is a very pervasive kind of information that is collected by several services, o

http://dl.acm.org/authorize?N33201

10. Learning certifiably optimal rule lists for categorical data

Elaine Angelino (UC Berkeley);Nicholas Larus-Stone (Harvard);Daniel Alabi (Harvard);Margo Seltzer (Harvard University);Cynthia Rudin (Duke)

We present the design and implementation of a custom discrete optimization technique for building rule lists over a categorical feature space. Our algorithm provides the optimal solution, with a certificate of optimality. By leveraging algorithmic bounds, efficient data structures, and computational reuse, we achieve several orders of magnitude speedup in time and a massive reduction of memory consumption. We demonstrate that our approach produces optimal rule lists on practical problems in seconds. This framework is a novel alternative to CART and other decision tree methods.

http://dl.acm.org/authorize?N33295

11. Meta-Graph Based Recommendation Fusion over Heterogeneous Information Networks

Huan Zhao (HKUST);Quanming Yao (HKUST);Jianda Li (HKUST);Yangqiu Song (HKUST);Dik Lee (HKUST)

Heterogeneous Information Network (HIN) is a natural and general representation of data shown in modern large commercial recommender systems with heterogeneous types of data. Recommendation based on HIN faces two challenges: how to represent the high-level semantics of recommendation and how to fuse the heterogeneous information to make recommendation. In this paper, we solve the two challenges by first introducing the concept of meta-graph to HIN-based recommendation, and then solving the information fusion problem with a

http://dl.acm.org/authorize?N33355

12. Discovering Reliable Approximate Functional Dependencies

Panagiotis Mandros (Max Planck Institute for Informatics);Mario Boley (Max Planck Institute for Informatics);Jilles Vreeken (Max Planck Institute for Informatics)

Given a database and a target attribute of interest, how can we tell whether there exists a functional, or approximately functional dependency of the target on any set of other attributes in the data? How can we reliably, without bias to sample size or dimensionality, measure the strength of such a dependency? And, how can we efficiently discover the optimal or alpha-approximate top-k dependencies? These are exactly the questions we answer in this paper.

As we want to be agnostic on the form of the dependency, we adopt an information-theoretic approach, and construct a reliable, bias correcting score that can be efficiently computed. Moreover, we give an effective optimistic estimator of this score, by which for the first time we can mine the approximate functional dependencies from data with guarantees of optimality. Empirical evaluation shows that the derived score achieves a good bias for variance trade-off, can be used within an efficient discovery algorithm, and indeed discovers meaningful dependencies. Most important, it remains reliable in the face of data sparsity.

http://dl.acm.org/authorize?N33227

13. Effective and Real-time In-App Activity Analysis in Encrypted Internet Traffic Streams

Junming Liu (Rutgers University);Yanjie Fu (Missouri University of Science and Technology);Jingci Ming (Rutgers University);Yong Ren (Futurewei Tech. Inc);Leilei Sun (Dalian University of Technology);Hui Xiong (Rutgers University)

The mobile in-App service analysis, aiming at classifying mobile internet traffic into different types of service usages, has become a challenging and emergent task for mobile service providers due to the increasing adoption of secure protocols for in-App services. While some efforts have been made for the classification of mobile internet traffic, existing methods reply on complex feature construction and large storage cache, which lead to low processing speed, and thus not practical for online real-time scenarios. To this end, we develop an iterative analyzer for classifying encrypted mobile traffic in a real-time way. Specifically, we first select an optimal set of most discriminative features from raw features extracted from traffic packet sequences by a novel Maximizing Inner activity similarity and Minimizing Different activity similarity (MIMD) measurement.

To develop the online analyzer, we first represent a traffic flow with a series of time windows, where each is described by the optimal feature vector and is updated iteratively at the packet level. Instead of extracting feature elements from a series of raw traffic packets, our feature elements are updated when a new traffic packet is observed and the storage of raw traffic packets is not required.

The time windows generated from the same service usage activity are grouped by our proposed method, namely recursive time continuity constrained KMeans clustering (rCKC). The feature vectors of cluster centers are then fed into a random forest classifier to identify corresponding service usages. Finally, we provide extensive experiments on real-world traffic data from Wechat, Whatsapp and Facebook to demonstrate the effectiveness and efficiency of our approach. The results show that the proposed analyzer provides high accuracy in real-world scenarios, and has low storage cache requirement as well as fast processing speed.

http://dl.acm.org/authorize?N33225

14. Local Higher-Order Graph Clustering

Hao Yin (Stanford University);Austin R. Benson (Stanford University);Jure Leskovec (Stanford University);David F. Gleich (Purdue University)

Local graph clustering methods aim to find a cluster of nodes by exploring a small region of the graph. These methods are attractive because they enable targeted clustering around a given seed node and are faster than traditional global graph clustering methods because their runtime does not depend on the size of the input graph. However, current local graph partitioning methods are not designed to account for the higher-order structures crucial to the network, nor can they effectively handle directed networks. Here we introduce a new class of local graph clustering methods that address these issues by incorporating higher-order network information captured by small subgraphs, also called network motifs. First, we show how to adapt the approximate personalized PageRank algorithm to find clusters containing a seed node with minimal motif conductance, a generalization of the conductance metric for network motifs. We also generalize existing theory to maintain the properties of fast running time (independent of the size of the graph) and cluster quality (in terms of motif conductance). For community detection tasks on both synthetic and real-world networks, our new framework outperforms the current edge-based personalized PageRank methodology. Second, we develop a theory of node neighborhoods for finding sets that have small motif conductance, where the motif is a clique. We apply these results to the case of finding good seed nodes to use as input to the personalized PageRank algorithm.

http://dl.acm.org/authorize?N33247

15. Functional Annotation of Human Protein Coding Isoforms via Non-convex Multi-Instance Learning

Tingjin Luo (National University of Defense Technology);Weizhong Zhang (Zhejiang University);Shuang Qiu (University of Michigan);Yang Yang (Beihang University);Dongyun Yi (National University of Defense Technology);Guangtao Wang (University of Michigan);Jieping Ye (University of Michigan);Jie Wang (University of Michigan)

Functional annotation of human genes is fundamentally important for understanding the molecular basis of various genetic diseases. A major challenge in determining the functions of human genes lies in the functional diversity of proteins, that is, a gene can perform different functions as it may consist of multiple protein coding isoforms (PCIs). Therefore, differentiating functions of PCIs can significantly deepen our understanding of the functions of genes. However, due to the lack of isoform-level gold-standards (ground-truth annotation), many existing functional annotation approaches are developed at gene-level. In this paper, we propose a novel approach to differentiate the functions of PCIs by integrating sparse simplex projection—-that is, a nonconvex sparsity-inducing regularizer—-with the framework of multi-instance learning (MIL). Specifically, we label the genes that are annotated to the function under consideration as *positive bags* and the genes without the function as *negative bags*. Then, by sparse projections onto simplex, we learn a mapping that embeds the original bag space to a discriminative feature space. Our framework is flexible to incorporate various smooth and nonsmooth loss functions such as logistic loss and hinge loss. To solve the resulting highly nontrivial non-convex and nonsmooth optimization problem, we further develop an efficient block coordinate decent algorithm. Extensive experiments on human genome data demonstrate that the proposed approaches significantly outperform the state-of-the-art methods in terms of functional annotation accuracy of human PCIs and efficiency.

http://dl.acm.org/authorize?N33226

16. FORA: Simple and Effective Approximate Single-Source Personalized PageRank

Sibo Wang (Nanyang Technological University);Renchi Yang (Nanyang Technological University);Xiaokui Xiao (Nanyang Technological University);Zhewei Wei (Renmin University of China);Yin Yang (Hamad Bin Khalifa University)

Given a graph G ,a source node s and a target node t, the Personalized PageRank (PPR) of t with respect to s is the probability that a random walk starting from s terminates at t . A single-source PPR (SSPPR) query enumerates all nodes in G , and returns the top-k nodes with the highest PPR values with respect to a given source node s . SSPPR has important applications in web search and social networks, e.g., in Twiter

http://dl.acm.org/authorize?N33232

17. A Minimal Variance Estimator for the Cardinality of Big Data Set Intersection

Reuven Cohen (Technion);Liran Katzir (Technion);Aviv Yehezkel (Technion)

In recent years there has been a growing interest in developing "streaming algorithms" for efficient processing and querying of continuous data streams. These algorithms seek to provide accurate results while minimizing the required storage and the processing time, at the price of a small inaccuracy in their output. A fundamental query of interest is the intersection size of two big data streams. This problem arises in many different application areas, such as network monitoring, database systems, data integration and information retrieval. In this paper we develop a new algorithm for this problem, based on the Maximum Likelihood (ML) method. We show that this algorithm outperforms all known schemes in terms of the estimation's quality (lower variance) and that it asymptotically achieves the optimal variance.

http://dl.acm.org/authorize?N33291

18. KATE: K-Competitive Autoencoder for Text

Yu Chen (RPI);Mohammed Zaki (RPI)

Autoencoders have been successful in learning meaningful representations from image datasets. However, their performance on text datasets has not been widely studied. Traditional autoencoders tend to learn possibly trivial representations of text documents due to their confounding properties such as high-dimensionality, sparsity and power-law word distributions. In this paper, we propose a novel k-competitive autoencoder, called KATE, for text documents. Due to the competition between the neurons in the hidden layer, each neuron becomes specialized in recognizing specific data patterns, and overall the model can learn meaningful representations of textual data. A comprehensive set of experiments show that KATE can learn better representations than traditional autoencoders including denoising, contractive, variational, and k-sparse autoencoders. Our model also outperforms deep generative models, probabilistic topic models, and even word representation models (e.g., Word2Vec) in terms of several downstream tasks such as document classification, regression, and retrieval.

http://dl.acm.org/authorize?N33290

19. Online Ranking with Constraints: A Primal-Dual Algorithm and Applications to Web Traffic-Shaping

Parikshit Shah (Yahoo Research);Akshay Soni (Yahoo Research);Troy Chevalier (Yahoo Research)

We study the online constrained ranking problem motivated by an application to web-traffic shaping: an online stream of sessions arrive in which, within each session, we are asked to rank items. The challenge involves optimizing the ranking in each session so that local vs. global objectives are controlled: within each session one wishes to maximize a reward (local) while satisfying certain constraints over the entire set of sessions (global). A typical application of this setup is that of page optimization in a web portal. We wish to rank items so that not only is user engagement maximized in each session, but also other business constraints (such as the number of views/clicks delivered to various publishing partners) are satisfied.

We describe an online algorithm for performing this optimization. A novel element of our approach is the use of linear programming duality and connections to the celebrated Hungarian algorithm. This framework enables us to determine a set of *shadow prices* for each traffic-shaping constraint that can then be used directly in the final ranking function to assign near-optimal rankings. The (dual) linear program can be solved off-line periodically to determine the prices. At serving time these prices are used as weights to compute weighted rank-scores for the items, and the simplicity of the approach facilitates scalability to web applications. We provide rigorous theoretical guarantees for the performance of our online algorithm and validate our approach using numerical experiments on real web-traffic data from a prominent internet portal.

http://dl.acm.org/authorize?N33222

20. Discrete Content-aware Matrix Factorization

Defu Lian (University of Electronic Science and Technology of China);Rui Liu (University of Electronic Science and Technology of China);Yong Ge (University of Arizona);Kai Zheng (University of Queensland);Xing Xie (Microsoft Research);Longbing Cao (University of Technology Sydney)

Precisely recommending relevant items from massive candidates to a large number of users is an indispensable yet computationally expensive task in many online platforms (e.g., Amazon.com and Netflix.com). A promising way is to project users and items into a Hamming space and then recommend items via Hamming distance. However, previous studies didn't address the cold-start challenges and couldn't make the best use of preference data like implicit feedback. To fill this gap, we propose a Discrete Content-aware Matrix Factorization (DCMF) model, 1) to derive compact yet informative binary codes at the presence of user/item content information; 2) to support the classification task based on a local upper bound of logit loss; 3) to introduce an interaction regularization for dealing with the sparsity issue. We further develop an efficient discrete optimization algorithm for parameter learning. Based on extensive experiments on three real-world datasets, we show that DCFM outperforms the state-of-the-arts on both regression and classification tasks.

http://dl.acm.org/authorize?N33214

21. Communication-Efficient Distributed Block Minimization for Nonlinear Kernel Machines

Si Si (UT austin);Cho-Jui Hsieh (UC Davis);Inderjit Dhillon (UT austin)

Nonlinear kernel machines often yield superior predictive performance on various tasks; however, they suffer from severe computational challenges. In this paper, we show how to overcome the important challenge of speeding up kernel machines using multiple computers. In particular, we develop a parallel block minimization framework, and demonstrate its good scalability in solving nonlinear kernel SVM and logistic regression. Our framework proceeds by dividing the problem into smaller subproblems by forming a block-diagonal approximation of the Hessian matrix. The subproblems are then solved approximately in parallel. After that, a communication efficient line search procedure is developed to ensure sufficient reduction of the objective function value by exploiting the problem structure of kernel machines. We prove global linear convergence rate of the proposed method with a wide class of subproblem solvers, and our analysis covers strongly convex and some non-strongly convex functions. We apply our algorithm to solve large-scale kernel SVM problems on distributed systems, and show a significant improvement over existing parallel solvers. As an example, on the covtype dataset with half-a-million samples, our algorithm can obtain an approximate solution with 96% accuracy in 20 seconds using 32 machines, while all the other parallel kernel SVM solvers require more than 2000 seconds to achieve a solution with 95% accuracy. Moreover, our algorithm is the first distributed kernel SVM solver that can scale to massive data sets. On the KDDB dataset (20 million samples and 30 million features), our parallel solver can compute the kernel SVM solution within half an hour using 32 machines with 640 cores in total, while existing solvers can not scale to this dataset.

http://dl.acm.org/authorize?N33216

22. Accelerating Innovation Through Analogy Mining

Tom Hope (Hebrew University of Jerusalem);Joel Chan (Carnegie Mellon University);Aniket Kittur (Carnegie Mellon University);Dafna Shahaf (Hebrew University of Jerusalem)

The availability of large idea repositories (e.g., the U.S. patent database) could significantly accelerate innovation and discovery by providing people with inspiration from solutions to analogous problems. However, finding useful analogies in these large, messy, real-world repositories remains a persistent challenge for either human or automated methods. Previous approaches include costly hand-created databases that have high relational structure (e.g., predicate calculus representations) but are very sparse. Simpler machine-learning/information-retrieval similarity metrics can

scale to large, natural-language datasets, but struggle to account for structural similarity, which is central to analogy. In this paper we explore the viability and value of learning simpler structural representations, specifically, “problem schemas”, which specify the purpose of a product and the mechanisms by which it achieves that purpose. Our approach combines crowdsourcing and recurrent neural networks to extract purpose and mechanism vector representations from product descriptions. We demonstrate that these learned vectors allow us to find analogies with higher precision and recall than traditional information-retrieval methods. In an ideation experiment, analogies retrieved by our models significantly increased people’s likelihood of generating creative ideas compared to analogies retrieved by traditional methods. Our results suggest a promising approach to enabling computational analogy at scale is to learn and leverage weaker structural representations.

http://dl.acm.org/authorize?N33215

23. Collaboratively Improving Topic Discovery and Word Embeddings by Coordinating Global and Local Contexts

Guangxu Xun (State University of New York at Buffalo);Yaliang Li (State University of New York at Buffalo);Jing Gao (State University of New York at Buffalo);Aidong Zhang (State University of New York at Buffalo)

A text corpus typically contains two types of context information—global context and local context. Global context carries topical information which can be utilized by topic models to discover topic structures from the text corpus, while local context can train word embeddings to capture semantic regularities reflected in the text corpus. This encourages us to exploit the useful information in both the global and the local context information. In this paper, we propose a unified language model based on matrix factorization techniques which 1) takes the complementary global and local context information into consideration simultaneously, and 2) models topics and learns word embeddings collaboratively. We empirically show that by incorporating both global and local context, this collaborative model can not only significantly improve the performance of topic discovery over the baseline topic models, but also learn better word embeddings than the baseline word embedding models. We also provide qualitative analysis that explains how the cooperation of global and local context information can result in better topic structures and word embeddings.

http://dl.acm.org/authorize?N33245

24. Coresets for Kernel Regression

Yan Zheng (University of Utah);Jeff Phillips (University of Utah)

Kernel regression is an essential and ubiquitous tool for non-parametric data analysis, particularly popular among time series and spatial data. However, the central operation which is performed many times, evaluating a kernel on the data set, takes linear time. This is impractical for modern large data sets.

In this paper we describe coresets for kernel regression: compressed data sets which can be used as proxy for the original data and have provably bounded worst case error. The size of the coresets are independent of the raw number of data points; rather they only depend on the error guarantee, and in some cases the size of domain and amount of smoothing. We evaluate our methods on very large time series and spatial data, and demonstrate that they incur negligible error, can be constructed extremely efficiently, and allow for great computational gains.

http://dl.acm.org/authorize?N33356

25. Towards an Optimal Subspace for K-Means

Dominik Mautz (Ludwig-Maximilians-Universität München);Wei Ye (Ludwig-Maximilians-Universität München);Claudia Plant (Universität Wien);Christian Böhm (Ludwig-Maximilians-Universität München)

Is there an optimal dimensionality reduction for k-means, revealing the prominent cluster structure hidden in the data? We propose SubKmeans, which extends the classic k-means algorithm. The goal of this algorithm is twofold: find a sufficient k-means-style clustering partition and transform the clusters onto a common subspace, which is optimal for the cluster structure. Our solution is able to pursue these two goals simultaneously. The dimensionality of this subspace is found automatically and therefore the algorithm comes without the burden of additional parameters. At the same time this subspace helps to mitigate the curse of dimensionality. The SubKmeans optimization algorithm is intriguingly simple and efficient. It is easy to implement and can readily be adopted to the current situation. Furthermore, it is compatible to many existing extensions and improvements of k-means.

http://dl.acm.org/authorize?N33228

26. Unsupervised Network Discovery for Brain Imaging Data

Zilong Bai (University of California, Davis);Peter Walker (Naval Medical Research Center);Anna Tschiffely (Naval Medical Research Center);Fei Wang (Cornell University);Ian Davidson (University of California, Davis)

A common problem with spatiotemporal data is how to simplify the data to discover an underlying network that consists of cohesive spatial regions (nodes) and relationships between those regions (edges). This network discovery problem naturally exists in a multitude of domains including climate data (dipoles), astronomical data (gravitational lensing) and the focus of this paper, fMRI scans of human subjects. Whereas previous work requires strong supervision, we propose an unsupervised matrix tri-factorization formulation with complex constraints and spatial regularization. We show that this formulation works well in controlled experiments with synthetic networks and is able to recover the underlying ground-truth network. We then show that for real fMRI data our approach can reproduce well known results in neurology regarding the default mode network in resting-state healthy and Alzheimer affected individuals.

http://dl.acm.org/authorize?N33297

27. Robust Top-k Multi-class SVM for Visual Category Recognition

Xiaojun Chang (Carnegie Mellon University);Yao-Liang Yu (University of Waterloo);Yi Yang (University of Technology Sydney)

Classification problems with a large number of classes inevitably involve overlapping or similar classes. In such cases it seems reasonable to allow the learning algorithm to make mistakes on similar classes, as long as the true class is still among the top-k (say) predictions. Likewise, in applications such as search engine or ad display, we are allowed to present k predictions at a time and the customer would be satisfied as long as her interested prediction is included. Inspired by the recent work of [**?**], we propose a very generic, robust multiclass SVM formulation that directly aims at minimizing a weighted and truncated combination of the ordered prediction scores. Our method includes many previous works as special cases. Computationally, using the Jordan decomposition Lemma we show how to rewrite our objective as the difference of two convex functions, based on which we develop an efficient algorithm that allows incorporating many popular regularizers (such as the ℓ_2 and ℓ_1 norms). We conduct extensive experiments on four real large-scale visual category recognition datasets, and obtain very promising performances.

http://dl.acm.org/authorize?N33299

28. Patient Subtyping via Time-Aware LSTM Networks

Inci Baytas (Michigan State University);Cao Xiao (IBM T. J. Watson Research Center);Xi Zhang (Cornell University);Fei Wang (Cornell University);Anil Jain (Michigan State University);Jiayu Zhou (Michigan State University)

In the study of various diseases, the heterogeneity among patients usually leads to different progression patterns and may require different types of therapeutic intervention. Therefore, it is important to study patient subtyping, the grouping of patients into disease characterizing subtypes. Subtyping from complex patient data is challenging because of the information heterogeneity and temporal dynamics. Long-Short Term Memory (LSTM) has been successfully used in many domains for processing sequential data, and recently applied for analyzing longitudinal patient records. The LSTM units are designed to handle data with constant elapsed times between consecutive elements of the sequence. Given that time lapse between successive elements in patient records can vary from days to months, the design of traditional LSTM may lead to suboptimal performance. In this paper, we propose a novel LSTM unit called Time Aware LSTM (T-LSTM) to handle irregular time intervals in longitudinal patient records. We learn a subspace decomposition of the cell memory which enables time decay to discount the memory content according to the elapsed time. We propose a patient subtyping model that leverages the proposed T-LSTM in an auto-encoder to learn a powerful single representation for sequential records of patients, which are then used to cluster patients into clinical subtypes. Experiments on synthetic and real world datasets show that the proposed T-LSTM architecture captures the underlying structures in the sequences with time irregularities.

http://dl.acm.org/authorize?N33298

29. AnnexML: Approximate Nearest Neighbor Search for Extreme Multi-label Classification

Yukihiro Tagami (Yahoo Japan Corporation)

Extreme multi-label classification methods have been widely used in Web-scale classification tasks such as Web page tagging and product recommendation. In this paper, we present a novel graph embedding method called AnnexML. At training step, AnnexML constructs k-nearest neighbor graph of the label vectors and attempts to reproduce the graph structure in the embedding space. The prediction is efficiently performed by using an approximate nearest neighbor search method which efficiently explores the learned k-nearest neighbor graph in the embedding space. We conducted evaluations on several large-scale real-world data sets and compared our method with recent state-of-the-art methods. Experimental results show our AnnexML can significantly improve prediction accuracy, especially on data sets that have larger label space. In addition, AnnexML improves the trade-off between prediction time and accuracy. At the same level of accuracy, the prediction time of AnnexML was up to 58 times faster than that of SLEEC, which is a state-of-the-art embedding-based method.

http://dl.acm.org/authorize?N33237

30. Graph Edge Partitioning via Neighborhood Heuristic

Chenzi Zhang (the University of Hong Kong);Fan Wei (Stanford University);Qin Liu (Huawei Noah's Ark Lab);Zhihao Gavin Tang (the University of Hong Kong);Zhenguo Li (Huawei Noah's Ark Lab)

We consider the edge partitioning problem that partitions the edges of an input graph into multiple balanced components, while minimizing the total number of vertices replicated (one vertex might appear in more than one partition). This problem is critical in minimiz- ing communication costs and running time for several large-scale distributed graph computation platforms (e.g., PowerGraph, Spark GraphX). We first prove that this problem is NP-hard, and then present a new partitioning heuristic with polynomial running time. We provide a worst-case upper bound of replication factor for our heuristic on general graphs. To our knowledge, we are the first to provide such bound for edge partitioning algorithms on general graphs. Applying this bound to random power-law graphs greatly improves the previous bounds of expected replication factor. Ex- tensive experiments demonstrated that our partitioning algorithm consistently produces much smaller replication factors on various benchmark data sets than the state-of-the-art. When deployed in the production graph engine, PowerGraph, in average it reduces replication factor, communication, and running time by 54

http://dl.acm.org/authorize?N33242

31. Unsupervised P2P Rental Recommendations via Integer Programming

Yanjie Fu (Missouri University of Science and Technology);Guannan Liu (Beihang University);Mingfei Teng (Rutgers University);Charu Aggarwal (IBM T. J. Watson Research Center)

Due to the sparseness of quality rating data, unsupervised recommender systems are used in many applications in Peer to Peer (P2P) rental marketplaces such as Airbnb, FlipKey, and HomeAway. We present an integer programming based recommender systems, where both accommodation benefit and community risk of lodging places are measured and are incorporated into objective function as utility measurements. More specifically, we first present an unsupervised fused scoring method for quantifying the accommodation benefit and community risk of a lodging with crowd-sourced geo-tagged data. In the view of maximizing the utility of recommendations, we formulate the unsupervised P2P rental recommendations as a constrained integer programming problem, where the accommodation benefit of recommendations is maximized and the community risk of recommendations is minimized, while maintaining constraints on personalization. Furthermore, we provide an efficient solution for the optimization problem by developing a learning to integer programming method for combining aggregated listwise learning to rank into branching variable selection. We apply the proposed approach to the Airbnb data of New York City and provide lodging recommendations to travelers. In empirical experiments, we demonstrate the effectiveness of our method in striking a trade-off among satisfaction time on market, number of reviews, and achieving a balance between positive and negative sides, as well as the efficiency enhancement of our methods.

http://dl.acm.org/authorize?N33208

32. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing

Kun Kuang (Tsinghua University);Peng Cui (Tainghua University);Bo Li (Tsinghua University);Meng Jiang (University of Illinois Urbana-Champaign);Shiqiang Yang (Tsinghua University)

Estimating treatment effect plays an important role on decision making in many fields, such as social marketing, healthcare, and public policy. The key challenge on estimating treatment effect in the wild observational studies is to handle confounding bias induced by imbalance of the confounder distributions between treated and control units. Traditional methods remove confounding bias by re-weighting units with supposedly accurate propensity score estimation under the unconfoundedness assumption. Controlling high-dimensional variables may make the unconfoundedness assumption more plausible, but poses new challenge on accurate propensity score estimation. One strand of recent literature seeks to directly optimize weights to balance confounder distributions, bypassing propensity score estimation. But existing balancing methods fail to do selection and differentiation among the pool of a large number of potential confounders, leading to possible underperformance in many high dimensional settings. In this paper, we propose a data-driven Differentiated Confounder Balancing (DCB) algorithm to jointly select confounders, differentiate weights of confounders and balance confounder distributions for treatment effect estimation in the wild high dimensional settings. The synergistic learning algorithm we proposed is more capable of reducing the confounding bias in many observational studies. To validate the effectiveness of our DCB algorithm, we conduct extensive experiments on both synthetic and real datasets. The experimental results clearly demonstrate that our DCB algorithm outperforms the state-of-the-art methods. We further show that the top features ranked by our algorithm generate accurate prediction of online advertising effect.

http://dl.acm.org/authorize?N33218

33. Similarity Forests

Saket Sathe (IBM T. J. Watson Research Center);Charu Aggarwal (IBM T. J. Watson Research Center)

Random forests are among the most successful methods used in data mining because of their extraordinary accuracy and effectiveness. However, their use is primarily limited to multidimensional

data because they sample features from the original data set. In this paper, we propose a method for extending random forests to work with any arbitrary set of data objects as long as similarities can be computed among the data objects. Furthermore, since it is understood that similarity computation between all $O(n^2)$ pairs of objects might be expensive, our method computes only a very small fraction of the $O(n^2)$ pairwise similarities between objects to construct the forests. Our results show that the proposed similarity forest approach is extremely efficient and is also very accurate on a wide variety of data sets. Therefore, this paper significantly extends the applicability of random forest methods to arbitrary data domains. Furthermore, the approach even outperforms traditional random forests on multidimensional data. In many cases, the similarity matrices learned from arbitrary applications are noisy, because of the difficulty in estimating similarity values between pairs of objects. Similarity forests are very robust to errors in classification. In many practical settings, the similarity values between objects are incompletely specified because of the difficulty in collecting such values. In such cases, the similarity forest approach can be naturally extended to a partially specified similarity matrix.

http://dl.acm.org/authorize?N33221

34. The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables

Himabindu Lakkaraju (Stanford University);Jon Kleinberg (Cornell University);Jure Leskovec (Stanford University);Jens Ludwig (University of Chicago);Sendhil Mullainathan (Harvard University)

Evaluating whether machines improve on human performance is one of the central questions of machine learning. However, there are many domains where the data is *selectively labeled* in the sense that the observed outcomes are themselves a consequence of the existing choices of the human decision-makers. For instance, in the context of judicial bail decisions, we observe the outcome of whether a defendant fails to return for their court appearance only if the human judge decides to release the defendant on bail. Comparing the performance of humans and machines on data with this type of bias can lead to erroneous estimates and wrong conclusions. Here we propose a novel framework for evaluating the performance of predictive models on selectively labeled data. We develop an evaluation methodology that is robust to the presence of unmeasured confounders (unobservables). We propose a metric that allows us to evaluate the effectiveness of any given black-box predictive model and benchmark it against the performance of human decision-makers. We also develop an approach called *contraction* which allows us to compute this metric without resorting to counterfactual inference by exploiting the heterogeneity of human decision-makers. Experimental results on real world datasets spanning diverse domains such as health care, insurance, and criminal justice demonstrate the utility of our evaluation metric in comparing human decisions and machine predictions. Experiments on synthetic data also show that our contraction technique produces accurate estimates of our evaluation metric.

http://dl.acm.org/authorize?N33219

35. HoORaYs: High-order Optimization of Rating Distance for Recommender Systems

Jingwei Xu (Nanjing University);Yuan Yao (Nanjing University);Hanghang Tong (City College, CUNY);Xianping Tao (Nanjing Univ.);Jian Lu (Nanjing University)

Latent factor models have become a prevalent method in recommender systems, to predict users' preference on items based on the historical user feedback. Most of the existing methods, explicitly or implicitly, are built upon the first-order rating distance principle, which aims to minimize the difference between the estimated and real ratings. In this paper, we generalize such first-order rating distance principle and propose a new latent factor model HoORaYs for recommender systems. The core idea of the proposed method is to explore high-order rating distance, which aims to minimize not only (i) the difference between the estimated and real ratings of the same (user, item) pair (i.e., the first-order rating distance), but also (ii) the difference between the estimated and real rating difference of the same user across different items (i.e., the second-order rating distance).

We formulate it as a regularized optimization problem, and propose an effective and scalable algorithm to solve it. Our analysis from the geometry and Bayesian perspectives indicate that by exploring the high-order rating distance, it helps to reduce the variance of the estimator, which in turns leads to better generalization performance (e.g., smaller prediction error). We evaluate the proposed method on four real-world data sets, two with explicit user feedback and the other two with implicit user feedback. Experimental results show that the proposed method consistently outperforms the state-of-the-art methods in terms of the prediction accuracy.

http://dl.acm.org/authorize?N33234

36. An Online Hierarchical Algorithm for Extreme Clustering

Ari Kobren (University of Massachusetts Amherst);Nicholas Monath (University of Massachusetts Amherst);Akshay Krishnamurthy (University of Massachusetts Amherst);Andrew McCallum (University of Massachusetts Amherst)

Many modern clustering methods scale well to a large number of data points, N, but not to a large number of clusters, K. This paper introduces PERCH, a new non-greedy, incremental algorithm for hierarchical clustering that scales to both massive N and K—-a problem setting we term extreme clustering. Our algorithm efficiently routes new data points to the leaves of an incrementally-built tree. Motivated by the desire for both accuracy and speed, our approach performs tree rotations for the sake of enhancing subtree purity and encouraging balancedness. We prove that, under a natural separability assumption, our non-greedy algorithm will produce trees with perfect dendrogram purity regardless of data arrival order. Our experiments demonstrate that PERCH constructs more accurate trees than other tree-building clustering algorithms and scales well with both N and K, achieving a higher quality clustering than the strongest flat clustering competitor in nearly half the time.

http://dl.acm.org/authorize?N33217

37. Is the Whole Greater Than the Sum of Its Parts?

Liangyue Li (Arizona State University);Hanghang Tong (Arizona Stete University);Yong Wang (Hong Kong University of Science and Technology);Conglei Shi (IBM Research);Nan Cao (Tongji University);Norbou Buchler (US Army Research Laboratory)

The part-whole relationship routinely finds itself in many disciplines, ranging from collaborative teams, crowdsourcing, autonomous systems to networked systems. From the algorithmic perspective, the existing work has primarily focused on predicting the outcomes of the whole and parts, by either separate models or linear joint models, which assume the outcome of the parts has a linear and independent effect on the outcome of the whole. In this paper, we propose a joint predictive method named PAROLE to simultaneously and mutually predict the part and whole outcomes. The proposed method offers two distinct advantages over the existing work. First (Model Generality), we formulate joint part-whole outcome prediction as a generic optimization problem, which is able to encode a variety of complex relationships between the outcome of the whole and parts, beyond the linear independence assumption. Second (Algorithm Efficacy), we propose an effective and efficient block coordinate descent algorithm, which is able to find the coordinate-wise optimum with a linear complexity in both time and space. Extensive empirical evaluations on real-world datasets demonstrate that the proposed PAROLE (1) leads to consistent prediction performance improvement by modeling the non-linear part-whole relationship as well as part-part interdependency, and (2) scales linearly in terms of the size of the training dataset.

http://dl.acm.org/authorize?N33211

38. metapath2vec: Scalable Representation Learning for Heterogeneous Networks

Yuxiao Dong (University of Notre Dame);Nitesh V. Chawla (University of Notre Dame);Ananthram Swami (Army Research Laboratory)

We study the problem of representation learning in heterogeneous networks. The unique challenges come from the existence of multiple types of nodes and links, which limit the feasibility of the conventional network embedding techniques. We develop two novel scalable representation learning models, namely metapath2vec and metapath2vec++. The metapath2vec model formalizes meta path based random walks to construct the heterogeneous neighborhood of a node and then leverages a heterogeneous skip-gram model to perform node embeddings. The metapath2vec++ model further enables the simultaneous modeling of structural and semantic correlations in heterogeneous networks. Extensive experiments show that metapath2vec and metapath2vec++ are able to not only outperform state-of-the-art embedding models in various heterogeneous network mining tasks, such as node classification, clustering, and similarity search, but also discern the structural and semantic correlations between diverse network objects.

http://dl.acm.org/authorize?N33205

39. Multi-Aspect Streaming Tensor Completion

Qingquan Song (Texas A&M University);Xiao Huang (Texas A&M University);Hancheng Ge (Texas A&M University);James Caverlee (Texas A&M University);Xia Hu (Texas A&M University)

Tensor completion has become an effective computational tool in many real-world data-driven applications. Beyond traditional static setting, with the increasing popularity of high velocity streaming data, it requires efficient online processing without reconstructing the whole model from scratch. Existing work on streaming tensor completion is usually built upon the assumption that tensors only grow in one mode. Unfortunately, the assumption does not hold in many real-world situations in which tensors may grow in multiple modes, i.e., multi-aspect streaming tensors. Efficiently modeling and completing these incremental tensors without sacrificing its effectiveness remains a challenging task due to the uncertainty of tensor mode changes and complex data structure of multi-aspect streaming tensors. To bridge this gap, we propose a Multi-Aspect Streaming Tensor completion framework MAST based on CANDECOMP/PARAFAC (CP) decomposition to track the subspace of general incremental tensors for completion. In addition, we investigate a special situation where time is one mode of the tensors, and leverage its extra structure information to improve the general framework towards higher effectiveness. Experimental results on four datasets collected from various real-world applications demonstrate the effectiveness and efficiency of the proposed framework.

http://dl.acm.org/authorize?N33235

40. Groups-Keeping Solution Path Algorithm for Sparse Regression with Automatic Feature Grouping

Bin Gu (University of Texas at Arlington);Guodong Liu (Univ. of Texas at Arlington);Heng Huang (University of Texas at Arlington)

Variable selection with identifying homogenous groups of features is crucial for high-dimensional data analysis. Octagonal shrinkage and clustering algorithm for regression (OSCAR) is an important sparse regression approach with automatic feature grouping by ℓ_1 norm and pairwise ℓ_∞ norm. However, due to over-complex representation of the penalty (especially the pairwise ℓ_∞ norm), until now OSCAR has no solution path algorithm which is mostly useful for tuning the model. To address this challenge, in this paper, we propose a groups-keeping solution path algorithm of OSCAR (OscarGKPath). Given a set of homogenous groups of features and an accuracy ε, OscarGKPath can fit the solutions in an interval of regularization parameters while keeping the feature groups. The entire solution path can be obtained by combining multiple such intervals. Theoretically, we prove that all solutions in the solution path produced by OscarGKPath can strictly satisfy the given accuracy ε. The experimental results on a variety of datasets not only confirm the effectiveness of our OscarGKPath, but also show the superiority of our OscarGKPath for cross validation compared with the batch algorithm.

http://dl.acm.org/authorize?N33200

41. EmbedJoin: Efficient Edit Similarity Joins via Embeddings

Haoyu Zhang (Indiana University Bloomington);Qin Zhang (Indiana University Bloomington)

We study the problem of edit similarity joins, where given a set of strings and a threshold value K, we need to output all the pairs of strings whose edit distances are at most K. Edit similarity join is a fundamental operation in numerous applications such as data cleaning/integration, bioinformatics, natural language processing, and has been identified as a primitive operator for database systems. This problem has been studied extensively in the literature. However, we observed that all the existing algorithms fall short on long strings and large distance thresholds. In this paper we propose an algorithm named ebdjoin that scales very well with string length and distance threshold. Our algorithm is built on the recent advance of metric embeddings for edit distance, and is very different from all of the previous approaches. We demonstrate via an extensive set of experiments that ebdjoin significantly outperforms the previous best algorithms on long strings and large distance thresholds. For example, on a collection of 20,000 real-world DNA sequences each of length 20,000 and a distance threshold that is 1% of the string length (1% errors), the previous best algorithms that we have tested cannot finish in 10 hours, while ebdjoin finished in less than 6 minutes. Moreover, ebdjoin scales very well up to 20% errors which is critical in applications such as bioinformatics, and is far beyond the reach of existing algorithms.

http://dl.acm.org/authorize?N33240

42. Collaborative Variational Autoencoder for Recommender Systems

Xiaopeng Li (The Hong Kong University of Science and Technology);James She (The Hong Kong University of Science and Technology)

Modern recommender systems usually employ collaborative filtering with rating information to recommend items to users due to its successful performance. However, because of the drawbacks of collaborative-based methods such as sparsity, cold start, etc., more attention has been drawn to hybrid methods that consider both the rating and content information. Most of the previous works in this area cannot learn a good representation from content for recommendation task or consider only text modality of the content, thus their methods are very limited in current multimedia scenario. This paper proposes a Bayesian generative model called collaborative variational autoencoder (CVAE) that considers both rating and content for recommendation in multimedia scenario. The model learns deep latent representations from content data in an unsupervised manner and also learns implicit relationships between items and users from both content and rating. Unlike previous works with denoising criteria, the proposed CVAE learns a latent distribution for content in latent space instead of observation space through an inference network and can be easily extended to other multimedia modalities other than text. Experiments show that CVAE is able to significantly outperform the state-of-the-art recommendation methods with more robust performance.

http://dl.acm.org/authorize?N33212

43. Tracking the Dynamics in Crowdfunding

Hongke Zhao (USTC);Hefu Zhang (University of Science and Technology of China);Yong Ge (The University of Arizona);Qi Liu (USTC);Enhong Chen (University of Science & Technology of China);Huayu Li (UNCC);Le Wu (HeFei University of Technology)

Crowdfunding is an emerging Internet fundraising mechanism by raising monetary contributions from the crowd for projects or ventures. In these platforms, the dynamics, i.e., daily funding amount on campaigns and perks (backing options with rewards), are the most concerned issue for creators, backers and platforms. However, tracking the dynamics in crowdfunding is very challenging and still underexplored. To that end, in this paper, we present a focused study on this problem. A special goal is to forecast the funding amount in the future days for a given campaign and its perks. Specifically, we formalize the dynamics in crowdfunding as a hierarchical time series, i.e., campaign level and perk level. Specific to each level, we develop a special regression model by modeling

the decision making process of the crowd (visitors and backing probability) and exploring various factors that impact the decision; on this basis, an enhanced switching regression is proposed at each level to address the heterogeneity of funding sequences. Further, we employ a revision matrix to combine the two-level base forecasts for the final forecasting. We conduct extensive experiments on a real-world crowdfunding data collected from Indiegogo. The experimental results clearly demonstrate the effectiveness of our approaches on tracking the dynamics in crowdfunding.

http://dl.acm.org/authorize?N33244

44. Human Mobility Synchronization and Trip Purpose Detection with Mixture of Hawkes Processes

Pengfei Wang (University of Chinese Academy of Science &amp; Computer Network Information Center, Chinese Academy of Sciences);Guannan Liu (Beihang University);Yanjie Fu (Missouri University of Science and Technology);Wenqing Hu (Missouri University of Science and Technology);Charu Aggarwal (IBM T. J. Watson Research Center)

While exploring human mobility can benefit many applications such as smart transportation, city planning, and urban economics, there are two key questions that need to be answered: (i) What is the nature of the spatial diffusion of human mobility across regions with different urban functions? (ii) How to spot and trace the trip purposes of human mobility trajectories? To answer these questions, we study large-scale and city-wide taxi trajectories; and furtherly organize them as arrival sequences according to the chronological arrival time. We figure out an important property across different regions from the arrival sequences, namely human mobility synchronization effect, which can be exploited to explain the phenomenon that two regions have similar arrival patterns in particular time periods if they share similar urban functions. In addition, the arrival sequences are mixed by arrival events with distinct trip purposes, which can be revealed by the regional environment of both the origins and destinations. To that end, in this paper, we develop a joint model that integrates Mixture of Hawkes Process (MHP) with a hierarchical topic model to capture the arrival sequences with mixed trip purposes. Essentially, the human mobility synchronization effect is encoded as a synchronization rate in the MHP; while the regional environment is modeled by introducing latent Trip Purpose and POI Topic to generate the Point of Interests (POIs) in the regions. Moreover, we provide an effective inference algorithm for parameter learning. Finally, we conduct intensive experiments on synthetic data and real-world data, and the experimental results have demonstrated the effectiveness of the proposed model.

http://dl.acm.org/authorize?N33231

45. Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data

David Hallac (Stanford University);Sagar Vare (Stanford University);Stephen Boyd (Stanford University);Jure Leskovec (Stanford University)

Subsequence clustering of multivariate time series is a useful tool for discovering repeated patterns in temporal data. Once these patterns have been discovered, seemingly complicated datasets can be interpreted as a temporal sequence of only a small number of states, or clusters. For example, raw sensor data from a fitness-tracking application can be expressed as a timeline of a select few actions (i.e., walking, sitting, running). However, discovering these patterns is challenging because it requires simultaneous segmentation and clustering of the time series. Furthermore, interpreting the resulting clusters is difficult, especially when the data is high-dimensional. Here we propose a new method of model-based clustering, which we call Toeplitz Inverse Covariance-based Clustering (TICC). Each cluster in the TICC method is defined by a correlation network, or Markov random field (MRF), characterizing the interdependencies between different observations in a typical subsequence of that cluster. Based on this graphical representation, TICC simultaneously segments and clusters the time series data. We solve the TICC problem through an expectation maximization (EM) algorithm. We derive closed-form solutions to efficiently solve both the E and M-steps in a scalable way, through dynamic programming and the alternating direction method of multipliers (ADMM), respectively. We validate our approach by comparing TICC to several state-of-the-art baselines in a series of synthetic experiments, and we then demonstrate on an automobile sensor dataset how TICC can be used to learn interpretable clusters in real-world scenarios.

http://dl.acm.org/authorize?N33203

46. SPARTan: Scalable PARAFAC2 for Large & Sparse Data

Ioakeim Perros (Georgia Institute of Technology);Evangelos E. Papalexakis (University of California, Riverside);Fei Wang (Weill Cornell Medicine);Richard Vuduc (Georgia Institute of Technology);Elizabeth Searles (Children's Healthcare of Atlanta);Michael Thompson (Children's Healthcare of Atlanta);Jimeng Sun (Georgia Institute of Technology)

In exploratory tensor mining, a common problem is how to analyze a set of variables across a set of subjects whose observations do not align naturally. For example, when modeling medical features across a set of patients, the number and duration of treatments may vary widely in time, meaning there is no meaningful way to align their clinical records across time points for analysis purposes. To handle such data, the state-of-the-art tensor model is the so-called PARAFAC2, which yields interpretable and robust output and can naturally handle sparse data. However, its main limitation up to now has been the lack of efficient algorithms that can handle large-scale datasets. In this work, we fill this gap by developing a scalable method to compute the PARAFAC2 decomposition of large and sparse datasets, called SPARTan. Our method exploits special structure within PARAFAC2, leading to a novel algorithmic reformulation that is both fast (in absolute time) and more memory-efficient than prior work. We evaluate SPARTan on both synthetic and real datasets, showing 22X performance gains over the best previous implementation and also handling larger problem instances for which the baseline fails. Furthermore, we are able to apply SPARTan to the mining of temporally-evolving phenotypes on data taken from real and medically complex pediatric patients. The clinical meaningfulness of the phenotypes identified in this process, as well as their temporal evolution over time for several patients, have been endorsed by clinical experts.

http://dl.acm.org/authorize?N33229

47. Weisfeiler-Lehman Neural Machine for Link Prediction

Muhan Zhang (Washington University in St. Louis);Yixin Chen (Washington University in St. Louis)

In this paper, we propose a next-generation link prediction method, Weisfeiler-Lehman Neural Machine (Wlnm), which learns topological features in the form of graph patterns that promote the formation of links. Wlnm has unmatched advantages including higher performance than state-of-the-art methods and universal applicability over various kinds of networks. Wlnm extracts an enclosing subgraph of each target link and encodes the subgraph as an adjacency matrix. The key novelty of the encoding comes from a fast hashing-based Weisfeiler-Lehman (WL) algorithm that labels the vertices according to their structural roles in the subgraph while preserving the subgraph

http://dl.acm.org/authorize?N33249

48. A Local Algorithm for Structure-Preserving Graph Cut

Dawei Zhou (Arizona State University);Si Zhang (Arizona State University);Mehmet Yigit Yildirim (Arizona State University);Scott Alcorn (Early Warnings LLC.);Hanghang Tong (Arizona State University);Hasan Davulcu (Arizona State University);Jingrui He (Arizona State University)

Nowadays, large-scale graph data is being generated in a variety of real-world applications, from social networks to co-authorship networks, from protein-protein interaction networks to road traffic networks. Many existing works on graph mining focus on the vertices and edges, with first-order Markov chain as the underlying model. They fail to explore the high-order network structures, which are of key importance in many high impact domains. For example, in bank customer personally identifiable information (PII) networks, the star structures often correspond to a set of synthetic identities; in financial transaction networks, the loop structures may indicate the existence of money laundering; in signed networks, the triangle structures play an essential role in the balance theory for edge sign prediction. In this paper, we focus on mining user specified high-order network structures, and aim to find a structure-rich sub-graph which does not break many such structures by separating the sub-graph from the rest. A key challenge associated with finding a structure-rich subgraph is the prohibitive computational cost. To address this problem, inspired by the family of local graph clustering algorithms for efficiently identifying a low conductance cut without exploring the entire graph, we propose to generalize the key idea to model high-order network structures. In particular, we start with a generic definition of high-order conductance and define the high-order diffusion core, which is based on a high-order random walk induced by any user-specified high-order network structures. Then we propose a novel High-Order Structure Preserving Local Clustering algorithm named HOSPLOC, which runs in a near linear time complexity with respect to the number of edges. It starts with a seed node and iteratively explores its neighborhood until it finds a sub-graph with a small high-order conductance. Furthermore, we analyze its performance in terms of both effectiveness and efficiency. The experimental results on both synthetic graphs and real graphs demonstrate the e

http://dl.acm.org/authorize?N33357

49. Linearized GMM Kernels and Normalized Random Fourier Features

Ping Li (Rutgers University)

The method of "random Fourier features (RFF)" has become a popular tool for approximating the "radial basis function (RBF)" kernel. The variance of RFF is actually very large. Interestingly, the variance can be substantially reduced by a simple normalization step as we theoretically demonstrate. We name the improved scheme as the "normalized RFF (NRFF)'', and we provide a technical proof of the theoretical variance of NRFF, as validated by simulations.

We also propose the "generalized min-max (GMM)" kernel as a measure of data similarity, where data vectors can have both positive and negative entries. GMM is positive definite as there is an associate hashing method named "generalized consistent weighted sampling (GCWS)" which linearizes this (nonlinear) kernel. We provide an extensive empirical evaluation of the RBF and GMM kernels on more than 50 datasets. For a majority of the datasets, the (tuning-free) GMM kernel outperforms the best-tuned RBF kernel.

We also conduct extensive experiments for comparing the linearized RBF kernel using NRFF hashing with the linearized GMM kernel using GCWS hashing. We observe that, in order to reach a similarity classification accuracy, GCWS typically requires substantially fewer samples than NRFF, even on datasets where the original RBF kernel outperforms the original GMM kernel. The training, storage, and processing costs are directly proportional to the sample size. Thus, our experiments help demonstrate that GCWS would be a much more practical scheme for large-scale machine learning.

The empirical success of GCWS (compared to NRFF) can also be explained theoretically, from at least two aspects. Firstly, the relative variance (normalized by the squared expectation) of GCWS is substantially smaller than that of NRFF, except for the very high similarity region (where the variances of both methods are close to zero). Secondly, if we make a general model assumption on the data, we can show analytically that GCWS exhibits much smaller variance than NRFF for estimating the same object (e.g., the RBF kernel), except for the high similarity region.

http://dl.acm.org/authorize?N33213

50. Anomaly Detection with Robust Deep Auto-encoders

Chong Zhou (Worcester Polytechnic Institute);Randy Paffenroth (Worcester Polytechnic Institute)

Deep auto-encoders and other deep neural networks have demonstrated their effectiveness in discovering non-linear features across many problem domains. However, in many real-world problems, large outliers and pervasive noise are commonplace and one may not have access to clean training data as required by standard deep denoising auto-encoders. Herein, we demonstrate novel extensions to deep auto-encoders which not only maintain a deep auto-encoders

http://dl.acm.org/authorize?N33358

51. Large-scale Collaborative Ranking in Near-Linear Time

Liwei Wu (University of California, Davis);Cho-Jui Hsieh (University of California, Davis);James Sharpnack (University of California, Davis)

In this paper, we consider the Collaborative Ranking (CR) problem for recommendation systems. Given a set of pairwise preferences between items for each user, collaborative ranking can be used to rank un-rated items for each user, and this ranking can be naturally used for recommendation. It is observed that collaborative ranking algorithms usually achieve better recommendations since it direct minimizes the ranking loss; however, they are rarely used in practice due to the poor scalability. All the existing CR algorithms have time complexity at least $O(|\Omega|r)$ per iteration, where r is the target rank and $|\Omega|$ is number of pairs that grows quadratically with number of ratings per user. For example, the Netflix data contains totally 20 billion rating pairs, and in this scale all the current algorithms have to work on subsamples, resulting in poor prediction on testing data. In this paper, we propose a new collaborative ranking algorithm called Primal-CR that reduces the time complexity to $O(|\Omega| + d_1\bar{d}_2r)$, where d_1 is number of users and $\bar{d}_2$ is the averaged number of items rated by a user. Note that d_1 $\bar{d}_2$ is strictly smaller and often much smaller than $|\Omega|$. Furthermore, by exploiting the fact that most data is in the form of numerical ratings instead of pairwise comparisons, we propose Primal-CR++ with $O(d_1\bar{d}_2(r + \log \bar{d}_2))$ time complexity. Both algorithms have better theoretical time complexity than existing approaches and also outperform existing approaches in terms of NDCG and pairwise error on real data sets. To the best of our knowledge, we are the first one in the collaborative ranking setting to apply the algorithm to the full Netflix dataset using all the 20 billion ratings, and this leads to a model with much better recommendation compared with previous models trained on subsamples. Finally, compared with classical matrix factorization algorithm which also requires $O(d_1\bar{d}_2r)$ time, our algorithm has almost the same efficiency while making much better recommendation since we consider the ranking loss.

http://dl.acm.org/authorize?N33233

52. Efficient Correlated Topic Modeling with Topic Embedding

Junxian He (Carnegie Mellon University);Zhiting Hu (carnegie mellon university);Taylor Berg-Kirkpatrick (carnegie mellon university);Ying Huang (Shanghai Jiaotong University);Eric Xing (carnegie mellon university)

Correlated topic modeling has been limited to small model and problem sizes due to their high computational cost and poor scaling. In this paper, we propose a new model which learns compact topic embeddings and captures topic correlations through the closeness between the topic vectors. Our method enables efficient inference in the low-dimensional embedding space, reducing previous cubic or quadratic time complexity to linear w.r.t the topic size. We further speedup variational inference with a fast sampler to exploit sparsity of topic occurrence. Extensive experiments show that our approach is capable of handling model and data scales which are several orders of magnitude larger than existing correlation results, without sacrificing modeling quality by providing competitive or superior performance in document classification and retrieval.

http://dl.acm.org/authorize?N33204

53. Long Short Memory Process: Modeling Growth Dynamics of Microscopic Social Connectivity

Chengxi Zang (Department of Computer Science, Tsinghua University);Peng Cui (Department of Computer Science, Tsinghua University);Christos Faloutsos (Computer Science Department, Carnegie Mellon University);Wenwu Zhu (Department of Computer Science, Tsinghua University)

How do people make friends dynamically in social networks? What are the temporal patterns for an individual increasing its social connectivity? What are the basic mechanisms governing the formation of these temporal patterns? No matter cyber or physical social systems, their structure and dynamics are mainly driven by the connectivity dynamics of each individual. However, due to the lack of empirical data, little is known about the empirical dynamic patterns of social connectivity at microscopic level, let alone the regularities or models governing these microscopic dynamics. We

examine the detailed growth process of "WeChat", the largest online social network in China, with 300 million users and 4.75 billion links spanning two years. We uncover a wide range of long-term power law growth and short-term bursty growth for the social connectivity of different users. We propose three key ingredients that govern the observed growth patterns at microscopic level. As a result, we propose the long short memory process incorporating these ingredients, demonstrating that it successfully reproduces the complex growth patterns observed in the empirical data. By analyzing modeling parameters, we discover statistical regularities underlying the empirical growth dynamics. Our model and discoveries provide a foundation for the microscopic mechanisms of network growth dynamics, potentially leading to implications for prediction, clustering and outlier detection on human dynamics.

http://dl.acm.org/authorize?N33248

54. Ego-splitting Framework: from Non-Overlapping to Overlapping Clusters

Alessandro Epasto (Google);Silvio Lattanzi (Google);Renato Paes Leme (Google)

We propose a new framework called Ego-Splitting for detecting clusters in complex networks which leverage the local structures known as ego-nets (i.e. the subgraph induced by the neighborhood of each node) to de-couple overlapping clusters. Ego-Splitting is a highly scalable and flexible framework, with provable theoretical guarantees, that reduces the complex overlapping clustering problem to a simpler and more amenable non-overlapping (partitioning) problem. We can solve community detection in graphs with tens of billions of edges and outperform previous solutions based on ego-nets analysis.

More precisely, our framework works in two steps: a local ego-net analysis phase, and a global graph partitioning phase . In the local step, we first partition the nodes' ego-nets using a partitioning algorithm. We then use the computed clusters to split each node into its persona nodes that represent the instantiations of the node in its communities. Then, in the global step, we partition the newly created graph to obtain an overlapping clustering of the original graph.

http://dl.acm.org/authorize?N33206

55. Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking

Gabriele Tolomei (Yahoo);Fabrizio Silvestri (Facebook);Andrew Haines (Yahoo Inc);Mounia Lalmas (Yahoo)

Machine-learned models are often described as "black boxes". In many real-world applications, however, models may have to sacrifice predictive power in favour of human-interpretability. When this is the case, feature engineering becomes a crucial albeit expensive task, requiring manual and time-consuming analysis. In addition, whereas some features are inherently static as they represent properties that are fixed (e.g., the age of an individual), other capture characteristics that could be adjusted (e.g., the daily amount of carbohydrates taken). Nonetheless, once a model is learned from the data, each prediction it makes on new instances is irreversible, therefore assuming every instance to be a static point located in the chosen feature space. There are many circumstances, instead, where it is important to understand (i) why a model outputs a certain prediction on a given instance, (ii) which adjustable features of that instance should be modi

http://dl.acm.org/authorize?N33238

56. Contextual Motifs: Increasing the Utility of Motifs using Contextual Data

Ian Fox (CSE, University of Michigan);Lynn Ang (Department of Internal Medicine, Division of Metabolism, Endocrinology and Diabetes, University of Michigan);Mamta Jaiswal (Department of Internal Medicine, Division of Metabolism, Endocrinology and Diabetes, University of Michigan);Rodica Pop-Busui (Department of Internal Medicine, Division of Metabolism, Endocrinology and Diabetes, University of Michigan);Jenna Wiens (CSE, University of Michigan)

Motifs are a powerful tool for analyzing long physiological signals. Standard motif methods, however, ignore important contextual information (e.g., what the patient was doing at the time the data were collected). We hypothesize that these additional contextual data could increase the utility of motifs. Thus, we propose an extension to motifs, *contextual motifs*, that incorporates context. We present methods to discover contextual motifs, both with observed and inferred contextual information. Oftentimes, we may not observe context, or collecting context may simply be too burdensome. In this setting, we present methods to jointly infer motifs and context. Through experiments on simulated data, we illustrate the potential discriminative power of contextual motifs across a range of settings, improving discriminative performance, measured using AUROC, by up to 11 percentage points over a contextless baseline. We further validate the proposed approach on a dataset of continuous glucose monitor data collected from type 1 diabetics. Applied to the task of predicting hypo- and hyper-glycemic events, use of contextual motifs led to a 7.2 percentage point improvement in AUROC compared with a state-of-the-art baseline.

http://dl.acm.org/authorize?N33207

57. Structural Deep Brain Network Mining

Shen Wang (University of Illinois at Chicago);Lifang He (University of Illinois at Chicago);Bokai Cao (University of Illinois at Chicago);Chun-Ta Lu (University of Illinois at Chicago);Philip Yu (University of Illinois at Chicago);Ann Ragin (Northwestern University)

Mining from neuroimaging data is becoming increasingly popular in the field of healthcare and bioinformatics, due to its potential to discover clinically meaningful structure patterns that could facilitate the understanding and diagnosis of neurological and neuropsychiatric disorders. Most recent researches concentrate on applying subgraph mining techniques to discover connected subgraph patterns in the brain network. However, the underlying brain network structure is complicated. As a shallow linear model, subgraph mining can not capture the highly non-linear structures, resulting in sub-optimal patterns. Therefore, how to learn representations that can capture the highly non-linearity of brain networks and preserve the underlying structures is a critical problem. In this paper, we propose a Structural Deep Brain Network mining method, namely SDBN, to learn highly non-linear and structure-preserving representations of brain networks. Specifically, we first introduce a novel graph normalization approach based on module identification, which rearranges the order of the nodes to preserve the modular structure of the graph.Next, we perform structural augmentation to further enhance the spatial information of the normalized graph. Then we propose a deep feature learning framework for combining supervised learning and unsupervised learning in a small-scale setting, by augmenting Convolutional Neural Network (CNN) with decoding pathways for reconstruction. With the help of the multiple layers of non-linear mapping, the proposed SDBN approach can capture the highly non-linear structure of brain networks. Besides, it has better generalization capability for high-dimensional brain network and is suitable for small sample learning. To evaluate the proposed SDBN method , we conduct extensive experiments on four real brain network datasets for disease diagnoses. The experiment results show that SDBN can capture discriminative and meaningful structural graph representations for brain disorder diagnosis.

http://dl.acm.org/authorize?N33239

58. Randomized Feature Engineering as a Fast and Accurate Alternative to Kernel Methods

Suhang Wang (Arizona State University);Charu Aggarwal (IBM);Huan Liu (Professor, Arizona State University)

Feature engineering has found increasing interest in recent years because of its ability to improve the effectiveness of various machine learning models. Although tailored feature engineering methods have been designed for various domains, there are few that simulate the consistent effectiveness of kernel methods. At the core, the success of kernel methods is achieved by using similarity functions that emphasize local variations in similarity. Unfortunately, this ability comes at the price of the high level of computational resources required and the inflexibility of the representation as it only provides the similarity of two data points instead of vector representations of each data point; while

the vector representations can be readily used as input to facilitate various models for different tasks. Furthermore, kernel methods are also highly susceptible to overfitting and noise and it cannot capture the variety of data locality. In this paper, we first analyze the inner working and weaknesses of kernel method, which serves as guidance for designing feature engineering. With the guidance, we explore the use of randomized methods for feature engineering by capturing multi-granular locality of data. This approach has the merit of being time and space efficient for feature construction. Furthermore, the approach is resistant to overfitting and noise because the randomized approach naturally enables fast and robust ensemble methods. Extensive experiments on a number of real world datasets are conducted to show the effectiveness of the approach for various tasks such as clustering, classification and outlier detection.

http://dl.acm.org/authorize?N33230

59. struc2vec: Learning Node Representations from Structural Identity

Leonardo F. R. Ribeiro (Federal University of Rio de Janeiro), Pedro H. P. Saverese (Federal University of Rio de Janeiro), Daniel R. Figueiredo (Federal University of Rio de Janeiro)

Structural identity is a concept of symmetry in which network nodes are identified according to the network structure and their relationship to other nodes. Structural identity has been studied in theory and practice over the past decades, but has only recently been addressed with techniques from representational learning. This work presents struc2vec, a novel and flexible framework for learning latent representations of node's structural identity. struc2vec assesses structural similarity without using node or edge attributes, uses a hierarchy to measure similarity at different scales, and constructs a multilayer graph to encode the structural similarities and generate structural context for nodes. Numerical experiments indicate that state-of-the-art techniques for learning node representations fail in capturing stronger notions of structural identity, while struc2vec exhibits much superior performance in this task, as it overcomes limitations of prior techniques.

http://dl.acm.org/authorize?N33220

60. Network Inference via the Time-Varying Graphical Lasso

David Hallac (Stanford University);Youngsuk Park (Stanford University);Stephen Boyd (Stanford University);Jure Leskovec (Stanford University)

Many important problems can be modeled as a system of interconnected entities, where each entity is recording time-dependent observations or measurements. In order to spot trends, detect anomalies, and interpret the temporal dynamics of such data, it is essential to understand the relationships between the different entities and how these relationships evolve over time. In this paper, we introduce the time-varying graphical lasso (TVGL), a method of inferring time-varying networks from raw time series data. We cast the problem in terms of estimating a sparse time-varying inverse covariance matrix, which reveals a dynamic network of interdependencies between the entities. Since dynamic network inference is a computationally expensive task, we derive a scalable message-passing algorithm based on the Alternating Direction Method of Multipliers (ADMM) to solve this problem in an efficient way. We also discuss several extensions, including a streaming algorithm to update the model and incorporate new observations in real time. Finally, we evaluate our TVGL algorithm on both real and synthetic datasets, obtaining interpretable results and outperforming state-of-the-art baselines in terms of both accuracy and scalability.

http://dl.acm.org/authorize?N33202

61. PReP: Path-Based Relevance from a Probabilistic Perspective in Heteogeneous Information Networks

Yu Shi (Dept. of Computer Science, University of Illinois at Urbana-Champaign);Po-Wei Chan (Dept. of Computer Science, University of Illinois at Urbana-Champaign);Honglei Zhuang (Dept. of Computer Science, University of Illinois at Urbana-Champaign);Huan Gui (Dept. of Computer Science, University of Illinois at Urbana-Champaign);Jiawei Han (Dept. of Computer Science, University of Illinois at Urbana-Champaign)

As a powerful representation paradigm for networked and multi-typed data, the heterogeneous information network (HIN) is ubiquitous. Meanwhile, defining proper relevance measures has always been a fundamental problem and of great pragmatic importance for network mining tasks. Inspired by the probabilistic interpretation of existing path-based relevance measures, we propose to study HIN relevance from a probabilistic perspective. We also identify, from real-world data, and propose to model cross-meta-path synergy, which is a characteristic important for defining path-based HIN relevance and has not been modeled by existing methods. A generative model is established to derive a novel path-based relevance measure, which is data-driven and tailored for each HIN. We develop an inference algorithm to find the maximum a posteriori (MAP) estimate of the model parameters, which entails non-trivial tricks. Experiments on two real-world datasets demonstrate the effectiveness of the proposed model and relevance measure.

http://dl.acm.org/authorize?N33224

62. The Co-Evolution Model for Social Network Evolving and Opinion Migration

Yupeng Gu (University of California, Los Angeles);Yizhou Sun (University of California, Los Angeles);Jianxi Gao (Northeastern University)

Almost all real-world social networks are dynamic and evolving with time, where new links may form and old links may drop, largely determined by the homophily of social actors (i.e., nodes in the network). Meanwhile, (latent) properties of social actors, such as their opinions, are changing along the time, partially due to social influence received from the network, which will in turn affect the network structure. Social network evolution and node property migration are usually treated as two orthogonal problems, and have been studied separately. In this paper, we propose a co-evolution model that closes the loop by modeling the two phenomena together, which contains two major components: (1) a network generative model when the node property is known; and (2) a property migration model when the social network structure is known. Simulation shows that our model has several nice properties: (1) it can model a broad range of phenomena such as opinion convergence (i.e., herding) and community-based opinion divergence; and (2) it allows us control the evolution via a set of factors such as social influence scope, opinion leader, and noise level. Finally, the usefulness of our model is demonstrated by an application of co-sponsorship prediction for legislative bills in Congress, which outperforms several state-of-the-art baselines.

http://dl.acm.org/authorize?N33209

63. Improved Degree Bounds and Full Spectrum Power Laws in Preferential Attachment Networks

Chen Avin (Ben Gurion University of the Negev);Zvi Lotker (Ben Gurion University of the Negev);Yinon Nahum (Weizmann Institute of Science);David Peleg (Weizmann Institute of Science)

Consider a random preferential attachment model $G(p)$ for network evolution that allows both node and edge arrivals. Starting with an arbitrary nonempty graph G_0, at each time step, there are two possible events: with probability $p > 0$ a new node arrives and a new edge is added between the new node and an existing node, and with probability $1-p$ a new edge is added between two existing nodes. In both cases, the involved existing nodes are chosen at random according to preferential attachment, i.e., with probability proportional to their degree. $G(p)$ is known to generate *power law networks*, i.e., the fraction of nodes with degree k is proportional to $k^{-\beta}$. Here $\beta = (4-p)/(2-p)$ is in the range $[2, 3]$.

Denoting the number of nodes of degree k at time t by $m_{k,t}$, we significantly improve some long-standing results. In particular, we show that $m_{k,t}$ is concentrated around its mean with a deviation of $O(\sqrt{t})$, which is independent of k. We also tightly bound the expectation $Em_{k,t}$ with an additive error of $O(1/k)$, which is independent of t. These new bounds allow us to tightly estimate $m_{k,t}$ for a considerably larger k values than before. This, in turn, enables us to estimate other important quantities, e.g., the size of the k-rich club, namely, the set of all nodes with a degree at least k.

Finally, we introduce a new generalized model, $G(p_t, r_t, q_t)$, which extends $G(p)$ by allowing also *time-varying* probabilities for node and edge arrivals, as well as the formation of new components.

We show that the extended model can produce power law networks with any exponent β in the range $(1, \infty)$. Furthermore, the concentration bounds established for $m_{k,t}$ in $G(p)$ also apply in $G(p_t, r_t, q_t)$.

http://dl.acm.org/authorize?N33296

64. A Framework for Guided Time Series Motif Discovery

Hoang Anh Dau (University of California, Riverside);Eamonn Keogh (University of California, Riverside)

Time series motif discovery has emerged as perhaps the most used primitive for time series data mining, and has seen applications to domains as diverse as robotics, medicine and climatology. There has been recent significant progress on the scalability of motif discovery. However, we believe that the current definitions of motif discovery are limited, and can create a mismatch between the user

http://dl.acm.org/authorize?N33294

65. Constructivism Learning: A Learning Paradigm for Transparent Predictive Analytics

Xiaoli Li (University of Kansas);Jun Huan (University of Kansas)

Developing transparent predictive analytics has attracted significant research attention recently. There have been multiple theories on how to model learning transparency but none of them aims to understand the internal and often complicated modeling processes. In this paper we adopt a contemporary philosophical concept called “constructivism‘, which is a theory regarding how human learns. We hypothesis that a critical aspect of transparent machine learning is to “reveal” model construction with two key process: (1) the assimilation process where we enhance our existing learning models and (2) the accommodation process where we create new learning models. With this intuition we propose a new learning paradigm using a Bayesian nonparametric to dynamically handle the creation of new learning tasks. Our empirical study on both synthetic and real data sets demonstrate that the new learning algorithm is capable of delivering higher quality models (as compared to base lines and state-of-the-art) and at the same time increasing the transparency of the learning process.

http://dl.acm.org/authorize?N33210

66. HyperLogLog Hyper Extended: Sketches for Concave Sublinear Frequency Statistics

Edith Cohen (Google Research)

One of the most common statistics computed over data elements is the number of distinct keys. A thread of research pioneered by Flajolet and Martin three decades ago culminated in the design of optimal approximate counting sketches, which have size that is double logarithmic in the number of distinct keys and provide estimates with a small relative error. Moreover, the sketches are composable, and thus suitable for streamed, parallel, or distributed computation.

We consider here all statistics of the frequency distribution of keys, where a contribution of a key to the aggregate is concave and grows (sub)linearly with its frequency. These fundamental aggregations are very common in text, graphs, and logs analysis and include logarithms, low frequency moments, and capping statistics.

We design composable sketches of double-logarithmic size for all concave sublinear statistics. Our design combines theoretical optimality and practical simplicity. In a nutshell, we specify tailored mapping functions of data elements to output elements so that our target statistics on the data elements is approximated by the (max-) distinct statistics of the output elements, which can be approximated using off-the-shelf sketches. Our key insight is relating these target statistics to the *complement Laplace* transform of the input frequencies.

http://dl.acm.org/authorize?N33292

67. Fast Enumeration of Large k-Plexes

Alessio Conte (University of Pisa);Donatella Firmani (Roma Tre University);Caterina Mordente (Be Think Solve Execute);Maurizio Patrignani (Roma Tre University);Riccardo Torlone (Roma Tre University)

The k-plex is a formal yet flexible way of defining communities in any kind of network. It generalizes the notion of clique and is more appropriate in most real cases: while a node of a clique C is connected to all other nodes of C, a node of a k-plex may miss k connections. Unfortunately, computing all maximal k-plexes is a gruesome task and state of the art algorithms can only process small-size networks. In this paper we propose a new approach for enumerating large k-plexes in networks that speeds up the search of several orders of magnitude by leveraging: (i) methods for strongly reducing the search space and (ii) efficient techniques for the computation of maximal cliques. Several experiments show that our strategy is effective and is able to increase the size of the networks for which the computation of large k-plexes is feasible from a few hundred to several hundred thousand nodes.

http://dl.acm.org/authorize?N33293

68. On Finding Socially Tenuous Groups for Online Social Networks

Chih-Ya Shen (National Tsing Hua University);Liang-Hao Huang (Academia Sinica);De-Nian Yang (Academia Sinica);Hong-Han Shuai (National Chiao Tung University);Wang-Chien Lee (The Pennsylvania State University);Ming-Syan Chen (National Taiwan University)

Existing research on finding social groups mostly focuses on dense subgraphs in social networks. However, finding socially tenuous groups also has many important applications. In this paper, we introduce the notion of k-triangles to measure the tenuity of a group. We then formulate a new research problem, Minimum k-Triangle Disconnected Group (MkTG), to find a socially tenuous group from online social networks. We prove that MkTG is NP-Hard and inapproximable within any ratio in arbitrary graphs but polynomial-time tractable in threshold graphs. Two algorithms, namely TERA and TERA-ADV, are designed to exploit graph-theoretical approaches for solving MkTG on general graphs effectively and efficiently. Experimental results on seven real datasets manifest that the proposed algorithms outperform existing approaches in both efficiency and solution quality.

http://dl.acm.org/authorize?N33223

69. A Parallel and Primal-Dual Sparse Method for Extreme Classification

Ian Yen (Carnegie Mellon University);Xiangru Huang (University of Texas at Austin);Wei Dai (Carnegie Mellon University);Pradeep Ravikumar (Carnegie Mellon University);Inderjit Dhillon (University of Texas at Austin);Eric Xing (Carnegie Mellon University)

Extreme Classification considers the problem of multiclass or multilabel prediction when there is a huge number of classes: a scenario that occurs in many real-world applications such as text and image tagging. In this setting, standard classification methods with complexity linear to the number of classes become intractable, while enforcing structural constraints among classes (such as low-rank or tree-structured) to reduce complexity often sacrifices accuracy for efficiency. The recent *PD-Sparse* method addresses this issue to gives an algorithm that is sublinear in the number of variables by exploiting *primal-dual* sparsity inherent in the max-margin loss. However, the objective requires training models of all classes together, which incurs large memory consumption and prohibits it from the simple parallelization scheme that a one-versus-all method can easily take advantage of. In this work, we propose a primal-dual sparse method that enjoys the same parallelizability and space efficiency of one-versus-all approach, while having complexity sublinear to the number of classes. On several large-scale benchmark data sets, the proposed method achieves accuracy competitive to state-of-the-art methods while reducing training time from days to tens of minutes compared to existing parallel or sparse methods on a cluster of 100 cores.

http://dl.acm.org/authorize?N33246

70. Incremental Dual-memory LSTM in Land Cover Prediction

Xiaowei Jia (University of Minnesota);Ankush Khandelwal (University of Minnesota);Guruprasad Nayak (University of Minnesota);James Gerber (University of Minnesota);Kimberly Carlson (University of Hawaii Manoa);Paul West (University of Minnesota);Vipin Kumar (University of Minnesota)

Land cover prediction is essential for monitoring global environmental change. Unfortunately, traditional classification models are plagued by temporal variation and emergence of novel/unseen land cover classes in the prediction process. In this paper, we propose an LSTM-based spatio-temporal learning framework with a dual-memory structure. The dual-memory structure captures both long-term and short-term temporal variation patterns, and is updated incrementally to adapt the model to the ever-changing environment. Moreover, we integrate zero-shot learning to identify unseen classes even without labelled samples. Experiments on both synthetic and real-world datasets demonstrate the superiority of the proposed framework over multiple baselines in land cover prediction.

http://dl.acm.org/authorize?N33377

71. An Alternative to NCD for Large Sequences, Lempel-Ziv Jaccard Distance

Edward Raff (Laboratory for Physical Sciences);Charles Nicholas (University of Maryland Baltimore County)

The Normalized Compression Distance (NCD) has been used in a number of domains to compare objects with varying feature types. This flexibility comes from the use of general purpose compression algorithms as the means of computing distances between byte sequences. Such flexibility makes NCD particularly attractive for cases where the right features to use are not obvious, such as malware classification. However, NCD can be computationally demanding, thereby restricting the scale at which it can be applied. We introduce an alternative metric also inspired by compression, the Lempel-Ziv Jaccard Distance (LZJD). We show that this new distance has desirable theoretical properties, as well as comparable or superior performance for malware classification, while being easy to implement and orders of magnitude faster in practice.

http://dl.acm.org/authorize?N33381

72. Distributed Local Outlier Detection in Big Data

Yizhou Yan (Worcester Polytechnic Institute);Lei Cao (Massachusetts Institute of Technology);Caitlin Kuhlman (Worcester Polytechnic Institute);Elke Rundensteiner (Worcester Polytechnic Institute)

In this work, we present the first distributed solution for the Local Outlier Factor (LOF) method—a popular outlier detection technique shown to be very effective for datasets with skewed distributions. As datasets increase radically in size, highly scalable LOF algorithms leveraging modern distributed infrastructures are required. This poses significant challenges due to the complexity of the LOF definition, and a lack of access to the entire dataset at any individual compute machine. Our solution features a distributed LOF pipeline framework, called DLOF. Each stage of the LOF computation is conducted in a fully distributed fashion by leveraging our invariant observation for intermediate value management. Furthermore, we propose a data assignment strategy which ensures that each machine is self-sufficient in all stages of the LOF pipeline, while minimizing the number of data replicas. Based on the convergence property derived from analyzing this strategy in the context of real world datasets, we introduce a number of data-driven optimization strategies. These strategies not only minimize the computation costs within each stage, but also eliminate unnecessary communication costs by aggressively pushing the LOF computation into the early stages of the DLOF pipeline. Our comprehensive experimental study using both real and synthetic datasets confirms the efficiency and scalability of our approach to terabyte level data.

http://dl.acm.org/authorize?N33303

73. Recurrent Poisson Factorization for Temporal Recommendation

Seyed Abbas Hosseini (Sharif University of Technology);Keivan Alizadeh (Sharif University of Technology);Ali Khodadadi (Sharif University of Technology);Ali Arabzadeh (Sharif University of Technology);Mehrdad Farajtabar (Georgia Institute of Technology);Hongyuan Zha (Georgia Institute of Technology);Hamid R. Rabiee (Sharif University of Technology)

Poisson factorization is a probabilistic model of users and items for recommendation systems, where the so-called implicit consumer data is modeled by a factorized Poisson distribution. There are many variants of Poisson factorization methods who show state-of-the-art performance on real-world recommendation tasks. However, most of them do not explicitly take into account the temporal behavior and recurrent activities of users which is essential to recommend the right item to the right user at the right time. In this paper, we introduce Recurrent Poisson Factorization (RPF) framework that generalizes the classical PF methods by utilizing a Poisson process for modeling the implicit feedback.

RPF treats time as a natural constituent of the model and offers a rich family of time-sensitive factorization models. To elaborate, we instantiate several variants of RPF who are capable of handling dynamic user preferences and item specification (DRPF), modeling the social-aspect of product adoption (SRPF), and capturing the consumption heterogeneity among users and items (HRPF). We also develop a variational algorithm for approximate posterior inference that scales up to massive data sets. Furthermore, we demonstrate RPF's superior performance over many state-of-the-art methods on synthetic dataset, and large scale real-world datasets on music streaming logs, and user-item interactions in M-Commerce platforms.

http://dl.acm.org/authorize?N33375

74. Detecting Network Effects: Randomizing Over Randomized Experiments

Martin Saveski (MIT);Jean Pouget-Abadie (Harvard University);Guillaume Saint-Jacques (MIT);Weitao Duan (LinkedIn);Souvik Ghosh (LinkedIn);Ya Xu (LinkedIn);Edo Airoldi (Harvard University)

Randomized experiments—A/B tests—are the standard approach for evaluating the effect of new product features. They rely on the "stable treatment value assumption" (SUTVA) which states that treatment only affects treated users and does not spill over to their friends. Violations of SUTVA, common in features that exhibit network effects, result in inaccurate estimates of the treatment effect. In this paper, we leverage a new experimental design for testing whether SUTVA holds, without making any assumptions on how treatment effects may spill over between the treatment and the control group. We do so by simultaneously running completely randomized and cluster-based randomized experiments and comparing the difference of resulting estimates, detailing known theoretical bounds on the Type I error rate. We provide practical guidelines for implementing this design on large-scale experimentation platforms. Finally, we deploy this design to LinkedIn's experimentation platform and apply it to two online experiments, highlighting the presence of network effects and bias in standard A/B testing approaches in a "real-world" setting.

http://dl.acm.org/authorize?N33383

75. Unsupervised Discovery of Drug Side-Effects From Heterogeneous Data Sources

Fenglong Ma (SUNY Buffalo);Chuishi Meng (SUNY Buffalo);Houping Xiao (SUNY Buffalo);Qi Li (SUNY Buffalo);Jing Gao (SUNY Buffalo);Lu Su (SUNY Buffalo);Aidong Zhang (SUNY Buffalo)

Drug side-effects have become a worldwide public health concern, which are the fourth leading cause of death in the United States. Pharmaceutical industry has paid tremendous efforts to identify drug side-effects during the drug development. However, it is impossible and impractical to identify all of them. Fortunately, drug side-effects can also be reported on heterogeneous data sources, such as FDA Adverse Event Reporting System and various online communities. However, existing supervised and semi-supervised approaches are not practical as annotating labels are expensive in the medical field. In this paper, we propose a novel and effective unsupervised model Sifter to automatically discover drug side-effects. Sifter enhances the estimation on drug side-effects by learning

from various online platforms and measuring platform-level and user-level quality simultaneously. In this way, Sifter demonstrates better performance compared with existing approaches in terms of correctly identifying drug side-effects. Experimental results on five real-world datasets show that Sifter can significantly improve the performance of identifying side-effects compared with the state-of-the-art approaches.

http://dl.acm.org/authorize?N33387

76. Federated Tensor Factorization for Computational Phenotyping

Yejin Kim (POSTECH);Jimeng Sun (Georgia Institute of Technology);Hwanjo Yu (POSTECH);Xiaoqian Jiang (University of California San Diego)

Tensor factorization models offer an effective approach to convert massive electronic health records into meaningful clinical concepts (phenotypes) for data analysis. These models need a large amount of diverse samples to avoid population bias. An open challenge is how to derive phenotypes jointly across multiple hospitals, in which direct patient-level data sharing is not possible (e.g., due to institutional policies). In this paper, we developed a novel solution to enable federated tensor factorization for computational phenotyping without sharing patient-level data. We developed secure data harmonization and federated computation procedures based on alternating direction method of multipliers (ADMM). Using this method, the multiple hospitals iteratively update tensors and transfer secure summarized information to a central server, and the server aggregates the information to generate phenotypes. We demonstrated with real medical datasets that our method resembles the centralized training model (based on combined datasets) in terms of accuracy and phenotypes discovery while respecting privacy.

http://dl.acm.org/authorize?N33379

77. Optimized Risk Scores

Berk Ustun (Massachusetts Institute of Technology);Cynthia Rudin (Duke University)

Risk scores are simple classification models that let users quickly assess risk by adding, subtracting, and multiplying a few small numbers. Such models are widely used in healthcare and criminology, but are still built ad hoc. In this paper, we present a new approach to learn risk scores that are fully optimized for feature selection, integer coefficients, and operational constraints. We formulate the risk score problem as a mixed integer nonlinear program, and present a new cutting plane algorithm to efficiently recover its optimal solution. Our approach can learn optimized risk scores in a way that scales linearly in the sample size of a dataset, provides a proof of optimality, and accommodates complex constraints without parameter tuning. We illustrate these benefits by building a customized risk score for ICU seizure prediction, as well as an extensive set of numerical experiments.

http://dl.acm.org/authorize?N33393

78. DenseAlert: Incremental Dense-Subtensor Detection in Tensor Streams

Kijung Shin (Carnegie Mellon University);Bryan Hooi (Carnegie Mellon University);Jisu Kim (Carnegie Mellon University);Christos Faloutsos (Carnegie Mellon University)

Consider a stream of retweet events - how can we spot fraudulent lock-step behavior in such multi-aspect data (i.e., tensors) evolving over time? Can we detect it in real time, with an accuracy guarantee? Past studies have shown that dense subtensors tend to indicate anomalous or even fraudulent behavior in many tensor data including social media, Wikipedia, and TCP dumps. Thus, several approaches have been proposed for detecting dense subtensors rapidly and accurately. However, all these methods assume static tensors, while tensors evolve over time in many real-world applications such as social media and web. We propose DenseStream, an incremental algorithm that maintains and updates dense subtensors in a tensor stream (i.e., sequences of changes in a tensor), and DenseAlert, an incremental algorithm spotting the sudden appearances of dense subtensors. Our methods are: (1) Fast and

http://dl.acm.org/authorize?N33396

79. DeepMood: Modeling Mobile Phone Typing Dynamics for Mood Detection

Bokai Cao (University of Illinois at Chicago);Lei Zheng (University of Illinois at Chicago);Chenwei Zhang (University of Illinois at Chicago);Philip S. Yu (University of Illinois at Chicago);Andrea Piscitello (University of Illinois at Chicago);John Zulueta (University of Illinois at Chicago);Olu Ajilore (University of Illinois at Chicago);Kelly Ryan (University of Michigan);Alex Leow (University of Illinois at Chicago)

The increasing use of electronic forms of communication presents new opportunities in the study of mental health, including the ability to investigate the manifestations of psychiatric diseases unobtrusively and in the setting of patients' daily lives. A pilot study to explore the possible connections between bipolar affective disorder and mobile phone usage was conducted. In this study, participants were provided a mobile phone to use as their primary phone. This phone was loaded with a custom keyboard that collected metadata consisting of keypress entry time and accelerometer movement. Individual character data with the exceptions of the backspace key and space bar were not collected due to privacy concerns. We propose an end-to-end deep architecture based on late fusion, named DeepMood, to model the multi-view metadata for the prediction of mood scores. Experimental results show that 90.31

http://dl.acm.org/authorize?N33365

80. Construction of Directed 2K Graphs

Balint Tillman (University of California, Irvine);Athina Markopoulou (University of California, Irvine);Carter T. Butts (University of California, Irvine);Minas Gjoka (Google)

We study the problem of constructing synthetic graphs that resemble real-world directed graphs in terms of their degree correlations. We define the problem of directed 2K construction (D2K) that takes as input the directed degree sequence (DDS) and a joint degree and attribute matrix (JDAM) so as to capture degree correlation specifically in directed graphs. We provide necessary and sufficient conditions to decide whether a target D2K is realizable, and we design an efficient algorithm that creates realizations with that target D2K. We evaluate our algorithm in creating synthetic graphs that target real-world directed graphs (such as Twitter) and we show that it brings significant benefits compared to state-of-the-art approaches.

http://dl.acm.org/authorize?N33392

81. Prospecting the Career Development of Talents: A Survival Analysis Perspective

Huayu Li (UNC Charlotte);Yong Ge (University of Arizona);Hengshu Zhu (Baidu Talent Intelligence Center);Hui Xiong (Rutgers University);Hongke Zhao (University of Sci. & Tech. of China)

The study of career development has become more important during a time of rising competition. Even with the help of newly available big data in the field of human resources, it is challenging to prospect the career development of talents in an effective manner, since the nature and structure of talent careers can change quickly. To this end, in this paper, we propose a novel survival analysis approach to model the talent career paths, with a focus on two critical issues in talent management, namely turnover and career progression. Specifically, for modeling the talent turnover behaviors, we formulate the prediction of survival status at a sequence of time intervals as a multi-task learning problem by considering the prediction at each time interval as a task. Also, we impose the ranking constraints to model both censored and uncensored data, and capture the intrinsic properties exhibited in general lifetime modeling with non-recurrent and recurrent events. Similarly, for modeling the talent career progression, each task concerns the prediction of a relative occupational level at each time interval. The ranking constraints imposed on different occupational levels can help to reduce the prediction error. Finally, we evaluate our approach with several state-of-the-art baseline methods on real-world talent data. The experimental results clearly demonstrate the effectiveness of the proposed models for predicting the turnover and career progression of talents.

http://dl.acm.org/authorize?N33372

82. Anarchists, Unite: Practical Entropy Approximation for Distributed Streams

Moshe Gabel (Technion);Daniel Keren (Haifa University);Assaf Schuster (Technion)

Entropy is a fundamental property of data and a key metric in many scientific and engineering fields. Entropy estimation has been extensively studied, but almost always under the assumption that there is a single data stream, seen in its entirety by one node running the estimation algorithm. Multiple distributed data sources are becoming increasingly common, however, with applications in signal processing, computer science, medicine, physics, and more. Centralizing all data can be infeasible, for example in networks of battery or bandwidth limited sensors, so entropy estimation in distributed streams requires new, communication-efficient approaches. We propose a practical communication-efficient algorithm for continuously approximating the entropy of distributed streams, with deterministic, user-defined error bounds. Unlike previous streaming methods, it supports deletions and variable-sized time-based sliding windows, while still avoiding communication when possible. Moreover, it optionally incorporates a state-of-the-art entropy sketch, allowing for both bandwidth reduction and monitoring very high dimensional problems. Finally, it provides the approximation to all nodes, rather than to a centralized location, which is important in settings such as wireless sensor networks. Evaluation on several public datasets from real application domains shows that our adaptive algorithm can reduce the number of messages by up to three orders of magnitude compared to centralizing all data in one node.

http://dl.acm.org/authorize?N33364

83. Visualizing Attributed Graphs via Terrain Metaphor

Yang Zhang (The Ohio State University);Yusu Wang (The Ohio State University);Srinivasan Parthasarathy (The Ohio State University)

The value proposition of a dataset often resides in the implicit interconnections or explicit relationships (patterns) among individual entities, and is often modeled as a graph. Effective visualization of such graphs can lead to key insights uncovering such value. In this article we propose a visualization method to explore attributed graphs with numerical attributes associated with nodes (or edges). Such numerical attributes can represent raw content information, similarities, or derived information reflecting important network measures such as triangle density and centrality. The proposed visualization strategy seeks to simultaneously uncover the relationship between attribute values and graph topology, and relies on transforming the network to generate a terrain map. A key objective here is to ensure that the terrain map reveals the overall distribution of components-of-interest (e.g. dense subgraphs, k-cores) and the relationships among them while being sensitive to the attribute values over the graph. We also design extensions that can capture the relationship across multiple numerical attributes. We demonstrate the efficacy of our method on several real-world data science tasks while scaling to large graphs with millions of nodes.

http://dl.acm.org/authorize?N33313

84. SPOT: Sparse Optimal Transformations for High Dimensional Variable Selection and Exploratory Regression Analysis

Qiming Huang (Purdue University);Michael Zhu (Purdue Univeristy)

We develop a novel method called SParse Optimal Transformations (SPOT) to simultaneously select important variables and explore relationships between the response and predictor variables in high dimensional nonparametric regression analysis. Not only are the optimal transformations identified by SPOT interpretable, they can also be used for response prediction. We further show that SPOT achieves consistency in both variable selection and parameter estimation. Numerical experiments and real data applications demonstrate that SPOT outperforms other existing methods and can serve as an effective tool in practise.

http://dl.acm.org/authorize?N33376

85. Algorithmic decision making and the cost of fairness

Sam Corbett-Davies (Stanford University);Emma Pierson (Stanford University);Avi Feller (University of California, Berkeley);Sharad Goel (Stanford University);Aziz Huq (University of Chicago)

Algorithms are now regularly used to decide whether defendants awaiting trial are too dangerous to be released back into the community. In some cases, black defendants are substantially more likely than white defendants to be incorrectly classified as high risk. To mitigate such disparities, several techniques recently have been proposed to achieve algorithmic fairness. Here we reformulate algorithmic fairness as constrained optimization: the objective is to maximize public safety while satisfying formal fairness constraints designed to reduce racial disparities. We show that for several past definitions of fairness, the optimal algorithms that result require detaining defendants above race-specific risk thresholds. We further show that the optimal unconstrained algorithm requires applying a single, uniform threshold to all defendants. The unconstrained algorithm thus maximizes public safety while also satisfying one important understanding of equality: that all individuals are held to the same standard, irrespective of race. Because the optimal constrained and unconstrained algorithms generally differ, there is tension between improving public safety and satisfying prevailing notions of algorithmic fairness. By examining data from Broward County, Florida, we show that this trade-off can be large in practice. We focus on algorithms for pretrial release decisions, but the principles we discuss apply to other domains, and also to human decision makers carrying out structured decision rules.

http://dl.acm.org/authorize?N33360

86. LEAP: Learning to Prescribe Effective and Safe Treatment Combinations for Multimorbidity

Yutao Zhang (Tsinghua University);Robert Chen (Georgia Institute of Technology);Jie Tang (Tsinghua University);Jimeng Sun (Georgia Institute of Technology);Walter Stewart (Sutter Health)

Managing patients with complex multimorbidity has long been recognized as a difficult problem due to complex disease and medication dependencies and the potential risk of adverse drug interactions. Existing work either uses complicated hard-coded protocols which are hard to implement and maintain, or use simple statistical models that treat each disease independently which might lead to sub-optimal or even harmful drug combinations. In this work, we propose the LEAP (LEArn to Prescribe) algorithm to decompose the treatment recommendation into a sequential decision making process while automatically determines the appropriate number of medications. A recurrent decoder is used to model label dependencies and content-based attention is used to capture label instance mapping. We further leverage reinforcement learning to fine tune the model parameters to ensure accuracy and completeness. We incorporate external clinical knowledge into the design of the reinforcement reward to effectively prevent generating unfavorable drug combinations. Both quantitative experiments and qualitative case studies are conducted on two real world electronic health record datasets to verify the effectiveness of our solution. On both datasets, model significantly outperforms baselines by up to 10-30

http://dl.acm.org/authorize?N33312

87. Evaluating U.S. Electoral Representation with a Joint Statistical Model of Congressional Roll-Calls, Legislative Text, and Voter Registration Data

Zhengming Xing (Criteo Labs);Sunshine Hillygus (Duke University);Lawrence Carin (Duke University)

Extensive information on 3 million randomly sampled United States citizens is used to construct a statistical model of constituent preferences for each U.S. congressional district. This model is linked to the legislative voting record of the legislator from each district, yielding an integrated model for constituency data, legislative roll-call votes, and the text of the legislation. The model is used to examine the extent to which legislators' voting records are aligned with constituent preferences, and the implications of that alignment (or lack thereof) on subsequent election outcomes. The analysis is based on a Bayesian formalism, with fast inference via a stochastic variational Bayesian analysis.

http://dl.acm.org/authorize?N33301

88. Distributed Multi-Task Relationship Learning

Sulin Liu (Nanyang Technological University, Singapore);Sinno Jialin Pan (Nanyang Technological University, Singapore);Qirong Ho (Petuum, Inc.)

Multi-task learning aims to learn multiple tasks jointly by exploiting their relatedness to improve the generalization performance for each task. Traditionally, to perform multi-task learning, one needs to centralize data from all the tasks to a single machine. However, in many real-world applications, data of different tasks may be geo-distributed over different local machines. Due to heavy communication caused by transmitting the data and the issue of data privacy and security, it is impossible to send data of different task to a master machine to perform multi-task learning. Therefore, in this paper, we propose a distributed multi-task learning framework that simultaneously learns predictive models for each task as well as task relationships between tasks alternatingly in the parameter server paradigm. In our framework, we first offer a general dual form for a family of regularized multi-task relationship learning methods. Subsequently, we propose a communication-efficient primal-dual distributed optimization algorithm to solve the dual problem by carefully designing local subproblems to make the dual problem decomposable. Moreover, we provide a theoretical convergence analysis for the proposed algorithm, which is specific for distributed multi-task relationship learning. We conduct extensive experiments on both synthetic and real-world datasets to evaluate our proposed framework in terms of effectiveness and convergence.

http://dl.acm.org/authorize?N33374

89. Inferring the strength of social ties: a community-driven approach

Polina Rozenshtein (Aalto University);Nikolaj Tatti (Aalto University);Aristides Gionis (Aalto University)

Online social networks are growing and becoming denser. The social connections of a given person may have very high variability: from close friends and relatives to acquaintances to people who hardly know. Inferring the strength of social ties is an important ingredient for modeling the interaction of users in a network and understanding their behavior. Furthermore, the problem has applications in computational social science, viral marketing, and people recommendation.

In this paper we study the problem of inferring the strength of social ties in a given network. Our work is motivated by a recent approach [?], which leverages the *strong triadic closure* (STC) principle, a hypothesis rooted in social psychology [?]. To guide our inference process, in addition to the network structure, we also consider as input a collection of *tight* communities. Those are sets of vertices that we expect to be connected via strong ties. Such communities appear in different situations, e.g., when being part of a community implies a strong connection to one of the existing members.

We consider two related problem formalizations that reflect the assumptions of our setting: small number of STC violations and strong-tie connectivity in the input communities. We show that both problem formulations are NP-hard. We also show that one problem formulation is hard to approximate, while for the second we develop an algorithm with approximation guarantee. We validate the proposed method on real-world datasets by comparing with baselines that optimize STC violations and community connectivity separately.

http://dl.acm.org/authorize?N33382

90. Mixture Factorized Ornstein-Uhlenbeck Processes for Time-Series Forecasting

Guo-Jun Qi (UCF);Jiliang Tang (MSU);Jingdong Wang (Microsoft);Jiebo Luo (University of Rochester)

Forecasting the future observations of time-series data can be performed by modeling the trend and fluctuations from the observed data. Many classical time-series analysis models like Autoregressive model (AR) and its variants have been developed to achieve such forecasting ability. While they are often based on the white noise assumption to model the data fluctuations, a more general

Brownian motion has been adopted that results in Ornstein-Uhlenbeck (OU) process. The OU process has gained huge successes in predicting the future observations over many genres of time series, however, it is still limited in modeling simple diffusion dynamics driven by a single persistent factor that never evolves over time. However, in many real problems, a mixture of hidden factors are usually present, and when and how frequently they appear or disappear are unknown ahead of time. This imposes a challenge that inspires us to develop a Mixture Factorized OU process (MFOUP) to model evolving factors. The new model is able to capture the changing states of multiple mixed hidden factors, from which we can infer their roles in driving the movements of time series. We conduct experiments on three forecasting problems, covering sensor and market data streams. The results show its competitive performance on predicting future observations and capturing evolution patterns of hidden factors as compared with the other algorithms.

http://dl.acm.org/authorize?N33389

91. Privacy-Preserving Distributed Multi-Task Learning with Asynchronous Updates

Liyang Xie (michigan state university);Inci Baytas (michigan state university);Kaixiang Lin (michigan state university);Jiayu Zhou (michigan state university)

Many data mining applications involve a set of related learning tasks. Multi-task learning (MTL) is a learning paradigm that improves generalization performance by transferring knowledge among those tasks. MTL has attracted so much attention in the community, and various algorithms have been successfully developed. Recently, distributed MTL has also been studied for related tasks whose data is distributed across different geographical regions. One prominent challenge of the distributed MTL frameworks is to maintain the privacy of the data. The distributed data may contain sensitive and private information such as patient records and registers of a company. In such cases, distributed MTL frameworks are required to preserve the privacy of the data. In this paper, we propose a novel privacy-preserving distributed MTL framework to address this challenge. A privacy-preserving proximal gradient algorithm, which asynchronously updates models of the learning tasks, is introduced to solve a general class of MTL formulations. The proposed asynchronous approach is robust against network delays and provides a guaranteed differential privacy through carefully designed perturbation. Theoretical guarantees of the proposed algorithm are derived and supported by the extensive experimental results.

http://dl.acm.org/authorize?N33300

92. Revisiting power-law distributions in spectra of real world networks

Nicole Eikmeier (Purdue University);David Gleich (Purdue University)

By studying a large number of real world graphs, we find empirical evidence that most real world graphs have a statistically significant power-law distribution with a cutoff in the singular values of the adjacency matrix and eigenvalues of the Laplacian matrix in addition to the commonly conjectured power-law in the degrees. Among these results, power-laws in the singular values appear more consistently than in the degree distribution. The exponents of the power-law distributions are much larger than previously observed. We find a surprising direct relationship between the power-law in the degree distribution and the power-law in the eigenvalues of the Laplacian that was theorized in simple models but is extremely accurate in practice. We investigate these findings in large networks by studying the cutoff value itself, which shows a scaling law for the number of elements involved in these power-laws. Using the scaling law enables us to compute only a subset of eigenvalues of large networks, up to tens of millions of vertices and billions of edges, where we find that those too show evidence of statistically significant power-laws.

http://dl.acm.org/authorize?N33362

93. Point of Interest Demand Modeling with Human Mobility Patterns

Yanchi Liu (Rutgers University);Chuanren Liu (Drexel University);Xinjiang Lu (Northwestern Polytechnical University, China);Mingfei Teng (Rutgers University);Hengshu Zhu (Baidu Talent Intelligence Center);Hui Xiong (Rutgers University)

Point-of-Interest (POI) demand modeling in urban regions is critical for many applications such as business site selection and real estate investment. While some efforts have been made for the demand analysis of some specific POI categories, such as restaurants, it lacks of a systematic means to support POI demand modeling. To this end, in this paper, we develop a systematic POI demand modeling framework, named Region POI Demand Identification (RPDI), to model POI demands by exploiting the daily needs of people identified from their large-scale mobility data. Specifically, we first partition the urban space into spatially differentiated neighborhood regions formed by many small local communities. Then, the daily activity patterns of people traveling in the city will be extracted from human mobility data. Since the trip activities, even aggregated, are sparse and insufficient to directly identify the POI demands, especially for underdeveloped regions, we develop a latent factor model that integrates human mobility data, POI profiles, and demographic data to robustly model the POI demand of urban regions in a holistic way. In this model, POI preferences and supplies are used together with demographic features to estimate the POI demands simultaneously for all the urban regions interconnected in the city. Moreover, we also design efficient algorithms to optimize the latent model for large-scale data. Finally, experimental results on real-world data in New York City (NYC) show that our method is effective for identifying POI demands for different regions.

http://dl.acm.org/authorize?N33385

94. Small Batch or Large Batch: Gaussian Walk with Rebound Can Teach

Peifeng Yin (IBM Research Almaden);Ping Luo (Institute of Computing Technology, CAS);Taiga Nakamura (IBM Research Almaden)

Efficiency of large-scale learning is a hot topic in both academic and industry. The *stochastic gradient descent (SGD)* algorithm, and its extension *mini-batch SGD*, allow the model to be updated without scanning the whole data set. However, the use of approximate gradient leads to the uncertainty issue, slowing down the decreasing of objective function. Furthermore, such uncertainty may result in a high frequency of meaningless update on the model, causing a communication issue in parallel learning environment. In this work, we develop a *batch-adaptive stochastic gradient descent (BA-SGD)* algorithm, which can dynamically choose a proper batch size as learning proceeds. Particularly on the basis of Taylor extension and central limit theorem, it models the decrease of objective value as a random walk game with a Gaussian dice. In this game, a heuristic strategy of determining batch size is adopted to maximize the utility of each incremental sampling. By evaluation on multiple real data sets, we demonstrate that by smartly choosing the batch size, the BA-SGD not only conserves the fast convergence of SGD algorithm but also avoids too frequent model updates.

http://dl.acm.org/authorize?N33318

95. Adversary Resistant Deep Neural Networks with an Application to Malware Detection

Qinglong Wang (Pennsylvania State University);Wenbo Guo (Pennsylvania State University);Kaixuan Zhang (Pennsylvania State University);Alexander Ororbia (Pennsylvania State University);Xinyu Xing (Pennsylvania State University);Lee Giles (Pennsylvania State University);Xue Liu (McGill University)

Beyond its highly publicized victories in Go, there have been numerous successful applications of deep learning in information retrieval, computer vision and speech recognition. In cybersecurity, an increasing number of companies have become excited about the potential of deep learning, and have started to use it for various security incidents, the most popular being malware detection. These companies assert that deep learning (DL) could help turn the tide in the battle against malware infections. However, deep neural networks (DNNs) are vulnerable to adversarial samples, a flaw that plagues most if not all statistical learning models. Recent research has demonstrated that those with malicious intent can easily circumvent deep learning-powered malware detection by exploiting this flaw.

In order to address this problem, previous work has developed various defense mechanisms that either augmenting training data or enhance model's complexity. However, after a thorough analysis of the fundamental flaw in DNNs, we discover that the effectiveness of current defenses is limited and, more importantly, cannot provide theoretical guarantees as to their robustness against adversarial sampled-based attacks. As such, we propose a new adversary resistant technique that obstructs attackers from constructing impactful adversarial samples by randomly nullifying features within samples. In this work, we evaluate our proposed technique against a real world dataset with 14,679 malware variants and 17,399 benign programs. We theoretically validate the robustness of our technique, and empirically show that our technique significantly boosts DNN robustness to adversarial samples while maintaining high accuracy in classification. To demonstrate the general applicability of our proposed method, we also conduct experiments using the MNIST and CIFAR-10 datasets, generally used in image recognition research.

http://dl.acm.org/authorize?N33305

96. PAMAE: Parallel k-Medoids Clustering with High Accuracy and Efficiency

Hwanjun Song (KAIST);Jae-Gil Lee (KAIST);Wook-Shin Han (POSTECH)

The k-medoids algorithm is one of the best-known clustering algorithms. Despite this, however, it is not as widely used for big data analytics as the k-means algorithm, mainly because of its high computational complexity. Many studies have attempted to solve the efficiency problem of the k-medoids algorithm, but all such studies have improved efficiency at the expense of accuracy. In this paper, we propose a novel parallel k-medoids algorithm, which we call PAMAE, that achieves both high accuracy and high efficiency. We identify two factors—-"global search" and "entire data"—-that are essential to achieving high accuracy, but are also very time-consuming if considered simultaneously. Thus, our key idea is to apply them individually through two phases: parallel seeding and parallel refinement, neither of which is costly. The first phase performs global search over sampled data, and the second phase performs local search over entire data. Our theoretical analysis proves that this serial execution of the two phases leads to an accurate solution that would be achieved by global search over entire data. In order to validate the merit of our approach, we implement PAMAE on Spark as well as Hadoop and conduct extensive experiments using various real-world data sets on 12 Microsoft Azure machines (48 cores). The results show that PAMAE significantly outperforms most of recent parallel algorithms and, at the same time, produces a clustering quality as comparable as the previous most-accurate algorithm. The source code and data are available at https://github.com/jaegil/k-Medoid.

http://dl.acm.org/authorize?N33399

97. Statistical Emerging Pattern Mining with Multiple Testing Correction

Junpei Komiyama (The University of Tokyo);Masakazu Ishihata (Hokkaido University);Hiroki Arimura (Hokkaido University);Takashi Nishibayashi (VOYAGE GROUP, Inc.);Shin-Ichi Minato (Hokkaido University)

Emerging patterns are patterns whose support significantly differs between two databases. We study the problem of listing emerging patterns with a multiple testing guarantee. Recently, Terada et al. proposed the Limitless Arity Multiple-testing Procedure (LAMP) that controls the family-wise error rate (FWER) in statistical association mining. LAMP reduces the number of "untestable" hypotheses without compromising its statistical power. Still, FWER is restrictive, and as a result, its statistical power is inherently unsatisfying when the number of patterns is large.

On the other hand, the false discovery rate (FDR) is less restrictive than FWER, and thus controlling FDR yields a larger number of significant patterns. We propose two emerging pattern mining methods: the first one controls FWER, and the second one controls FDR. The effectiveness of the methods is verified in computer simulations with real-world datasets.

http://dl.acm.org/authorize?N33370

98. Recommending Items with the Most Valuable Aspects Based on User Reviews

Konstantin Bauman (Stern School of Business, New York University);Bing Liu (University of Illinois at Chicago);Alexander Tuzhilin (Stern School of Business, New York University)

In this paper, we propose a recommendation technique that not only can recommend items of interest to the user as traditional recommendation systems do but also specific aspects of consumption of the items to further enhance the user experience with those items. For example, it can recommend the user to go to a specific restaurant (item) and also order some specific foods there, e.g., seafood (an aspect of consumption). Our method is called *Sentiment Utility Logistic Model* (SULM). As its name suggests, SULM uses sentiment analysis of user reviews. It first predicts the sentiment that the user may have on the item based on what he/she might express about the aspects of the item and then identifies the most valuable aspects of the user's potential experience with that item. Furthermore, the method can recommend items together with those most important aspects over which the user has control and can potentially select them, such as the time to go to a restaurant, e.g. lunch vs. dinner, and what to order there, e.g., seafood. We tested the proposed method on three applications (restaurant, hotel, and beauty,spa) and experimentally showed that those users who followed our recommendations of the most valuable aspects while consuming the items, had better experiences, as defined by the overall rating.

http://dl.acm.org/authorize?N33352

99. Functional Zone Based Hierarchical Demand Prediction For Bike System Expansion

Junming Liu (Rutgers University);Leilei Sun (Dalian University of Technology);Qiao Li (Rutgers University);Jingci Ming (Rutgers University);Yanchi Liu (Rutgers University);Hui Xiong (Rutgers University)

Bike sharing systems, aiming at providing the missing links in public transportation systems, are becoming popular in urban cities. Many providers of bike sharing systems are ready to expand their bike stations from the existing service area to surrounding regions. A key to success for a bike sharing systems expansion is the bike demand prediction for expansion areas. There are two major challenges in this demand prediction problem: 1. the bike transition records are not available for the expansion area and 2. station level bike demand have big variances across the urban city. Previous research mainly focus on discovering global features, assuming the station bike demands react equally to the global features, which brings large prediction error when the urban area is large and highly diversified. To address these challenges, in this paper, we develop a hierarchical station bike demand predictor which analyzes bike demands from functional zone level to station level. Specifically, we first divide the studied bike stations into functional zones by a novel Bi-clustering algorithm which is designed to cluster bike stations with similar POI characteristics and close geographical distances together. Then, the functional zone's hourly bike check-ins and check-outs are predicted by integrating three influential factors: distance preference, zone-to-zone preference and zone characteristics. The station demand is estimated by studying the demand distributions among the stations within the same functional zone. Finally, the extensive experimental results on the NYC Citi Bike system with two expansion stages show the advantages of our approach on station demand and balance prediction for bike sharing system expansions.

http://dl.acm.org/authorize?N33386

100. Decomposed Normalized Maximum Likelihood Codelength Criterion for Selecting Hierarchical Latent Variable Models

Tianyi Wu (University of Tokyo);Shinya Sugawara (University of Tokyo);Kenji Yamanishi (University of Tokyo)

We propose a new model selection criterion based on the minimum description length principle in a name of the decomposed normalized maximum likelihood criterion. Our criterion can be applied

to a large class of hierarchical latent variable models, such as the Naive Bayes models, stochastic block models and latent Dirichlet allocations, for which many conventional information criteria cannot be straightforwardly applied due to irregularity of latent variable models. Our method also has an advantage that it can be exactly evaluated without asymptotic approximation with small time complexity. Our experiments using synthetic and real data demonstrated validity of our method in terms of computational efficiency and modelselection accuracy, while our criterion especially dominated the other criteria when sample size is small and when data are noisy.

http://dl.acm.org/authorize?N33307

101. Structural Diversity and Homophily: A Study Across More Than One Hundred Big Networks

Yuxiao Dong (University of Notre Dame);Reid Johnson (University of Notre Dame);Jian Xu (University of Notre Dame);Nitesh Chawla (University of Notre Dame)

A widely recognized organizing principle of networks is structural homophily, which suggests that people with more common neighbors are more likely to connect with each other. However, what influence the diverse structures embedded in common neighbors have on link formation is much less well understood. To explore this problem, we begin by characterizing the structural diversity of common neighborhoods. Using a collection of 120 large-scale networks, we demonstrate that the impact of the common neighborhood diversity on link existence can vary substantially across networks, such as its positive effect in Facebook and negative one in LinkedIn, corresponding to different networking needs in these networks. We also discover striking cases where diversity violates the principle of homophily, that is, fewer mutual connections may lead to higher tendency to link with each other. We then leverage structural diversity to develop a common neighborhood signature (CNS) for a network, which we use to uncover distinct network superfamilies not discoverable by conventional methods. Our findings shed light on the pursuit to understand the ways in which network structures are organized and formed, pointing to potential advancement in designing random graph models and recommender systems.

http://dl.acm.org/authorize?N33361

102. Automatic Synonym Discovery with Knowledge Bases

Meng Qu (University of Illinois at Urbana-Champaign);Xiang Ren (University of Illinois at Urbana-Champaign);Jiawei Han (University of Illinois at Urbana-Champaign)

Recognizing entity synonyms from text has become a crucial task in many entity-leveraging applications. However, discovering entity synonyms from domain-specific text corpora (*e.g.*, news articles, scientific papers) are rather challenging. Current systems take an entity name string as input to find out other names that are synonymous, ignoring the fact that often times a name string can refer to multiple entities (*e.g.*, "apple" could refer to both *Apple Inc* and the fruit *apple*). Moreover, most existing methods require training data manually created by domain experts to construct supervised-learning systems. In this paper, we study the problem of automatic synonym discovery with knowledge bases, that is, identifying synonyms for *knowledge base entities* in a given domain-specific corpus. The manually-curated synonyms for each entity stored in a knowledge base not only form a set of name strings to *disambiguate* the meaning for each other, but also can serve as "*distant*" supervision to help determine important features for the task. We propose a novel framework, called **DPE**, to integrate two kinds of mutually-complementing signals for synonym discovery, *i.e.*, *distributional features* based on corpus-level statistics and *textual patterns* based on local contexts. In particular, DPE jointly optimizes the two kinds of signals in conjunction with distant supervision, so that they can mutually enhance each other in the training stage. At the inference stage, both signals will be utilized to discover synonyms for the given entities. Experimental results prove the effectiveness of the proposed framework.

http://dl.acm.org/authorize?N33380

103. A Context-aware Attention Network for Interactive Question Answering

Huayu Li (UNC Charlotte);Martin Renqiang Min (NEC Laboratories America);Yong Ge (University of Arizona);Asim Kadav (NEC Laboratories America)

Neural network based sequence-to-sequence models in an encoder-decoder framework have been successfully applied to solve Question Answering (QA) problems, predicting answers from statements and questions. However, almost all previous models have failed to consider detailed context information and unknown states under which systems do not have enough information to answer given questions. These scenarios with incomplete or ambiguous information are very common in the setting of Interactive Question Answering (IQA). To address this challenge, we develop a novel model, employing context-dependent word-level attention for more accurate statement representations and question-guided sentence-level attention for better context modeling, and design a new IQA dataset, which will be made publicly available, to test our model. Employing these attention mechanisms, our model accurately understands when it can output an answer or when it requires generating a supplementary question for additional input depending on different contexts. When available, user's feedback is encoded and directly applied to update sentence-level attention to infer an answer. Extensive experiments on QA and IQA datasets demonstrate quantitatively the effectiveness of our model with significant improvement over state-of-the-art conventional QA models.

http://dl.acm.org/authorize?N33373

104. REMIX: Automated Exploration for Interactive Outlier Detection

Yanjie Fu (Missouri S&amp;T);Charu Aggarwal (IBM TJ Watson Research);Srinivasan Parthasarathy (IBM TJ Watson Research);Deepak Turaga (IBM TJ Watson Research);Hui Xiong (Rutgers University)

Outlier detection is the identification of points in a dataset that do not conform to the norm. Outlier detection is highly sensitive to the choice of the detection algorithm and the feature subspace used by the algorithm. Extracting domain-relevant insights from outliers needs systematic exploration of these choices since diverse outlier sets could lead to complementary insights. This challenge is especially acute in an interactive setting, where the choices must be explored in a time-constrained manner. In this work, we present REMIX, the first system to address the problem of outlier detection in an interactive setting. REMIX uses a novel mixed integer programming (MIP) formulation for automatically selecting and executing a diverse set of outlier detectors within a time limit. This formulation incorporates multiple aspects such as (i) an upper limit on the total execution time of detectors, (ii) diversity in the space of algorithms and features, and (iii) meta-learning for evaluating the cost and utility of detectors. REMIX provides two distinct ways for the analyst to consume its results: (i) a partitioning of the detectors explored by REMIX into perspectives through low-rank non-negative matrix factorization; each perspective can be easily visualized as an intuitive heatmap of experiments versus outliers, and (ii) an ensembled set of outliers which combines outlier scores from all detectors. We demonstrate the benefits of REMIX through extensive empirical validation on real-world data.

http://dl.acm.org/authorize?N33363

105. Anomaly Detection in Streams with Extreme Value Theory

Alban Siffer (IRISA);Pierre-Alain Fouque (IRISA);Alexandre Termier (IRISA);Christine Largouët (IRISA)

Anomaly detection in time series has attracted considerable attention due to its importance in many real-world applications including intrusion detection, energy management and finance. Most approaches for detecting outliers rely on either manually set thresholds or assumptions on the distribution of data according to Chandola, Banerjee and Kumar.

Here, we propose a new approach to detect outliers in streaming univariate time series based on Extreme Value Theory that does not require to hand-set thresholds and makes no assumption on the distribution: the main parameter is only the risk, controlling the number of false positives. Our approach can be used for outlier detection, but more generally for automatically setting thresholds,

making it useful in wide number of situations. We also experiment our algorithms on various real-world datasets which confirm its soundness and efficiency.

http://dl.acm.org/authorize?N33397

106. Structural Event Detection from Log Messages

Fei Wu (penn state);Pranay Anchuri (NEC Labs America);Zhenhui Li (Penn State University)

A wide range of modern web applications are only possible because of the composable nature of the web services they are built upon. It is, therefore, often critical to ensure proper functioning of these web services. As often, the server-side of web services is not directly accessible, several log message based analysis have been developed to monitor the status of web services. Existing techniques focus on using clusters of messages (log patterns) to detect important system events. We argue that meaningful system events are often representable by groups of cohesive log messages and the relationships among these groups. We propose a novel method to mine structural events as directed workflow graphs (where nodes represent log patterns, and edges represent relations among patterns). The structural events are inclusive and correspond to interpretable episodes in the system. The problem is non-trivial due to the nature of log data: (i) Individual log messages contain limited information, and (ii) Log messages in a large scale web system are often interleaved even though the log messages from individual components are ordered. As a result, the patterns and relationships mined directly from the messages and their ordering can be errorenous and unreliable in practice. Our solution is based on the observation that meaningful log patterns and relations often form workflow structures that are connected. Our method directly models the overall quality of structural events. Through both qualitative and quantitative experiments on real world datasets, we demonstrate the effectiveness and the expressiveness of our event detection method.

http://dl.acm.org/authorize?N33308

107. Bridging Collaborative Filtering and Semi-Supervised Learning: A Neural Approach for POI recommendation

Carl Yang (University of Illinois, Urbana Champaign);Lanxiao Bai (University of Illinois, Urbana Champaign);Chao Zhang (University of Illinois, Urbana Champaign);Quan Yuan (University of Illinois, Urbana Champaign);Jiawei Han (University of Illinois, Urbana Champaign)

Recommender system is one of the most popular data mining topics that keep drawing extensive attention from both academia and industry. Among them, POI (point of interest) recommendation is extremely practical but challenging: it greatly benefits both users and businesses in real-world life, but it is hard due to data scarcity and various context. While a number of algorithms attempt to tackle the problem w.r.t. specific data and problem settings, they often fail when the scenarios change. In this work, we propose to devise a general and principled SSL (semi-supervised learning) framework, to alleviate data scarcity via smoothing among neighboring users and POIs, and treat various context by regularizing user preference based on context graphs. To enable such a framework, we develop PACE (Preference And Context Embedding), a deep neural architecture that jointly learns the embeddings of users and POIs to predict both user preference over POIs and various context associated with users and POIs. We show that PACE successfully bridges CF (collaborative filtering) and SSL by generalizing the *de facto* methods matrix factorization of CF and graph Laplacian regularization of SSL. Extensive experiments on two real location-based social network datasets demonstrate the effectiveness of PACE.

http://dl.acm.org/authorize?N33315

108. Sparse Compositional Local Metric Learning

Joseph St.Amand (University of Kansas);Jun Huan (University of Kansas)

Mahalanobis distance metric learning becomes an especially challenging problem as the dimension of the feature space p is scaled upwards. The number of parameters to optimize grows with O (p 2) complexity, making storage infeasible, interpretability poor, and causing the model to have a high tendency to overfit. Additionally, optimization while maintaining feasibility of the solution becomes prohibitively expensive, requiring a projection onto the positive semi-definite cone after every iteration. In addition to the obvious space and computational challenges, vanilla distance metric learning is unable to model complex and multi-modal trends in the data.

Inspired by the recent resurgence of Frank-Wolfe style optimization, we propose a new method for sparse compositional local Mahalanobis distance metric learning. Our proposed technique learns a set of distance metrics which are composed of local and global components. We capture local interactions in the feature space, while ensuring that all metrics share a global component, which may act as a regularizer. We optimize our model using an alternating pairwise Frank-Wolfe style algorithm. This serves a dual purpose, we can control the sparsity of our solution, and altogether avoid any expensive projection operations. We conduct an empirical evaluation of our technique on 5 separate datasets and find that in some cases our proposed technique is capable of outperforming current state of the art methods.

http://dl.acm.org/authorize?N33390

109. Learning from Multiple Teacher Networks

Shan You (Peking University);Chang Xu (University of Technology Sydney);Chao Xu (Peking University);Dacheng Tao (University of Sydney)

Training thin deep networks following the student-teacher learning paradigm has received intensive attention because of its excellent performance. However, to the best of our knowledge, most existing work mainly considers one single teacher network. In practice, a student may access multiple teachers, and multiple teacher networks together provide comprehensive guidance that are beneficial for training the student network. In this paper, we present a method to train a thin deep network by incorporating multiple teacher networks not only in output layer by averaging the softened outputs (dark knowledge) from different networks, but also in the intermediate layers by imposing a constraint about the dissimilarity among examples. We suggest that the relative dissimilarity between intermediate representations of different examples serves as a more flexible and appropriate guidance from teacher networks. Then triplets are utilized to encourage the consistence of these relative dissimilarity relationships between the student network and teacher networks. Moreover, we leverage a voting strategy to unify multiple relative dissimilarity information provided by multiple teacher networks, which realizes their incorporation in the intermediate layers. Extensive experimental results demonstrated that our method is capable of generating a well-performed student network, with the classification accuracy comparable or even superior to all teacher networks, yet having much fewer parameters and being much faster in running.

http://dl.acm.org/authorize?N33319

110. Let's See Your Digits: Anomalous-State Detection using Benford's Law

Samuel Maurus (Technical University of Munich);Claudia Plant (University of Vienna)

Benford's Law explains a curious "naturally-occurring" phenomenon in which the leading digits of numerical data are distributed in a precise fashion. In this paper we begin by showing that system metrics generated by many modern information systems like Twitter, Wikipedia, YouTube and GitHub obey this law. We then propose a novel unsupervised approach called BenFound that exploits this property to detect anomalous system events. BenFound tracks the "Benfordness" of key system metrics, like the follower counts of tweeting Twitter users or the change deltas in Wikipedia page edits. It then applies a novel Benford-conformity test in real-time to identify "non-Benford events". We investigate a variety of such events, showing that they correspond to unnatural and often undesired system interactions like spamming, hashtag-hijacking and denial-of-service attacks. The result is a technically-uncomplicated and effective "red flagging" technique. Although not without its limitations, it is highly efficient and requires neither obscure parameters, nor text streams, nor natural-language processing.

http://dl.acm.org/authorize?N33388

111. Inductive Semi-supervised Multi-Label Learning with Co-Training

Wang Zhan (Southeast University);Min-Ling Zhang (School of Computer Science and Engineering, Southeast University)

In multi-label learning, each training example is associated with multiple class labels and the task is to learn a mapping from the feature space to the power set of label space. It is generally demanding and time-consuming to obtain labels for training examples, especially for multi-label learning task where a number of class labels need to be annotated for the instance. To circumvent this difficulty, semi-supervised multi-label learning aims to exploit the readily-available unlabeled data to help build multi-label predictive model. Nonetheless, most semi-supervised solutions to multi-label learning work under transductive setting, which only focus on making predictions on existing unlabeled data and cannot generalize to unseen instances. In this paper, a novel approach named COINS is proposed to learning from labeled and unlabeled data by adapting the well-known co-training strategy which naturally works under inductive setting. In each co-training round, a dichotomy over the feature space is learned by maximizing the diversity between the two classifiers induced on either dichotomized feature subset. After that, pairwise ranking predictions on unlabeled data are communicated between either classifier for model refinement. Extensive experiments on a number of benchmark data sets show that COINS performs favorably against state-of-the-art multi-label learning approaches.

http://dl.acm.org/authorize?N33311

112. Robust Spectral Clustering for Noisy Data

Aleksandar Bojchevski (Technical University of Munich);Yves Matkovic (Technical University of Munich);Stephan Günnemann (Technical University of Munich)

Spectral clustering is one of the most prominent clustering approaches. However, it is highly sensitive to noisy input data. In this work, we propose a robust spectral clustering technique able to handle such scenarios. To achieve this goal, we propose a sparse and latent decomposition of the similarity graph used in spectral clustering. In our model, we jointly learn the spectral embedding as well as the corrupted data—thus, enhancing the clustering performance overall. We propose algorithmic solutions to all three established variants of spectral clustering, each showing linear complexity in the number of edges. Our experimental analysis confirms the significant potential of our approach for robust spectral clustering.

http://dl.acm.org/authorize?N33354

113. Multi-task Function-on-function Regression with Co-grouping Structured Sparsity

Pei Yang (South China University of Technology);Qi Tan (South China Normal University);Jingrui He (Arizona State University)

The growing importance of functional data has fueled the rapid development of functional data analysis, which treats the infinite-dimensional data as continuous functions rather than discrete, finite-dimensional vectors. On the other hand, heterogeneity is an intrinsic property of functional data due to the variety of sources to collect the data. In this paper, we propose a novel multi-task function-on-function regression approach to model both the functionality and heterogeneity of data. The basic idea is to simultaneously model the relatedness among tasks and correlations among basis functions by using the co-grouping structured sparsity to encourage similar tasks to behave similarly in shrinking the basis functions. The resulting optimization problem is challenging due to the non-smoothness and non-separability of the co-grouping structured sparsity. We present an efficient algorithm to solve the problem, and prove its separability, convexity, and global convergence. The proposed algorithm is applicable to a wide spectrum of structured sparsity regularized techniques, such as structured $\ell_{2,p}$ and structured Schatten-p norms. The effectiveness of the proposed approach is verified on benchmark functional data sets.

http://dl.acm.org/authorize?N33316

114. Multi-Modality Disease Modeling via Collective Deep Matrix Factorization

Qi Wang (Michigan State University);Mengying Sun (Michigan State University);Liang Zhan (University of Wisconsin-Stout);Paul Thompson (University of Southern California);Shuiwang Ji (Washington State University);Jiayu Zhou (Michigan State University)

Alzheimer's disease (AD), one of the most common causes of dementia, is a severe irreversible neurodegenerative disease that results in loss of mental functions. The transitional stage between the expected cognitive decline of normal aging and AD, mild cognitive impairment (MCI), has been widely regarded as a suitable time for possible therapeutic intervention. The challenging task of MCI detection is therefore of great clinical importance, where the key is to effectively fuse predictive information from multiple heterogeneous data sources collected from the patients. In this paper, we propose a framework to fuse multiple data modalities for predictive modeling using deep matrix factorization, which explores the non-linear interactions among the modalities and exploits such interactions to transfer knowledge and enable high performance prediction. Specifically, the proposed collective deep matrix factorization decomposes all modalities simultaneously to capture non-linear structures of the modalities in a supervised manner, and learns a modality specific component for each modality and a modality invariant component across all modalities. The modality invariant component serves as a compact feature representation of patients that has high predictive power. The modality specific components provide an effective means to explore imaging genetics, yielding insights into how imaging and genotype interact with each other non-linearly in the AD pathology. Extensive empirical studies using various data modalities provided by Alzheimer's Disease Neuroimaging Initiative (ADNI) demonstrate the effectiveness of the proposed method for fusing heterogeneous modalities.

http://dl.acm.org/authorize?N33306

115. A Location-Sentiment-Aware Recommender System for Both Home-Town and Out-of-Town Mobile Users

Hao Wang (Institute of Software, Chinese Academy of Sciences);Yanmei Fu (Institute of Software, Chinese Academy of Sciences);Qinyong Wang (Institute of Software, Chinese Academy of Science);Changying Du (Institute of Software, Chinese Academy of Sciences);Hongzhi Yin (University of Queensland);Hui Xiong (Rutgers University)

Spatial item recommendation has become an important means to help people discover interesting locations, especially when people pay a visit to unfamiliar regions. Some current researches are focusing on modelling individual and collective geographical preferences for spatial item recommendation based on users' check-in records, but they fail to explore the phenomenon of user interest drift across geographical regions, i.e., users have different interests when they travel in different regions. Besides, they ignore the sentiment influence of crowds' reviews for users' check-in behaviors. Specifically, it is intuitive that users would not check in a spatial item whose overall history reviews seem negative, though it might satisfy their interests. Therefore, it should recommend the item to the user at the location. In this paper, we propose a latent probabilistic generative model called LSARS to mimic the decision-making process of users' check-in activities both in home-town and out-of-town scenarios by adapting to user interest drift and implicit sentiment for spatial items, which can learn location-aware and sentiment-aware individual interests according to the contents of spatial items and crowds' reviews. As users' check-in records left in out-of-town regions are extremely sparse, LSARS incorporates the crowds' preferences learned from local users' check-in behaviors. We deploy LSARS to two application scenarios: spatial item recommendation and target user discovery. Extensive experiments on two large-scale location-based social networks (LBSNs) datasets show that LSARS achieves better performance than existing state-of-the-art competing methods.

http://dl.acm.org/authorize?N33394

116. On Sampling Strategies for Neural Network-based Collaborative Filtering

Ting Chen (University of California, Los Angeles);Yizhou Sun (University of California, Los Angeles);Yue Shi (Yahoo! Research);Liangjie Hong (Etsy Inc.)

Recent advances in neural networks have inspired people to design hybrid recommendation algorithms that can take care of both (1) user-item interaction information and (2) content information including images, audios, and text, without tedious feature engineering. Despite their promising results, neural network-based recommendation algorithms pose extensive computational costs, making it harder to scale and improve upon.

In this paper, we propose a general neural network-based recommendation framework, which subsumes several existing state-of-the-art neural network-based recommendation algorithms, and address the efficiency issue by investigating sampling strategies in the stochastic gradient descent algorithm for the framework. We tackle this issue by first establishing a connection between the loss functions and the user-item interaction bipartite graph, where loss functions are defined on links while costly computation are on nodes. Based on this insight, three novel node-based sampling strategies are proposed, which can significantly improve the training efficiency of the proposed framework (up to ×30 times speedup in our experiments), as well as improving the recommendation performance. Theoretical analysis is also provided for both the computational cost and the convergence. We believe our study of sampling strategies have further implications on general graph-based loss functions, and would also enable more research under the neural network-based recommendation framework.

http://dl.acm.org/authorize?N33367

117. GRAM: Graph-based Attention Model for Healthcare Representation Learning

Edward Choi (Georgia Institute of Technology);Mohammad Taha Bahadori (Georgia Institute of Technology);Le Song (Georgia Institute of Technology);Walter Stewart (Sutter Health);Jimeng Sun (Georgia Institute of Technology)

Deep learning methods exhibit promising performance for predictive modeling in healthcare, but two important challenges remain:

<ul>

<li>Data insufficiency: Often in healthcare predictive modeling, the sample size is insufficient for deep learning methods to achieve satisfactory results.</li>

<li>Interpretation: The representations learned by deep learning methods should align with medical knowledge.</li>

</ul>

To address these challenges, we propose a GRaph-based Attention Model, GRAM that supplements electronic health records (EHR) with hierarchical information inherent to medical ontologies. Based on the data volume and the ontology structure, GRAM represents a medical concept as a combination of its ancestors in the ontology via an attention mechanism. We compared predictive performance (i.e. accuracy, data needs, interpretability) of GRAM to various methods including the recurrent neural network (RNN) in two sequential diagnoses prediction tasks and one heart failure prediction task. Compared to the basic RNN, GRAM achieved 10

http://dl.acm.org/authorize?N33369

118. Bolt: Accelerated Data Mining with Fast Vector Compression

Davis Blalock (MIT);John Guttag (MIT)

Vectors of data are at the heart of machine learning and data mining. Recently, vector quantization methods have shown great promise in reducing both the time and space costs of operating on vectors. We introduce a vector quantization algorithm that can compress vectors up to 12x faster than existing techniques while also accelerating approximate vector operations such as distance

and dot product computations by over 10x. Because it can encode over two megabytes of vectors per millisecond (2 GB/s), it makes vector quantization cheap enough to employ in many more circumstances. As an example, using our technique to compute approximate dot products in a nested loop can multiply matrices faster than a state-of-the-art BLAS implementation, even when our algorithm must first compress the matrices.

In addition to showing the above speedups, we show experimentally that our approach can be used to accelerate nearest neighbor search and maximum inner product search by up to 140x compared to floating point operations and 10x compared to other vector quantization methods. Our approximate Euclidean distance and dot product computations are not only faster than those of related algorithms with slower encodings, but also faster than Hamming distance computations, which have direct hardware support on the tested platforms. We also assess the errors of our algorithm's approximate distances and dot products, and find that it is competitive with existing, slower vector quantization algorithms.

http://dl.acm.org/authorize?N33353

119. When is a Network a Network? Multi-Order Graphical Model Selection in Pathways and Temporal Networks

Ingo Scholtes (ETH Zurich)

We introduce a framework for the modeling of sequential data capturing pathways of varying lengths observed in a network. Such data are important, e.g., when studying click streams in information networks, travel patterns in transportation systems, information cascades in social networks, biological pathways or time-stamped social interactions. While it is common to apply graph analytics and network analysis to such data, recent works have shown that temporal correlations can invalidate the results of such methods. This raises a fundamental question: when is a network abstraction of sequential data justified? Addressing this open question, we propose a framework which combines Markov chains of multiple, higher orders into a multi-layer graphical model that captures temporal correlations in pathways at multiple length scales simultaneously. We develop a model selection technique to infer the optimal number of layers of such a model and show that it outperforms previously used Markov order detection techniques. An application to eight real-world data sets on pathways and temporal networks shows that it allows to infer graphical models which capture both topological and temporal characteristics of such data. Our work highlights fallacies of network abstractions and provides a principled answer to the open question when they are justified. Generalizing network representations to multi-order graphical models, it opens perspectives for new data mining and knowledge discovery algorithms.

http://dl.acm.org/authorize?N33384

120. A Temporally Heterogeneous Survival Framework with Application to Social Behavior Dynamics

Linyun Yu (Tsinghua University);Peng Cui (Tsinghua University);Chaoming Song (University of Miami);Tianyang Zhang (Tsinghua University);Shiqiang Yang (Tsinghua University)

Social behavior dynamics is one of the central building blocks in understanding and modeling complex social dynamic phenomena, such as information spreading, opinion formation, and social mobilization. While a wide range of models for social behavior dynamics have been proposed in recent years, the essential ingredients and the minimum model for social behavior dynamics is still largely unanswered. Here, we find that human interaction behavior dynamics exhibit rich complexities over the response time dimension and natural time dimension by exploring a large scale social communication dataset. To tackle this challenge, we develop a temporal Heterogeneous Survival framework where the regularities in response time dimension and the circadian rhythm in natural time dimension are organically integrated. We apply our model in two online social communication datasets. Our model can successfully regenerate the interaction patterns in the social communication datasets, and the results demonstrate that the proposed method can significantly outperform other state-of-the-art baselines. Meanwhile, the learnt parameters and discovered statistical regularities can lead to multiple potential applications.

http://dl.acm.org/authorize?N33310

121. End-to-end Learning for Short Text Expansion

Jian Tang (University of Michigan);Yue Wang (University of Michigan);Kai Zheng (University of California, Irvine);Qiaozhu Mei (University of Michigan)

Effectively making sense of short texts is a critical task for many real world applications such as search engines, social media services, and recommender systems. The task is particularly challenging as a short text contains very sparse information, often too sparse for a machine learning algorithm to pick up useful signals. A common practice for analyzing short text is to first expand it with external information, which is usually harvested from a large collection of longer texts. In literature, short text expansion has been done with all kinds of heuristics. We propose an end-to-end solution that automatically learns how to expand short text to optimize a given learning task. A novel deep memory network is proposed to automatically find relevant information from a collection of longer documents and reformulate the short text through a gating mechanism. Using short text classification as a demonstrating task, we show that the deep memory network significantly outperforms classical text expansion methods with comprehensive experiments on real world data sets.

http://dl.acm.org/authorize?N33391

122. Semi-Supervised Techniques for Mining Learning Outcomes and Prerequisites

Igor Labutov (Carnegie Mellon University);Yun Huang (University of PIttsburgh);Peter Brusilovsky (University of PIttsburgh);Daqing He (University of PIttsburgh)

Educational content of today no longer only resides in textbooks and classrooms; more and more learning material is found in a free, accessible form on the Internet. Our long-standing vision is to transform this web of educational content into an adaptive, web-scale “textbook"‘, that can guide its readers to most relevant “pages” according to their learning goal and current knowledge. In this paper, we address one core, long-standing problem towards this goal: identifying outcome and prerequisite concepts within a piece of educational content (e.g., a tutorial). Specifically, we propose a novel approach that leverages textbooks as a source of distant supervision, but learns a model that can generalize to arbitrary documents (such as those on the web). As such, our model can take advantage of any existing textbook, without requiring expert annotation. At the task of predicting outcome and prerequisite concepts, we demonstrate improvements over a number of baselines on six textbooks, especially in the regime of little to no ground-truth labels available. Finally, we demonstrate the utility of a model learned using our approach at the task of identifying prerequisite documents for adaptive content recommendation—- an important step towards our vision of the “web as a textbook"‘.

http://dl.acm.org/authorize?N33371

123. Learning from Labeled and Unlabeled Vertices in Networks

Wei Ye (University of Munich);Linfei Zhou (University of Munich);Dominik Mautz (University of Munich);Claudia Plant (University of Vienna);Christian Böhm (University of Munich)

Networks such as social networks, citation networks, protein-protein interaction networks, etc., are prevalent in real world. However, only very few vertices have labels compared to large amount of unlabeled vertices. For example, in social networks, not every user provides his/her profile information such as the personal interests which are better for social networking advertising. Can we leverage the limited user information and friendship network wisely to infer the likely labels of

unlabeled users? In this paper, we propose a semi-supervised learning framework called weighted-vote Geometric Neighbor classifier (wvGN) to infer the likely labels of unlabeled vertices. wvGN exploits random walks to explore not only local but also global neighborhood information of a vertex. Then the label of the vertex is determined by the accumulated local and global neighborhood information. Specifically, wvGN optimizes a proposed objective function by a search strategy which is based on the gradient and coordinate descent methods. The search strategy iteratively conducts a coarse search and a fine search to escape from local optima. Extensive experiments on various synthetic and real-world data verifies the effectiveness of wvGN compared to the state-of-the-art approaches.

http://dl.acm.org/authorize?N33317

124. Convex Factorization Machine for Toxicogenomics Prediction

Makoto Yamada (Kyoto University);Wenzhao Lian (Vicarius);Amit Goyal (Yahoo Research);Jianhui Chen (Yahoo Research);Kishan Wimalawarne (Kyoto University);Suleiman Kahn (University of Helsinki);Samuel Kaski (Aalto University);Hiroshi Mamitsuka (Kyoto University);Yi Chang (Huawei Research)

We propose the convex factorization machine (CFM), which is a convex variant of the widely used Factorization Machines (FMs). Specifically, we employ a linear+quadratic model and regularize the linear term with the L2-regularizer and the quadratic term with the trace norm regularizer. Then, we formulate the CFM optimization as a semidefinite programming problem and propose an efficient optimization procedure with Hazan's algorithm. A key advantage of CFM over existing FMs is that it can find a globally optimal solution, while FMs may get a poor locally optimal solution since the objective function of FMs is non-convex. In addition, the proposed algorithm is simple yet effective and can be implemented easily. Finally, CFM is a general factorization method and can also be used for other factorization problems including multi-view matrix factorization and tensor completion problems. Through synthetic and movielens datasets, we first show that the proposed CFM achieves results competitive to FMs. Furthermore, in a toxicogenomics prediction task, we show that CFM outperforms a state-of-the-art tensor factorization method.

http://dl.acm.org/authorize?N33302

125. Post Processing Recommender Systems for Diversity

Arda Antikacioglu (Carnegie Mellon University);R Ravi (Tepper School of Business, Carnegie Mellon University)

Collaborative filtering is a broad and powerful framework for building recommendation systems that has seen widespread adoption. Over the past decade, the propensity of such systems for favoring popular products and thus creating echo chambers have been observed. This has given rise to an active area of research that seeks to diversify recommendations generated by such algorithms. We address the problem of increasing diversity in recommendation systems that are based on collaborative filtering that use past ratings to predicting a rating quality for potential recommendations. Following our earlier work, we formulate recommendation system design as a subgraph selection problem from a candidate super-graph of potential recommendations where both diversity and rating quality are explicitly optimized: (1) On the modeling side, we define a new flexible notion of diversity that allows a system designer to prescribe the number of recommendations each item should receive, and smoothly penalizes deviations from this distribution. (2) On the algorithmic side, we show that minimum-cost network flow methods yield fast algorithms in theory and practice for designing recommendation subgraphs that optimize this notion of diversity. (3) On the empirical side, we show the effectiveness of our new model and method to increase diversity while maintaining high rating quality in standard rating data sets from Netflix and MovieLens.

http://dl.acm.org/authorize?N33351

126. Fast Newton Hard Thresholding Pursuit for Sparsity Constrained Nonconvex Optimization

Jinghui Chen (University of Virginia);Quanquan Gu (University of Virginia)

We propose a fast Newton hard thresholding pursuit algorithm for sparsity constrained nonconvex optimization. Our proposed algorithm reduces the per-iteration time complexity to be linear in the data dimension d compared with cubic time complexity in Newton's method, while preserving faster computational and statistical convergence rates. In particular, we prove that the proposed algorithm converges to the unknown true model parameter at a composite rate, namely quadratic at first and linear when it gets close to the true parameter, up to the minimax optimal statistical precision of the underlying model. Thorough experiments on both synthetic and real datasets demonstrate that our algorithm outperforms the state-of-the-art optimization algorithms for sparsity constrained optimization.

http://dl.acm.org/authorize?N33366

127. Retrospective Higher-Order Markov Processes for User Trails

Tao Wu (Purdue University);David Gleich (Purdue University)

Users form trails as they browse the web, checkin with a geolocation, rate items, or consume media. A common problem is to estimate what a user might do next for the purposes of guidance, recommendation, or prefetching. First-order and higher-order Markov chains have been one of the most widely used methods to study such sequences of data. First-order Markov chains are easy to estimate, but lack accuracy when history matters. Higher-order Markov chains suffer from overfitting due to their large numbers of parameters and the sparsity in the training data. Regularized fitting only offers mild improvements to the accuracy. In this paper we propose the retrospective higher-order Markov process (RHOMP) as a low-parameter model for such sequences. This model is a special case of a higher-order Markov chain where the transitions depend retrospectively on a single history state instead of an arbitrary combination of history states. There are two immediate computational advantages: the model complexity only grows linear with the order of the Markov chains and such model scales to large state spaces. Furthermore, by providing a specific structure to the higher-order chain, RHOMPs improve the model accuracy by efficiently utilizing history states without risks of overfitting the data. We demonstrate how to estimate a RHOMP from data and we demonstrate the effectiveness of our method on various real application datasets spanning geolocation data, review sequences, and locations. The RHOMP model uniformly outperforms higher-order Markov chains, Kneser-Ney regularization, and tensor factorizations in terms of prediction accuracy.

http://dl.acm.org/authorize?N33309

128. Effective Evaluation using Logged Bandit Feedback from Multiple Loggers

Aman Agarwal (Cornell University);Soumya Basu (Cornell University);Tobias Schnabel (Cornell University);Thorsten Joachims (Cornell University)

Accurately evaluating new policies (e.g. ad-placement models, ranking functions, recommendation functions) is one of the key problems in improving interactive systems. While the conventional approach to evaluation relies on online A/B tests, recent work has shown that counterfactual estimators can provide an inexpensive and fast alternative, since they can be applied offline using log data that was collected from a different policy fielded in the past. In this paper, we address the question of how to estimate the performance of a new policy when we have log data from multiple historic policies. This question is of great relevance in practice, since policies get updated frequently in most online systems. We show that naively combining data from multiple logging policies is highly suboptimal. In particular, we find that the standard Inverse Propensity Score (IPS) estimator suffers especially when logging and evaluation policies diverge—to a point where throwing away data improves the variance of the estimator. We therefore propose two alternative estimators which we characterize theoretically and compare experimentally. We find empirically that the new estimators can provide substantially improved estimation accuracy.

http://dl.acm.org/authorize?N33359

129. ReasoNet: Learning to Stop Reading in Machine Comprehension

Yelong Shen (Microsoft Research);Po-Sen Huang (Microsoft Research);Jianfeng Gao (Microsoft Research);Weizhu Chen (Microsoft Research)

Teaching a computer to read a document and answer general questions pertaining to the document is a challenging yet unsolved problem. In this paper, we describe a novel neural network architecture called the Reasoning Network (ReasoNet) for machine comprehension tasks. ReasoNets make use of multiple turns to effectively exploit and then reason over the relation among queries, documents, and answers. Different from previous approaches using a fixed number of turns during inference, ReasoNets introduce a termination state to relax this constraint on the reasoning depth. With the use of reinforcement learning, ReasoNets can dynamically determine whether to continue the comprehension process after digesting intermediate results, or to terminate reading when it concludes that existing information is adequate to produce an answer. ReasoNets have achieved exceptional performance in machine comprehension datasets, including unstructured CNN and Daily Mail datasets, the Stanford SQuAD dataset, and a structured Graph Reachability dataset.

http://dl.acm.org/authorize?N33395

130. Relay-Linking Models for Prominence and Obsolescence in Evolving Networks

Mayank Singh (Indian Institute of Technology, Kharagpur);Rajdeep Sarkar (Indian Institute of Technology, Kharagpur);Pawan Goyal (Indian Institute of Technology, Kharagpur);Animesh Mukherjee (Indian Institute of Technology, Kharagpur);Soumen Chakrabarti (Indian Institute of Technology, Bombay)

The rate at which nodes in evolving social networks acquire links (friends, citations) shows complex temporal dynamics. Preferential attachment and link copying models, while elegant and simple, only capture rich-gets-richer effects, not aging and decline. Recent aging models are complex and heavily parameterized; most involve estimating 1

http://dl.acm.org/authorize?N33398

131. Achieving Non-Discrimination in Data Release

Lu Zhang (University of Arkansas);Yongkai Wu (University of Arkansas);Xintao Wu (University of Arkansas)

Discrimination discovery and prevention/removal are increasingly important tasks in data mining. Discrimination discovery aims to unveil discriminatory practices on the protected attribute (e.g., gender) by analyzing the dataset of historical decision records, and discrimination prevention aims to remove discrimination by modifying the biased data before conducting predictive analysis. In this paper, we show that the key to discrimination discovery and prevention is to find the meaningful partitions that can be used to provide quantitative evidences for the judgment of discrimination. With the support of the causal graph, we present a graphical condition for identifying a meaningful partition. Based on that, we develop a simple criterion for the claim of non-discrimination, and propose discrimination removal algorithms which accurately remove discrimination while retaining good data utility. Experiments using real datasets show the effectiveness of our approaches.

http://dl.acm.org/authorize?N33314

132. MetaPAD: Meta Patten Discovery from Massive Text Corpora

Meng Jiang (University of Illinois at Urbana-Champaign);Jingbo Shang (University of Illinois at Urbana-Champaign);Taylor Cassidy (Army Research Lab);Xiang Ren (University of Illinois at Urbana-Champaign);Lance Kaplan (Army Research Lab);Timothy Hanratty (Army Research Lab);Jiawei Han (University of Illinois at Urbana-Champaign)

Mining textual pattens in news, tweets, papers, and many other kinds of text corpora has been an active theme in text mining and NLP research. Previous studies adopt a dependency parsing-based patten discovery approach. However, the parsing results lose rich around entities in the patten, and the process is costly for a corpus of large scale. In this study, we propose a novel typed textual patten

structure, called meta patten, which is extended to a frequent, informative, and precise subsequence patten in certain context. We propose an efficient framework, called MetaPAD, which discovers meta patten from massive corpora with three techniques: (1) it develops a context segmentation method to carefully determine the boundaries of patten with a learnt patten quality assessment function, which avoids dependency parsing and high-quality patten; (2) it identifies and groups synonymous meta patten from multiple facets—-their types, contexts, and extractions; and (3) it examines type distributions of entities in the instances extracted by each group of patten, and looks for appropriate type levels to make discovered precise. Experiments demonstrate that our proposed framework discovers high-quality typed textual patten efficiently from different genres of massive corpora and facilitates information extraction.

http://dl.acm.org/authorize?N33378

133. Tripoles: A New Class of Relationships in Time Series Data

Saurabh Agrawal (University of Minnesota);Gowtham Atluri (University of Cincinnati);Anuj Karpatne (University of Minnesota);William Haltom (University of Minnesota);Stefan Liess (University of Minnesota);Snigdhansu Chatterjee (University of Minnesota);Vipin Kumar (Univesity of Minnesota)

Relationship mining in time series data is one of the research directions that is of immense interest to several disciplines. Traditionally relationships that are studied in spatio-temporal data are between pairs of distant locations or regions. In this work, we define a novel relationship pattern over three time series which we refer to as a *tripole* that involves three time series. We show that tripoles can capture interesting relationships in the data that cannot be captured using traditionally studied pair-wise relationships. We propose a novel approach for finding tripoles in a given time-series dataset and demonstrate its computationally efficiency compared to the brute-force search on a real-world dataset from climate science domain. In addition, we show that tripoles could be found in real-world datasets from various domains including climate science and neuroscience. Furthermore, we found that most of the discovered tripoles are statistically significant and reproducible across multiple datasets that were completely independent to the original datasets that were used to find the tripoles. One of such discovered tripoles in climate data led to the discovery of a new climate teleconnection between Siberia and Pacific Ocean that was previously unknown in the climate domain.

http://dl.acm.org/authorize?N33350

134. Unsupervised Feature Selection in Signed Social Networks

Kewei Cheng (arizona state university);Jundong Li (arizona state university);Huan Liu (arizona state university)

The rapid growth of social media services brings large amounts of high-dimensional social media data at an unprecedented rate. Feature selection has shown to be powerful to prepare high-dimensional data for effective machine learning tasks. A majority of existing feature selection algorithms for social media data exclusively focus on positive interactions among linked instances. However, in many real-world social networks, instances may also be negatively interconnected. Recent work shows that the leverage of negative links could improve various learning tasks. To take advantage of negative links, we study a novel problem of unsupervised feature selection in signed social networks and propose a novel framework SignedFS. In particular, we provide a principled way to model positive and negative links for user preference learning. Then we embed the user preference learning into feature selection. Also, we revisit the homophily effect and balance theory in signed social networks and incorporate signed graph regularization into the feature selection framework to capture the first-order proximity and the second-order proximity in signed social networks. Experiments on real-world signed social networks demonstrate the effectiveness of our proposed framework. Further experiments are conducted to understand the impacts of negative links for feature selection.

http://dl.acm.org/authorize?N33368

135. Scalable Top-n Local Outlier Detection

Yizhou Yan (Worcester Polytechnic Institute);Lei Cao (Massachusetts Institute of Technology);Elke Rundensteiner (Worcester Polytechnic Institute)

Local Outlier Factor (LOF) method that labels all points with their respective LOF scores to indicate their status is known to be very effective for identifying outliers in datasets with a skewed distribution. Since outliers by definition are the absolute minority in a dataset, the concept of Top-N local outlier was proposed to discover the n points with the largest LOF scores. The detection of the Top-N local outliers is prohibitively expensive, since it requires huge number of high complexity k-nearest neighbor (kNN) searches. In this work, we present the first scalable Top-N local outlier detection approach called TOLF. The key innovation of *TOLF* is a multi-granularity pruning strategy that quickly prunes most points from the set of potential outlier candidates without computing their exact LOF scores or even without conducting any kNN search for them. Our customized density-aware indexing structure not only effectively supports the pruning strategy, but also accelerates the kNN search. Our extensive experimental evaluation on OpenStreetMap, SDSS, and TIGER datasets demonstrates the effectiveness of TOLF − up to 35 times faster than the state-of-the-art methods.

http://dl.acm.org/authorize?N33304

136. LiJAR: A System for Job Application Redistribution towards Efficient Career Marketplace

Fedor Borisyuk (LinkedIn Corporation);Liang Zhang (LinkedIn Corporation);Krishnaram Kenthapadi (LinkedIn Corporation)

Online professional social networks such as LinkedIn serve as a marketplace, wherein job seekers can find right career opportunities and job providers can reach out to potential candidates. LinkedIn's job recommendations product is a key vehicle for efficient matching between potential candidates and job postings. However, we have observed in practice that a subset of job postings receive too many applications (due to several reasons such as the popularity of the company, nature of the job, etc.), while some other job postings receive too few applications. Both cases can result in job poster dissatisfaction and may lead to discontinuation of the associated job posting contracts. At the same time, if too many job seekers compete for the same job posting, each job seeker's chance of getting this job will be reduced. In the long term, this reduces the chance of users finding jobs that they really like on the site. Therefore, it becomes beneficial for the job recommendation system to consider values provided to both job seekers as well as job posters in the marketplace.

In this paper we propose the job application redistribution problem, with the goal of ensuring that job postings do not receive too many or too few applications, while still providing job recommendations to users with the same level of relevance. We present a dynamic forecasting model to estimate the expected number of applications at the job expiration date, and algorithms to either promote or penalize jobs based on the output of the forecasting model. We also describe the system design and architecture for LiJAR, LinkedIn's Job Applications Forecasting and Redistribution system, which we have implemented and deployed in production. We perform extensive evaluation of LiJAR through both offline and online A/B testing experiments. Our production deployment of this system as part of LinkedIn's job recommendation engine has resulted in significant increase in the engagement of users for underserved jobs (6.5

http://dl.acm.org/authorize?N33329

137. Cascade Ranking for Operational E-commerce Search

Shichen Liu (Alibaba Group);Fei Xiao (Alibaba Group);Wenwu Ou (Alibaba Group);Luo Si (Alibaba Group)

In the 'Big Data' era, many real-world applications like search involve the ranking problem for a large number of items. It is important to obtain effective ranking results and at the same time obtain the results efficiently in a timely manner for providing good user experience and saving computational costs. Valuable prior research has been conducted for learning to efficiently rank like

the cascade ranking (learning) model, which uses a sequence of ranking functions to progressively filter some items and rank the remaining items. However, most existing research of learning to efficiently rank in search is studied in a relatively small computing environments with simulated user queries.

This paper presents novel research and thorough study of designing and deploying a Cascade model in a Large-scale Operational E-commerce Search application (CLOES), which deals with hundreds of millions of user queries per day with hundreds of servers. The challenge of the real-world application provides new insights for research: 1). Real-world search applications often involve multiple factors of preferences or constraints with respect to user experience and computational costs such as search accuracy, search latency, size of search results and total CPU cost, while most existing search solutions only address one or two factors; 2). Effectiveness of e-commerce search involves multiple types of user behaviors such as click and purchase, while most existing cascade ranking in search only models the click behavior. Based on these observations, a novel cascade ranking model is designed and deployed in an operational e-commerce search application. An extensive set of experiments demonstrate the advantage of the proposed work to address multiple factors of effectiveness, efficiency and user experience in the real-world application.

http://dl.acm.org/authorize?N33345

138. A Century of Science: Globalization of Scientific Collaborations, Citations, and Innovations

Yuxiao Dong (University of Notre Dame);Hao Ma (Microsoft Research);Zhihong Shen (MSR);Kuansan Wang (MSR)

Progress in science has advanced the development of human society across history, with dramatic revolutions shaped by information theory, genetic cloning, and artificial intelligence among the many scientific achievements produced in the 20th century. However, the way that science advances itself is much less well-understood. In this work we study the evolution of scientific development over the past century by presenting an anatomy of 89 million digitalized papers published between 1900 and 2015. We find that science has benefited from the shift from individual work to collaborative effort, with over 90

http://dl.acm.org/authorize?N33323

139. "Not All Passes Are Created Equal:" Objectively Measuring the Risk and Reward of Passes in Soccer from Tracking Dat

Paul Power (STATS LLC);Héctor Ruiz (STATS);Patrick Lucey (STATS);Xinyu Wei (STATS)

In soccer, the most frequent event that occurs is a pass. For a trained eye, there are a myriad of adjectives which could describe this event (e.g., "majestic pass", "conservative" to "poor-ball"). However, as these events are needed to be coded live and in real-time (most often by human annotators), the current method of grading passes is restricted to the binary labels 0 (unsuccessful) or 1 (successful). Obviously, this is sub-optimal because the quality of a pass needs to be measured on a continuous spectrum (i.e., 0 → 100

http:/dl.acm.org/authorize?N33340

140. No Longer Sleeping with a Bomb: A Duet System for Protecting Urban Safety from Dangerous Goods

Jingyuan Wang (Beihang University);Chao Chen (Beihang University);Junjie Wu (Beihang University);Zhang Xiong (Beihang University)

Recent years have witnessed the continuous growth of megalopolises worldwide, which makes urban safety a top priority in modern city life. Among various threats, dangerous goods such as gas and hazardous chemicals transported through and around cities have increasingly become the deadly "bomb" we sleep with every day. In both academia and government, tremendous efforts have been dedicated to dealing with dangerous goods transportation (DGT) issues, but further study is still in great need to quantify the problem and explore its intrinsic dynamics in a big data perspective. In this paper, we present a novel system called DGeye, which features a "duet" between DGT trajectory data and human mobility data for risky zones identification. Moreover, DGeye innovatively takes risky patterns as the keystones in DGT management, and builds causality networks among them for pain points identification, attribution and prediction. Experiments on both Beijing and Tianjin cities demonstrate the effectiveness of DGeye. In particular, DGeye after deployment has driven the Beijing government to lay down gas pipelines for the famous Guijie food street.

http://dl.acm.org/authorize?N33457

141. MARAS: Signaling Multi-Drug Adverse Reactions.

Xiao Qin (Worcester Polytechnic Institute & Oak Ridge Institute of Science);Tabassum Kakar (Worcester Polytechnic Institute & Oak Ridge Institute of Science);Susmitha Wunnava (Worcester Polytechnic Institute);Elke Rundensteiner (Worcester Polytechnic Institute);Cao Lei (Massachusetts Institute of Technology)

There is a growing need for computing-supported methods that facilitate the automated signaling of Adverse Drug Reactions (ADRs) otherwise left undiscovered from the exploding amount of ADR reports filed by patients, medical professionals and drug manufacturers. In this research, we design a Multi-Drug Adverse Reaction Analytics Strategy, called MARAS, to signal severe unknown ADRs triggered by the usage of a combination of drugs, also known as Multi-Drug Adverse Reactions (MDAR). First, MARAS features an efficient signal generation algorithm based on association rule learning that extracts non-spurious MDAR associations. Second, MARAS incorporates contextual information to detect drug combinations that are strongly associated with a set of ADRs. It groups related associations into Contextual Association Clusters (CACs) that then avail contextual information to evaluate the significance of the discovered MDAR Associations. Lastly, we use this contextual significance to rank discoveries by their notion of interestingness to signal the most compelling MDARs. To demonstrate the utility of MARAS, it is compared with state-of-the-art techniques and evaluated via case studies on datasets collected by U.S. Food and Drug Administration Adverse Event Reporting System (FAERS).

http://dl.acm.org/authorize?N33341

142. MOLIERE: Automatic Biomedical Hypothesis Generation System

Justin Sybrandt (Clemson University);Michael Shtutman (University of South Carolina);Ilya Safro (Clemson University)

Hypothesis generation is becoming a crucial time-saving technique which allows biomedical researchers to quickly discover implicit connections between important concepts. Typically, these systems operate on domain-specific fractions of public medical data. MOLIERE, in contrast, utilizes information from over 24.5 million documents. At the heart of our approach lies a multi-modal and multi-relational network of biomedical objects extracted from several heterogeneous datasets from the National Center for Biotechnology Information (NCBI). These objects include but are not limited to scientific papers, keywords, genes, proteins, diseases, and diagnoses. We model hypotheses using Latent Dirichlet Allocation applied on abstracts found near shortest paths discovered within this network, and demonstrate the effectiveness of MOLIERE by performing hypothesis generation on historical data. Our network, implementation, and resulting data are all publicly available for the broad scientific community.

http://dl.acm.org/authorize?N33343

143. A Data Science Approach to Understanding Residential Water Contamination in Flint

Jacob Abernethy (University of Michigan, Ann Arbor);Alex Chojnaki (Michigan Data Science Team);Chengyu Dai (Michigan Data Science Team);Arya Farahi (Michigan Data Science Team);Eric Schwartz (Ross School of Busi-

ness);Jared Webb (Brigham Young University);Guangsha Shi (University of Michigan);Daniel T. Zhang (Michigan Data Science Team)

The Flint Water Crisis was followed by a huge investment by residents and government officials to sample and test the water in Flint homes in order to understand the causes and extent of the lead contamination. This trove of data, most of which was made publicly available, is by far the largest dataset collected on lead in a municipality water system. In this paper we study several aspects of Flint's water troubles, and we lay out a number of analytical and algorithmic results on lead poisoning, many of which generalize well beyond one city. For example, we show that elevated lead risks are surprisingly predictable, to a reasonable extent, and we explore various factors associated with elevated lead. These risk assessments, developed in large part via a crowdsourced prediction challenge at the University of Michigan, have been incorporated into an informational web and mobile application, funded by Google.org, designed to target Flint residents. We also explore questions of self-selection in the residential testing program, and what factors induce residents to voluntarily sample their water, when they test, and how often.

http://dl.acm.org/authorize?N33320

144. Peeking at A/B Tests: Why it matters, and what to do about it

David Walsh (Stanford University);Ramesh Johari (Stanford University);Leonid Pekelis (Stanford University)

This paper reports on novel statistical methodology, which has been deployed by the commercial A/B testing platform Optimizely to communicate experimental results to their customers. Our methodology addresses the issue that traditional p-values and confidence intervals give unreliable inference. This is because users of A/B testing software are known to *continuously monitor* these measures as the experiment is running. We provide *always valid* p-values and confidence intervals that are provably robust to this effect. Not only does this make it safe for a user to continuously monitor, but it empowers her to detect true effects more efficiently. This paper provides simulations and numerical studies on Optimizely's data, demonstrating an improvement in detection performance over traditional methods.

http://dl.acm.org/authorize?N33331

145. Pharmacovigilance via Baseline Regularization with Large-Scale Longitudinal Observational Data

Zhaobin Kuang (University of Wisconsin, Madison);Peggy Peissig (Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI);Vitor Santos Costa (Universidade do Porto);Richard Maclin (University of Minnesota, Duluth);David Page (Department of Computer Sciences and Department of Biostatistics, University of Wisconsin, Madison, WI)

Several prominent public health hazards that occurred at the beginning of this century due to adverse drug events (ADEs) have raised international awareness of governments and industries about pharmacovigilance (PhV), the science and activities to monitor and prevent adverse events caused by pharmaceutical products after they are introduced to the market. A major data source for PhV is large-scale longitudinal observational databases (LODs) such as electronic health records (EHRs) and medical insurance claim databases. Inspired by the Self-Controlled Case Series (SCCS) model, arguably the leading method for ADE discovery from LODs, we propose baseline regularization, a regularized generalized linear model that leverages the diverse health profiles available in LODs across different individuals at different times. We apply the proposed method as well as SCCS to the Marshfield Clinic EHR. Experimental results suggest that the proposed method outperforms SCCS under various settings in identifying benchmark ADEs from the Observational Medical Outcomes Partnership ground truth.

http://dl.acm.org/authorize?N33333

146. The Simpler The Better: A Unified Approach to Predicting Original Taxi Demands on Large-Scale Online Platforms

Yongxin Tong (Beihang University);Yuqiang Chen (4Paradigm Inc.);Zimu Zhou (ETH Zurich);Lei Chen (Hong Kong University of Science and Technology);Jie Wang (Didi Research);Qiang Yang (Hong Kong University of Science and Technology);Jieping Ye (Didi Research)

Taxi-calling apps are gaining increasing popularity for their efficiency in dispatching idle taxis to passengers in need. To precisely balance the supply and the demand of taxis, it is essential for the online taxicab platforms to predict Unit Original Taxi Demands (UOTD), which refers to the number of taxi-calling requirements submitted per unit time (e.g. every hour) and per unit region (e.g. each POI). Prediction of UOTD is non-trivial for large-scale industrial online taxicab platforms because both accuracy and flexibility are essential. Complex non-linear models such as GBRT and deep learning are generally accurate, yet labor-intensive model redesign is indispensable after scenario changes (e.g. extra constraints due to new regulations). To accurately predict UOTD while remaining flexible to scenario changes, we propose LinUOTD, a unified linear regression model with more than 200 million dimensions of features. The simple model structure eliminates the need of repeatedly model redesign, while the high-dimensional features contribute to accurate UOTD prediction. Furthermore, we design a series of optimization techniques for efficient model training and updating. Evaluations on two large-scale datasets from an industrial online taxicab platform verify that LinUOTD outperforms popular non-linear models in accuracy. We envision our experiences to adopt simple linear models with high-dimensional features in UOTD prediction as a pilot study and can shed insights upon other industrial large-scale spatio-temporal prediction problems.

http://dl.acm.org/authorize?N33455

147. Deep Choice Model Using Pointer Networks for Airline Itinerary Prediction

Alejandro Mottini (Amadeus SAS);Rodrigo Acuna-Agost (Amadeus SAS)

Travel providers such as airlines and on-line travel agents are becoming more and more interested in understanding how passengers choose among alternative itineraries when searching for flights. This knowledge helps them better display and adapt their offer, taking into account market conditions and customer needs. Some common applications are not only filtering and sorting alternatives, but also changing certain attributes in real-time (e.g., changing the price). In this paper, we concentrate with the problem of modeling air passenger choices of flight itineraries. This problem has historically been tackled using classical Discrete Choice Modelling techniques. Traditional statistical approaches, in particular the Multinomial Logit model (MNL), is widely used in industrial applications due to its simplicity and general good performance. However, MNL models present several shortcomings and assumptions that might not hold in real applications. To overcome these difficulties, we present a new choice model based on Pointer Networks. Given an input sequence, this type of deep neural architecture combines Recurrent Neural Networks with the Attention Mechanism to learn the conditional probability of an output whose values correspond to positions in an input sequence. Therefore, given a sequence of different alternatives presented to a customer, the model can learn to point to the one most likely to be chosen by the customer. The proposed method was evaluated on a real dataset that combines on-line user search logs and airline flight bookings. Experimental results show that the proposed model outperforms the traditional MNL model on several metrics.

http://dl.acm.org/authorize?N33347

148. DeepSD: Generating High Resolution Climate Change Projections through Single Image Super-Resolution

Thomas Vandal (Northeastern University);Evan Kodra (risQ Inc.);Sangram Ganguly (Bay Area Environmental Research Institute / NASA Ames Research Center);Andrew Michaelis (University Corporation, Monterey Bay);Ramakrishna Nemani (NASA Ames Research Center);Auroop Ganguly (Northeastern University)

The impacts of climate change are felt by most critical systems, such as infrastructure, ecological systems, and power-plants. However, contemporary Earth System Models (ESM) are run at spatial resolutions too coarse for assessing effects this localized. Local scale projections can be obtained using statistical downscaling, a technique which uses historical climate observations to learn a low-resolution to high-resolution mapping. Depending on statistical modeling choices, downscaled projections have been shown to vary significantly terms of accuracy and reliability. The spatio-temporal nature of the climate system motivates the adaptation of super-resolution image processing techniques to statistical downscaling. In our work, we present DeepSD, a generalized stacked super resolution convolutional neural network (SRCNN) framework for statistical downscaling of climate variables. DeepSD augments SRCNN with multi-scale input channels to maximize predictability in statistical downscaling. We provide a comparison with Bias Correction Spatial Disaggregation as well as three Automated-Statistical Downscaling approaches in downscaling daily precipitation from 1 degree (100km) to 1/8 degrees (12.5km) over the Continental United States. Furthermore, a framework using the NASA Earth Exchange (NEX) platform is discussed for downscaling more than 20 ESM models with multiple emission scenarios.

http://dl.acm.org/authorize?N33456

149. An End-to-End Event Log Analysis Platform for System Management

Tao Li (Florida International University);Bin Xia (Nanjing University of Science and technology)

Many systems, such as distributed operating systems, complex networks, and high throughput web-based applications, are continuously generating large volume of event logs. These logs contain useful information to help system administrators to understand the system running status and to pinpoint the system failures. Generally, due to the scale and complexity of modern systems, the generated logs are beyond the analytic power of human beings. Therefore, it is imperative to develop a comprehensive log analysis system to support effective system management. Although a number of log mining techniques have been proposed to address specific log analysis use cases, few research and industrial efforts have been paid on providing integrated systems with an end-to-end solution to facilitate the log analysis routines. In this paper, we design and implement an integrated system, called FIU Log Analysis Platform (a.k.a. FLAP), that aims to facilitate the data analytics for system event logs. FLAP provides an end-to-end solution that utilizes advanced data mining techniques to assist log analysts to conveniently, timely, and accurately conduct event log knowledge discovery, system status investigation, and system failure diagnosis. Specifically, in FLAP, state-of-the-art template learning techniques are used to extract useful information from unstructured raw logs; advanced data transformation techniques are proposed and leveraged for event transformation and storage; effective event pattern mining, event summarization, event querying, and failure prediction techniques are designed and integrated for log analytics; and user-friendly interfaces are utilized to present the informative analysis results intuitively and vividly. Since 2015, FLAP has been used by Huawei Technologies Co. Ltd for internal event log analysis, and has provided effective support in its system operation and workflow optimization.

http://dl.acm.org/authorize?N33334

150. Luck is hard to beat: The Difficulty of Sports Prediction

Raquel Aoki (Universidade Federal de Minas Gerais);Renato Assunção (Universidade Federal de Minas Gerais);Pedro Vaz de Melo (Universidade Federal de Minas Gerais)

Predicting the outcome of sports events is a hard task. We quantify this difficulty with a coefficient that measures the distance between the observed final results of sports leagues and idealized perfectly balanced competitions in terms of skill. This indicates the relative presence of luck and skill. We collected and analyzed all games from 198 sports leagues comprising 1503 seasons from 84 countries of 4 different sports: basketball, soccer, volleyball and handball. We measured the competitiveness by countries and sports. We also identify in each season which teams, if removed from its league, result in a completely random tournament. Surprisingly, not many of them are needed. As another contribution of this paper, we propose a probabilistic graphical model to learn about the teams' skills and to decompose the relative weights of luck and skill in each game. We break down the skill component into factors associated with the teams' characteristics. The model also allows to estimate as 0.36 the probability that an underdog team wins in the NBA league, with a home advantage adding 0.09 to this probability. As shown in the first part of the paper, luck is substantially present even in the most competitive championships, which partially explains why sophisticated and complex feature-based models hardly beat simple models in the task of forecasting sports' outcomes.

http://dl.acm.org/authorize?N33326

151. Google Vizier: A Service for Black-Box Optimization

Daniel Golovin (Google, Inc.);Benjamin Solnik (Google, Inc.);Subhodeep Moitra (Google, Inc.);Greg Kochanski (Google, Inc.);John Karro (Google, Inc.);D. Sculley (Google, Inc.)

Any sufficiently complex system acts as a black box when it becomes easier to experiment with than to understand. Hence, black-box optimization has become increasingly important as systems have become more complex. In this paper we describe Google Vizier, a Google-internal service for performing black-box optimization that has become the de facto parameter tuning engine at Google. Google Vizier is used to optimize many of our machine learning models and other systems, and also provides core capabilities to Google's Cloud Machine Learning HyperTune subsystem. We discuss our requirements, infrastructure design, underlying algorithms, and advanced features such as transfer learning and automated early stopping that the service provides.

http://dl.acm.org/authorize?N33338

152. PNP: Fast Path Ensemble Method for Movie Design

Danai Koutra (University of Michigan);Abhilash Dighe (University of Michigan);Smriti Bhagat (Facebook);Udi Weinsberg (Facebook);Stratis Ioannidis (Northeastern University);Christos Faloutsos (Carnegie Mellon University);Jean Bolot (Technicolor)

How can we design a product or movie that will attract, for example, the interest of Pennsylvania adolescents or liberal newspaper critics? What should be the genre of that movie and who should be in the cast? In this work, we seek to identify how we can design new movies with features tailored to a specific user population. We formulate the movie design as an optimization problem over the inference of user-feature scores and selection of the features that maximize the number of attracted users. Our approach, PNP, is based on a heterogeneous, tripartite graph of users, movies and features (e.g., actors, directors, genres), where users rate movies and features contribute to movies. We learn the preferences by lever- aging user similarities defined through different types of relations, and show that our method outperforms state-of-the-art approaches, including matrix factorization and other heterogeneous graph-based analysis. We evaluate PNP on publicly available real-world data and show that it is highly scalable and effectively provides movie designs oriented towards different groups of users, including men, women, and adolescents.

http://dl.acm.org/authorize?N33332

153. Quick Access: Building a Smart Experience for Google Drive

Sandeep Tata (Google Inc.);Alexandrin Popescul (Google Inc.);Marc Najork (Google Inc.);Mike Colagrosso (Google Inc.);Julian Gibbons (Google Inc.);Alan Green (Google Inc.);Alexandre Mah (Google Inc.);Michael Smith (Google Inc.);Divanshu Garg (Google Inc.);Cayden Meyer (Google Inc.);Reuben Kan (Google Inc.)

Google Drive is a cloud storage and collaboration service used by hundreds of millions of users around the world. Quick Access is a new feature in Google Drive that surfaces the right documents when a user visits the home screen. Our metrics show that users locate their documents in half the time with this feature compared to previous approaches. The development of Quick Access illustrates many general challenges and constraints associated with practical machine learning such as protecting user privacy, working with data services that are not designed with machine-learning in mind, and evolving product definitions. We believe that the lessons learned from this experience will be useful to practitioners tackling a wide range of applied machine-learning problems.

http://dl.acm.org/authorize?N33344

154. Deep Embedding Forest: Forest-based Serving with Deep Embedding Features

Jie Zhu (Microsoft Corporation);Ying Shan (Microsoft Corporation);Jc Mao (Microsoft Corporation);Dong Yu (Microsoft Corporation);Holakou Rahmanian (University of California Santa Cruz);Yi Zhang (Microsoft Corporation)

Deep Neural Networks (DNN) have demonstrated superior ability to extract high level embedding vectors from low level features. Despite the success, the serving time is still the bottleneck due to expensive run-time computation of multiple layers of dense matrices. GPGPU, FPGA, or ASIC-based serving systems require additional hardware that are not in the mainstream design of most commercial applications. In contrast, tree or forest-based models are widely adopted because of low serving cost, but heavily depend on carefully engineered features. This work proposes a Deep Embedding Forest model that benefits from the best of both worlds. The model consists of a number of embedding layers and a forest/tree layer. The former maps high dimensional (hundreds of thousands to millions) and heterogeneous low-level features to the lower dimensional (thousands) vectors, and the latter ensures fast serving.

Built on top of a representative DNN model called Deep Crossing, and two forest/tree-based models including XGBoost and LightGBM, a two-step Deep Embedding Forest algorithm is demonstrated to achieve on-par or slightly better performance as compared with the DNN counterpart, with only a fraction of serving time on conventional hardware. After comparing with a joint optimization algorithm called partial fuzzification, also proposed in this paper, it is concluded that the two-step Deep Embedding Forest has achieved near optimal performance. Experiments based on large scale data sets (up to 1 billion samples) from a major sponsored search engine proves the efficacy of the proposed model.

http://dl.acm.org/authorize?N33450

155. KunPeng: Parameter Server based Distributed Learning Systems and Its Applications in Alibaba and Ant Financial

Jun Zhou (Ant Financial Group);Xiaolong Li (Ant Financial Group);Peilin Zhao (Ant Financial Group);Chaochao Chen (Ant Financial Group);Longfei Li (Ant Financial Group);Xinxing Yang (Ant Financial Group);Qing Cui (Alibaba Cloud);Jin Yu (Alibaba Cloud);Xu Chen (Alibaba Cloud);Yi Ding (Alibaba Cloud);Yuan Qi (Ant Financial Group)

In recent years, due to the emergence of Big Data (terabytes or petabytes) and Big Model (tens of billions of parameters), there has been an ever-increasing need of parallelizing machine learning (ML) algorithms in both academia and industry. Although there are some existing distributed computing systems, like Hadoop and Spark, for parallelizing ML algorithms, they only provide synchronous and coarse-grained operators (e.g., Map, Reduce, and Join, etc.), which may hinder developers from implementing more efficient algorithms. This motivated us to design a universal distributed platform termed KunPeng, that combines both distributed systems and parallel optimization algorithms to deal with the complexities that arise from large-scale ML. Specifically, KunPeng not only encapsulates the characteristics of data/model parallelism, load balancing, model sync-up, sparse representation, industrial fault-tolerance, etc., but also provides easy-to-use interface to empower users to focus on the core ML logics. Empirical results on terabytes of real datasets with billions of samples and features demonstrate that, such a design brings compelling performance improvements on ML programs ranging from Follow-the-Regularized-Leader Proximal algorithm to Sparse Logistic Regression and Multiple Additive Regression Trees. Furthermore, KunPeng's encouraging performance is also shown for several real-world applications including the Alibaba's Double 11 Online Shopping Festival.

http://dl.acm.org/authorize?N33459

156. HinDroid: An Intelligent Android Malware Detection System Based on Structured Heterogeneous Information Network

Yanfang Ye (West Virginia University);Shifu Hou (West Virginia University);Yangqiu Song (West Virginia University)

With explosive growth of Android malware and due to the severity of its damages to smart phone users, the detection of Android malware has become an increasingly important topic in cyber security. The increasing sophistication of Android malware calls for new defensive techniques that are harder to evade, and are capable of protecting users against novel threats. In this paper, to detect Android malware, instead of using Application Programming Interface (API) calls only, we further analyze the different relationships between them and create higher-level semantics which require more efforts for attackers to evade the detection. We represent the Android applications (apps), related APIs, and their rich relationships as a structured heterogeneous information network (HIN). Then we use a meta-path based approach to characterize the semantic relatedness of apps and APIs. We use each meta-path to formulate a similarity measure over Android apps, and aggregate different similarities using multi-kernel learning. Then each meta-path is automatically weighted by the learning algorithm to make predictions. To the best of our knowledge, this is the rest work to use structured HIN for Android malware detection. Comprehensive experiments on real sample collections from Comodo Cloud Security Center are conducted to compare various malware detection approaches. Promising experimental results demonstrate that our developed system HinDroid system outperforms other alternative Android malware detection techniques. HinDroid has already been incorporated into the scanning tool of Comodo Mobile Security product.

http://dl.acm.org/authorize?N33330

157. Compass: Spatio Temporal Sentiment Analysis of US Election

Debjyoti Paul (University of Utah);Feifei Li (University of Utah);Murali Krishna Teja Kilari (University of Utah);Xin Yu (University of Utah);Richie Frost (University of Utah)

With the widespread growth of various social network tools and platforms, analyzing and understanding societal response and crowd reaction to important and emerging social issues and events through social media data is increasingly an important problem. However, there are numerous challenges towards realizing this goal effectively and efficiently, due to the unstructured and noisy nature of social media data. The large volume of the underlying data also presents a fundamental challenge. Furthermore, in many application scenarios, it is often interesting, and in some cases critical, to discover patterns and trends based on geographical and/or temporal partitions, and keep track of how they will change overtime. This brings up the interesting problem of spatio-temporal sentiment analysis from large-scale social media data. This paper investigates this problem through a data science project called “US Election 2016, What Twitter Says”‘. The objective is to discover sentiment on twitter towards either the democratic or the republican party at US county and state levels over any arbitrary temporal intervals, using a large collection of geotagged tweets from a period of 6 months leading up to the US presidential election in 2016. Our results demonstrate that by integrating and developing a combination of machine learning and data management techniques, it is possible to do this at scale with effective outcomes. The results of our project have the potential to be adapted towards solving and influencing other interesting social issues such as building neighborhood happiness and health indicators.

http://dl.acm.org/authorize?N33348

158. Guided Deep List: Automating the Generation of Epidemiological Line Lists from Open Sources

Saurav Ghosh (Virginia Tech);Prithwish Chakraborty (Virginia Tech);Bryan Lewis (Virginia Tech);Maia Majumder (Massachusetts Institute of Technology);Emily Cohn (Boston Children’s Hospital);John Brownstein (Harvard Medical School);Madhav Marathe (Virginia Tech);Naren Ramakrishnan (Virginia Tech)

Real-time monitoring and responses to emerging public health threats rely on the availability of timely surveillance data. During the early stages of an epidemic, the ready availability of line lists with detailed tabular information about laboratory-confirmed cases can assist epidemiologists in making reliable inferences and forecasts. Such inferences are crucial to understand the epidemiology of a specific disease early enough to stop or control the outbreak. However, construction of such line lists requires considerable human supervision and therefore, difficult to generate in real-time. In this paper, we motivate Guided Deep List, the first tool for building automated line lists (in near real-time) from open source reports of emerging disease outbreaks. Specifically, we focus on deriving epidemiological characteristics of an emerging disease and the affected population from reports of illness. Guided Deep List uses distributed vector representations (ala word2vec) to discover a set of indicators for each line list feature. This discovery of indicators is followed by the use of dependency parsing based techniques for final extraction in tabular form. We evaluate the performance of Guided Deep List against a human annotated line list provided by HealthMap corresponding to MERS outbreaks in Saudi Arabia. We demonstrate that Guided Deep List extracts line list features with increased accuracy compared to a baseline method. We further show how these automatically extracted line list features can be used for making epidemiological inferences, such as inferring demographics and symptoms-to-hospitalization period of affected individuals.

http://dl.acm.org/authorize?N33337

159. Predicting Clinical Outcomes Across Changing Electronic Health Record Systems

Jen Gong (MIT);Tristan Naumann (MIT);Peter Szolovits (MIT);John Guttag (MIT)

Existing machine learning methods typically assume consistency in how information is encoded. However, the way information is recorded in databases differs across institutions and over time, rendering potentially useful data obsolescent. To address this problem, we map database-specific representations of information to a common set of semantic concepts, thus allowing models to transition across different databases.

http://dl.acm.org/authorize?N33339

160. Prognosis and Diagnosis of Parkinson's Disease Using Multi-Task Learning

Saba Emrani (SAS Institute Inc);Anya McGuirk (SAS Institute Inc.);Wei Xiao (SAS Institute Inc.)

Parkinson's disease (PD) is a debilitating neurodegenerative disease excessively affecting millions of patients. Early diagnosis of PD is critical as manifestation of symptoms occur many years after the onset of neurodegenration, when more than 60

http://dl.acm.org/authorize?N33335

161. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale

Adrian Albert (MIT & SLAC);Marta Gonzalez (MIT)

Urban planning applications (energy audits, investment, etc.) require an understanding of built infrastructure and its environment, i.e., both low-level, physical features (amount of vegetation, building area and geometry etc.), as well as higher-level concepts such as land use classes (which encode expert understanding of socio-economic end uses). This kind of data is expensive and labor-intensive to obtain, which limits its availability (particularly in developing countries). We analyze patterns in land use in urban neighborhoods using large-scale satellite imagery data (which is available worldwide from third-party providers) and state-of-the-art computer vision techniques based on deep convolutional neural networks. For supervision, given the limited availability of standard benchmarks for remote-sensing data, we obtain ground truth land use class labels carefully sampled from open-source surveys, in particular the Urban Atlas land classification dataset of

20 land use classes across 300 European cities. We use this data to train and compare deep architectures which have recently shown good performance on standard computer vision tasks (image classification and segmentation), including on geospatial data. Furthermore, we show that the deep representations extracted from satellite imagery of urban environments can be used to compare neighborhoods across several cities. We make our dataset available for other machine learning researchers to use for remote-sensing applications.

http://dl.acm.org/authorize?N33325

162. FIRST: Fast Interactive Attributed Subgraph Matching

Boxin Du (Arizona State University);Si Zhang (Arizona State University);Nan Cao (Tongji University);Hanghang Tong (Arizona State University)

Attributed subgraph matching is a powerful tool for explorative mining of large attributed networks. In many applications (e.g., network science of teams, intelligence analysis, finance informatics), the user might not know what exactly s/he is looking for, and thus require the user to constantly revise the initial query graph based on what s/he finds from the current matching results. A major bottleneck in such an interactive matching scenario is the efficiency, as simply re-running the matching algorithm on the revised query graph is computationally prohibitive. In this paper, we propose a family of effective and efficient algorithms (FIRST) to support interactive attributed subgraph matching. There are two key ideas behind the proposed methods. The first is to recast the attributed subgraph matching problem as a cross-network node similarity problem, whose major computation lies in solving a Sylvester equation for the query graph and the underlying data graph. The second key idea is to explore the smoothness between the initial and revised queries, which allows us to solve the new/updated Sylvester equation incrementally, without re-solving it from scratch. Experimental results show that our method can achieve (1) up to 16 times speed-up when applying on networks with 6M+ nodes; (2) preserving more than 90

http://dl.acm.org/authorize?N33324

163. A Dirty Dozen: Twelve Common Metric Interpretation Pitfalls in Online Controlled Experiments

Somit Gupta (Microsoft);Pavel Dmitriev (Microsoft);Garnet Vaz (Microsoft);Dong Woo Kim (Microsoft)

Online controlled experiments (e.g., A/B tests) are now regularly used to guide product development and accelerate innovation in software. Product ideas are evaluated as scientific hypotheses, and tested in web sites, mobile applications, desktop applications, services, and operating systems. One of the key challenges for organizations that run controlled experiments is to come up with the right set of metrics. Having good metrics, however, is not enough. In our experience of running thousands of experiments with many teams across Microsoft, we observed again and again how incorrect interpretations of metric movements may lead to wrong conclusions about the experiment's outcome, which if deployed could hurt the business by millions of dollars. Inspired by Steven Goodman's twelve p-value misconceptions [4], in this paper, we share twelve common metric interpretation pitfalls which we observed repeatedly in our experiments. We illustrate each pitfall with a puzzling example from a real experiment, and describe processes, metric design principles, and guidelines that can be used to detect and avoid the pitfall. With this paper, we aim to increase the experimenters' awareness of metric interpretation issues, leading to improved quality and trustworthiness of experiment results and better data-driven decisions.

http://dl.acm.org/authorize?N33322

164. A Data Mining Framework for Valuing Large Portfolios of Variable Annuities

Guojun Gan (Department of Mathematics, University of Connecticut);Jimmy Huang (York University)

A variable annuity is a tax-deferred retirement vehicle created to address concerns that many people have about outliving their assets. In the past decade, the rapid growth of variable annuities has posed great challenges to insurance companies especially when it comes to valuing the complex guarantees embedded in these products.

In this paper, we propose a data mining framework to address the computational issue associated with the valuation of large portfolios of variable annuity contracts. The data mining framework consists of two major components: a data clustering algorithm which is used to select representative variable annuity contracts, and a regression model which is used to predict quantities of interest for the whole portfolio based on the representative contracts. A series of numerical experiments are conducted on a portfolio of synthetic variable annuity contracts to demonstrate the performance of our proposed data mining framework in terms of accuracy and speed. The experimental results show that our proposed framework is able to produce accurate estimates of various quantities of interest and can reduce the runtime significantly compared to the state-of-the-art approaches.

http://dl.acm.org/authorize?N33336

165. Developing a comprehensive framework for multimodal feature extraction

Quinten McNamara (University of Texas at Austin);Alejandro de La Vega (University of Texas at Austin);Tal Yarkoni (University of Texas at Austin)

Feature extraction is a critical component of many applied data science workflows. In recent years, rapid advances in artificial intelligence and machine learning have led to an explosion of feature extraction tools and services that allow data scientists to cheaply and effectively annotate their data along a vast array of dimensions—-ranging from detecting faces in images to analyzing the sentiment expressed in coherent text. Unfortunately, the proliferation of powerful feature extraction services has been mirrored by a corresponding expansion in the number of distinct interfaces to feature extraction services. In a world where nearly every new service has its own API, documentation, and/or client library, data scientists who need to combine diverse features obtained from multiple sources are often forced to write and maintain ever more elaborate feature extraction pipelines.

To address this challenge, we introduce a new open-source framework for comprehensive multimodal feature extraction. Pliers is an open-source Python package that supports standardized annotation of diverse data types (video, images, audio, and text), and is expressly with both ease-of-use and extensibility in mind. Users can apply a wide range of pre-existing feature extraction tools to their data in just a few lines of Python code, and can also easily add their own custom extractors by writing modular classes. A graph-based API enables rapid development of complex feature extraction pipelines that output results in a single, standardized format. We describe the package's architecture, detail its major advantages over previous feature extraction toolboxes, and use a sample application to a large functional MRI dataset to illustrate how pliers can significantly reduce the time and effort required to construct sophisticated feature extraction workflows while increasing code clarity and maintainability.

http://dl.acm.org/authorize?N33346

166. TFX: A TensorFlow-Based Production-Scale Machine Learning Platform

Denis Baylor (Google Inc.);Eric Breck (Google Inc.);Heng-Tze Cheng (Google Inc.);Noah Fiedel (Google Inc.);Chuan Yu Foo (Google Inc.);Zakaria Haque (Google Inc.);Salem Haykal (Google Inc.);Mustafa Ispir (Google Inc.);Vihan Jain (Google Inc.);Levent Koc (Google Inc.);Chiu Yuen Koo (Google Inc.);Lukasz Lew (Google Inc.);Clemens Mewald (Google Inc.);Akshay Modi (Google Inc.);Neoklis Polyzotis (Google Inc.);Sukriti Ramesh (Google Inc.);Sudip Roy (Google Inc.);Steven Whang (Google Inc.);Martin Wicke (Google Inc.);Jarek Wilkiewicz (Google Inc.);Xin Zhang (Google Inc.);Martin Zinkevich (Google Inc.)

Creating and maintaining a platform for reliably producing and deploying machine learning models requires careful orchestration of many components—-a learner for generating models based on training data, modules for analyzing and validating both data as well as models, and finally infrastructure for serving models in production. This becomes particularly challenging when data changes over time and fresh models need to be produced continuously. Unfortunately, such orchestration is often done ad hoc using glue code and custom scripts developed by individual teams

for specific use cases, leading to duplicated effort and fragile systems with high technical debt. We present the anatomy of a general-purpose machine learning platform and one implementation of such a platform at Google. By integrating the aforementioned components into one platform, we were able to standardize the components, simplify the platform configuration, and reduce the time to production from the order of months to weeks, while providing platform stability that minimizes service disruptions. We present the case study of one deployment of the platform in the Google Play app store, where the machine learning models are refreshed continuously as new data arrive. Deploying the platform led to reduced custom code, faster experiment cycles, and a 2

http:////dl.acm.org/authorize?N33328

167. Backpage and Bitcoin: Uncovering Human Traffickers

Rebecca S. Portnoff (UC Berkeley);Danny Yuxing Huang (UC San Diego);Periwinkle Doerfler (NYU);Sadia Afroz (ICSI);Damon McCoy (NYU)

Sites for online classified ads selling sex are widely used by human traffickers to support their pernicious business. The sheer quantity of ads makes manual exploration and analysis unscalable. In addition, discerning whether an ad is advertising a trafficked victim or a independent sex worker is a very difficult task. Very little concrete ground truth (i.e., ads definitively known to be posted by a trafficker) exists in this space. In this work, we develop tools and techniques that can be used separately and in conjunction to group sex ads by their true owner (and not the claimed author in the ad). Specifically, we develop a machine learning classifier that uses stylometry to distinguish between ads posted by the same vs. different authors with 96

http://dl.acm.org/authorize?N33349

168. Estimation of recent ancestral origins of individuals on a large scale

Ross E Curtis (AncestryDNA);Ahna R Girshick (AncestryDNA)

The last ten years have seen an exponential growth of direct-to-consumer genomics tests. One popular feature of these tests is the report of a distant ancestral inference profile—a breakdown of the regions of the world where the test-takers' ancestors may have lived. While current methods and products generally focus on the more distant past (e.g., thousands of years ago), we have recently demonstrated that by leveraging network analysis tools such as community detection, more recent ancestry can be identified. However, using a network analysis tool like community detection on a large network with potentially millions of nodes is not feasible in a live production environment where hundreds or thousands of new genotypes need to be processed every day. In this study, we describe a classification method that leverages network features to assign individuals to communities in a large network corresponding to recent ancestry. We will be launching a version of this research as a new product feature at AncestryDNA.

http://dl.acm.org/authorize?N33321

169. A Practical Exploration System for Search Advertising

Parikshit Shah (Yahoo Research);Ming Yang (Yahoo);Sachidanand Alle (Yahoo);Adwait Ratnaparkhi (Yahoo Research);Ben Shahshahani (Yahoo Research);Rohit Chandra (Yahoo)

In this paper, we describe an exploration system that was implemented by the search-advertising team of a prominent web-portal to address the cold ads problem. The cold ads problem refers to the situation where, when new ads are injected into the system by advertisers, the system is unable to assign an accurate quality to the ad (in our case, the click probability). As a consequence, the advertiser may suffer from low impression volumes for these cold ads, and the overall system may perform sub-optimally if the click probabilities for new ads are not learnt rapidly. We designed a new exploration system that was adapted to search advertising and the serving constraints of the system. In this paper, we define the problem, discuss the design details of the exploration system, new evaluation criteria, and present the performance metrics that were observed by us.

http://dl.acm.org/authorize?N33342

170. A Quasi-experimental Estimate of the Impact of P2P Transportation Platforms on Urban Consumer Patterns

Zhe Zhang (Carnegie Mellon University);Beibei Li (Carnegie Mellon University)

With the pervasiveness of mobile technology and location-based computing, new forms of smart urban transportation, such as Uber &amp; Lyft, peer-to-peer new forms of urban infrastructure can influence individuals' movement frictions and patterns, in turn influencing local consumption patterns and the economic performance of local businesses. To gain insights about future impact of urban transportation changes, in this paper, we take advantage of a novel and individually-detailed dataset and use econometric and casual analysis methodsto examine how such peer-to-peer car sharing services may affect consumer mobility and consumption patterns.

http://dl.acm.org/authorize?N33458

171. Planning Bike Paths based on Sharing-Bikes' Trajectories

Jie Bao (Microsoft Research);Tianfu He (Harbin Institution of Technology);Sijie Ruan (Xidian University);Yanhua Li (Worcester Polytechnic Institute (WPI));Yu Zheng (Microsoft Research)

Cycling as a green transportation mode has been promoted by many governments all over the world. As a result, constructing effective bike lanes has become a crucial task for governments promoting the cycling life style, as well-planned bike paths can reduce traffic congestion and decrease safety risks for both cyclists and motor vehicle drivers. Unfortunately, existing trajectory mining approaches for bike lane planning do not consider key realistic government constraints: 1) budget imitations, 2) construction convenience, and 3) bike lane utilization.

In this paper, we propose a data-driven approach to develop bike lane construction plans based on large-scale real world bike trajectory data. We enforce these constraints to formulate our problem and introduce a flexible objective function to tune the benefit between coverage of the number of users and the length of their trajectories. We prove the NP-hardness of the problem and propose greedy-based heuristics to address it. Finally, we deploy our system on Microsoft Azure, providing extensive experiments and case studies to demonstrate the effectiveness of our approach.

http:/dl.acm.org/authorize?N33327

172. Automated Categorization of Onion Sites for Analyzing the Darkweb Ecosystem

Shalini Ghosh (SRI);Ariyam Das (UCLA);Phillip Porras (SRI);Vinod Yegneswaran (SRI International);Ashish Gehani (SRI International)

Onion sites on the darkweb operate using the Tor Hidden Service (HS) protocol to shield their locations on the Internet, which (among other features) enables these sites to host malicious and illegal content while being resistant to legal action and seizure. Identifying and monitoring such illicit sites in the darkweb is of high relevance to the Computer Security and Law Enforcement communities. We have developed an automated infrastructure that crawls and indexes content from onion sites into a large-scale data repository, called LIGHTS, with over 100M pages. In this paper we describe Automated Tool for Onion Labeling (ATOL), a novel scalable analysis service developed to conduct a thematic assessment of the content of onion sites in the LIGHTS repository. ATOL has three core components — (a) a novel keyword discovery mechanism (ATOLKeyword) which extends analyst-provided keywords for different categories by suggesting new descriptive and discriminative keywords that are relevant for the categories; (b) a classification framework (ATOLClassify) that uses the discovered keywords to map onion site content to a set of categories when sufficient labeled data is available; (c) a clustering framework (ATOLCluster) that can leverage information from multiple external heterogeneous knowledge sources, ranging from domain expertise to Bitcoin transaction data, to categorize onion content in the absence of sufficient supervised data. e paper presents empirical results of ATOL on onion datasets derived from the LIGHTS repository, and additionally benchmarks ATOL's algorithms on the publicly available 20 Newsgroups dataset to demonstrate the reproducibility of its results. On the LIGHTS dataset, ATOLClassify gives a 12

http://dl.acm.org/authorize?N33469

173. Matching Restaurant Menus to Crowdsourced Food Data, A Scalable Machine Learning Approach

Hesam Salehian (Under Armour);Chul Lee (Under Armour);Patrick Howell (Under Armour)

We study the problem of how to match a formally structured restaurant menu item to a large database of less structured food items that has been collected via crowd-sourcing. At first glance, this problem scenario looks like a typical text matching problem that might possibly be solved with existing text similarity learning approaches. However, due to the unique nature of our scenario and the need for scalability, our problem imposes certain restrictions on possible machine learning approaches that we can employ. We propose a novel, practical, and scalable machine learning solution architecture, consisting of two major steps. First we use a query generation approach, based on a Markov Decision Process algorithm, to reduce the time complexity of searching for matching candidates. That is then followed by a re-ranking step, using deep learning techniques, to meet our required matching quality goals. It is important to note that our proposed solution architecture has already been deployed in a real application system serving tens of millions of users, and shows great potential for practical cases of user-entered text to structured text matching, especially when scalability is crucial.

http://dl.acm.org/authorize?N33480

174. Dispatch with Confidence: Integration of machine learning, Optimization and Simulation for Open Pit Mines

Kosta Ristovski (Hitachi America Ltd.);Chetan Gupta (Hitachi America Ltd.);Kunihiko Harada (Hitachi America Ltd.);Hsiu-Khuern Tang (Hitachi America Ltd.)

Open pit mining operations require utilization of extremely expensive equipment such as large trucks, shovels and loaders. To remain competitive, mining companies are under pressure to increase equipment utilization and reduce operational costs. The key to this in mining operations is to have sophisticated truck assignment strategies which will ensure that equipment is utilized efficiently with minimum operating cost. To address this problem, we have implemented truck assignment approach which integrates machine learning, linear/integer programming and simulation. Our truck assignment approach takes into consideration the number of trucks and their sizes, shovels and dump locations as well as stochastic activity times during the operations. Machine learning is used to predict probability distributions of equipment activity duration. We have validated of the approach using data collected from two open pit mines. Our experimental results shows that our approach offers increase of 10

http://dl.acm.org/authorize?N33488

175. Learning to Generate Rock Descriptions from Multivariate Well Logs with Hierarchical Attention

Bin Tong (Research and Development Group, Hitachi, Ltd.);Martin Klinkigt (Research and Development Group, Hitachi, Ltd.);Makoto Iwayama (Research and Development Group, Hitachi, Ltd.);Toshihiko Yanase (Research and Development Group, Hitachi, Ltd.);Yoshiyuki Kobayashi (Research and Development Group, Hitachi, Ltd.);Anshuman Sahu (Big Data Laboratory, Hitachi America, Ltd.);Ravigopal Vennelakanti (Big Data Laboratory, Hitachi America, Ltd.)

In the shale oil and gas industry, operators are looking toward big data analytics to optimize operations and reduce cost. In this paper, we mainly focus on how to assist operators in understanding the subsurface formation, thereby helping them make optimal decisions. A large number of geology reports and well logs describing the sub-surface have been accumulated over years. Issuing geology reports is more time consuming and depends more on the expertise of engineers than acquiring the well logs. To assist in issuing geology reports, we propose an encoder-decoder-based model to automatically generate rock descriptions in human-readable format from multivariate well logs. Due to the different formats of data, this task differs dramatically from image and video captioning. The challenges are how to model structured rock descriptions and leverage the information

in multivariate well logs. To achieve this, we design a hierarchical structure and two forms of attention for the decoder. Extensive validations are conducted on public well data of North Dakota in the United States. We show that our model is effective in generating rock descriptions. These forms of attention enable the provision of a better insight into relations between well-log types and rock properties with our model from a data-driven perspective. This research is expected to be integrated into a customized solution for Hitachi regarding shale oil and gas.

http://dl.acm.org/authorize?N33483

176. Stock Price Prediction via Discovering Multi-Frequency Trading Patterns

Liheng Zhang (University of Central Florida);Charu Aggarwal (IBM T. J. Watson Research Center);Guo-Jun Qi (University of Central Florida)

Stock prices are formed based on short and/or long-term commercial and trading activities that reflect different frequencies of trading patterns. However, these patterns are often elusive as they are affected by many uncertain political-economic factors in the real world, such as corporate performances, government policies, and even breaking news circulated across markets. Moreover, time series of stock prices are non-stationary and non-linear, making the prediction of future price trends much challenging. To address them, we propose a novel State Frequency Memory (SFM) recurrent network to capture the multi-frequency trading patterns from past market data to make long and short term predictions over time. Inspired by Discrete Fourier Transform (DFT), the SFM decomposes the hidden states of memory cells into multiple frequency components, each of which models a particular frequency of latent trading pattern underlying the fluctuation of stock price. Then the future stock prices are predicted as a nonlinear mapping of the combination of these components in an Inverse Fourier Transform (IFT) fashion. Modeling multi-frequency trading patterns can enable more accurate predictions for various time ranges: while a short-term prediction usually depends on high frequency trading patterns, a long-term prediction should focus more on the low frequency trading patterns targeting at long-term return. Unfortunately, no existing model explicitly distinguishes between various frequencies of trading patterns to make dynamic predictions in literature. The experiments on the real market data also demonstrate more competitive performance by the SFM as compared with the state-of-the-art methods.

http://dl.acm.org/authorize?N33494

177. “The Leicester City Fairytale?”: Utilizing New Soccer Analytics Tools to Compare Performance in the 15/16 & 16/17 EPL Seasons

Héctor Ruiz (STATS);Paul Power (STATS);Xinyu Wei (STATS);Patrick Lucey (STATS)

The last two years have been somewhat of a rollercoaster for English Premier League (EPL) team Leicester City. In the 2015/16 season, against all odds and logic, they won the league to much fanfare. Fast-forward nine months later, and they are battling relegation. What could describe this fluctuating form? As soccer is a very complex and strategic game, common statistics (e.g., passes, shots, possession) do not really tell the full story on how a team succeeds and fails. However, using machine learning tools and a plethora of data, it is now possible to obtain some insights into how a team performs. To showcase the utility of these new tools (i.e., expected goal value, expected save value, strategy-plots and passing quality measures), we first analyze the EPL 2015/16 season which a specific emphasis on the champions Leicester City, and then compare it to the current one. Finally, we show how these features can be used to predict future performance.

http://dl.acm.org/authorize?N33489

178. An efficient bandit algorithm for realtime multivariate optimization

Daniel Hill (Amazon.com);Houssam Nassif (Amazon.com);Yi Liu (Amazon.com);Anand Iyer (Amazon.com);S. V. N. Vishwanathan (vishy@amazon.com)

Optimization is commonly employed to determine the content of web pages, such as to maximize conversions on landing pages or click-through rates on search engine result pages. Often the layout of these pages can be decoupled into several separate decisions. For example, the composition of a landing page may involve deciding which image to show, which wording to use, what color background to display, etc. Thus, optimization is a combinatorial problem over an exponentially large decision space. Randomized experiments do not scale well to this setting, and therefore, in practice, one is typically limited to optimizing a single aspect of a web page at a time. This represents a missed opportunity in both the speed of experimentation and the exploitation of possible interactions between layout decisions.

Here we focus on multivariate optimization of interactive web pages. We formulate an approach where the possible interactions between different components of the page are modeled explicitly. We apply bandit methodology to explore the layout space efficiently and use hill-climbing to select optimal content in realtime. Our algorithm also extends to contextualization and personalization of layout selection. Simulation results show the suitability of our approach to large decision spaces with strong interactions between content. We further apply our algorithm to optimize a message that promotes adoption of an Amazon service. After only a single week of online optimization, we saw a 21

http://dl.acm.org/authorize?N33461

179. Machine Learning for Encrypted Malware Traffic Classification: Accounting for Noisy Labels and Non-Stationarity

Blake Anderson (Cisco Systems, Inc.);David McGrew (Cisco Systems, Inc.)

The application of machine learning for the detection of malicious network traffic has been well researched over the past several decades; it is particularly appealing when the traffic is encrypted because traditional pattern-matching approaches cannot be used. Unfortunately, the promise of machine learning has been slow to materialize in the network security domain. In this paper, we highlight two primary reasons why this is the case: inaccurate ground truth and a highly non-stationary data distribution. To demonstrate and understand the effect that these pitfalls have on popular machine learning algorithms, we design and carry out experiments that show how six common algorithms perform when confronted with real network data.

With our experimental results, we identify the situations in which certain classes of algorithms underperform on the task of encrypted malware traffic classification. We offer concrete recommendations for practitioners given the real-world constraints outlined. From an algorithmic perspective, we find that the random forest ensemble method outperformed competing methods. More importantly, feature engineering was decisive; we found that iterating on the initial feature set, and including features suggested by domain experts, had a much greater impact on the performance of the classification system. For example, linear regression using the more expressive feature set easily outperformed the random forest method using a standard network traffic representation on all criteria considered. Our analysis is based on millions of TLS encrypted sessions collected over 12 months from a commercial malware sandbox and two geographically distinct, large enterprise networks.

http://dl.acm.org/authorize?N33452

180. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks

Fenglong Ma (SUNY Buffalo);Radha Chitta (Conduent);Jing Zhou (Conduent);Quanzeng You (University of Rochester);Tong Sun (United Technologies Research Center);Jing Gao (SUNY Buffalo)

Predicting the future health information of patients from the historical Electronic Health Records (EHR) is a core research task in the development of personalized healthcare. Patient EHR data consist of sequences of visits over time, where each visit contains multiple medical codes, including diagnosis, medication, and procedure codes. The most important challenges for this task are

to model the temporality and high dimensionality of sequential EHR data and to interpret the prediction results. Existing work solves this problem by employing recurrent neural networks (RNNs) to model EHR data and utilizing simple attention mechanism to interpret the results. However, RNN-based approaches suffer from the problem that the performance of RNN drops when the size of sequences is large, and the relationships between subsequent visits are ignored by current RNN-based approaches. To address these issues, we propose Dipole, an end-to-end, simple and robust model for predicting patients' future health information. Dipole employs bidirectional recurrent neural networks to remember all the information of both the past visits and the future visits, and it introduces three attention mechanisms to measure the relationships of different visits for the prediction. With the attention mechanisms, Dipole can interpret the prediction results efficiently. Dipole also allows us to interpret the learned medical code representations, which are confirmed positively by medical experts. Experimental results on two real world EHR datasets show that the proposed Dipole can significantly improve the prediction accuracy compared with the state-of-the-art diagnosis prediction approaches and provide clinically meaningful interpretation.

http://dl.acm.org/authorize?N33470

181. A Data-driven Process Recommender Framework

Sen Yang (Rutgers University);Xin Dong (Rutgers University);Leilei Sun (Dalian University of Technology);Yichen Zhou (Rutgers University);Richard A. Farneth (Children's National Medical Center);Hui Xiong (Rutgers University);Randall S. Burd (Children's National Medical Center);Ivan Marsic (Rutgers University)

We present an approach for improving the performance of complex knowledge-based processes by providing data-driven step-by-step recommendations. Our framework uses the associations between similar historic process performances and contextual information to determine the established procedures. We introduce a novel similarity metric for grouping traces into clusters that incorporates temporal information about activity performance and handles concurrent activities. Our data-driven recommender system selects the appropriate prototype performance of the process based on user-provided context attributes. Our approach for determining the prototypes discovers the commonly performed activities and their temporal relationships. We tested our system on data from three real-world medical processes and achieved recommendation accuracy up to 0.77 F1 score (compared to 0.37 F1 score using ZeroR) and 63.2

http://dl.acm.org/authorize?N33491

182. The Fake vs Real Goods Problem: Microscopy and Machine Learning to the Rescue

Ashlesh Sharma (Entrupy Inc);Vidyuth Srinivasan (Entrupy Inc);Vishal Kanchan (Entrupy Inc);Lakshminarayanan Subramanian (Entrupy Inc)

Counterfeiting of physical goods is a global problem amount-ing to nearly 7

http://dl.acm.org/authorize?N33481

183. Large scale sentiment learning with limited labels

Vasileios Iosifidis (Leibniz University of Hanover);Eirini Ntoutsi (Leibniz University of Hanover)

Sentiment analysis is an important task in order to gain insights over the huge amounts of opinions that are generated in the social media on a daily basis. Although there is a lot of work on sentiment analysis, there are no many datasets available which one can use for developing new methods and for evaluation. To the best of our knowledge, the largest dataset for sentiment analysis is TSentiment, a 1.6 millions machine-annotated tweets dataset covering a period of about 3 months in 2009. This dataset however is too short and therefore insufficient to study heterogeneous, fast evolving streams. Therefore, we annotated the Twitter dataset of 2015 (275 million tweets) and we make it publicly available for research. For the annotation we leveraged the power of unlabeled data,

together with labeled data which we derived using emoticons and emoticon-lexicons, using semi-supervised learning and in particular, Self-Learning and Co-Training. Our main contribution is the provision of the TSentiment15 dataset together with insights from the analysis, which includes both batch-and stream-processing of the data. In the former, all labeled and unlabeled data are available to the algorithms from the beginning, whereas in the later, they are revealed gradually based on their arrival time in the stream.

http://dl.acm.org/authorize?N33462

184. Learning Temporal State of Diabetes Patients via Combining Behavioral and Demographic Data

Houping Xiao (SUNY Buffalo);Jing Gao (SUNY Buffalo);Long Vu (IBM TJ Watson);Deepak Turaga (IBM TJ Watson)

In recent decades, diabetes has become a serious disease affecting a large number of people. Although there is no cure for diabetes, it can be managed. Especially, with the advance in sensor technology, lots of data may lead to the improvement of patient diabetes management, if properly mined. However, there usually exists noise or errors in the observed behavioral data which poses challenges for extracting meaningful knowledge. To overcome this challenge, we propose to learn the latent state which represents the patient's condition. Such states should be inferred from the behavioral data but unknown a priori. In this paper, we propose a novel framework to capture the trajectory of latent states for patients from behavioral data while exploiting their demographic difference and similarities to other patients. We conduct hypothesis testing to illustrate the importance of the demographic data in diabetes management, and validate that each behavioral feature follows an exponential or a Gaussian distribution. Integrating these aspects, we propose a restricted hidden Markov model (RHMM) to estimate the trajectory of latent states by integrating the demographic and behavioral data. In RHMM, the latent state is mainly determined by the previous state and the demographic features in a nonlinear way. Markov Chain Monte Carlo techniques are used for model parameter estimation. Experimental results on synthetic and real datasets demonstrate that the proposed RHMM is effective in diabetes management.

http://dl.acm.org/authorize?N33498

185. Predicting Optimal Facility Location Without Customer Locations

Emre Yilmaz (Bilkent University);Sanem Elbasi (Bilkent University);Hakan Ferhatosmanoglu (Bilkent University)

Deriving meaningful insight from location data helps businesses make better decisions. One critical decision made by a business is choosing a location for its new facility. Optimal location queries ask for a location to build a new facility that optimizes an objective function. All of the existing works on optimal location queries propose solutions to return best location when the set of existing facilities and the set of customers are given. However, most businesses do not know the locations of their customers. In this paper, we introduce a new problem setting for optimal location queries by removing the assumption that the customer locations are known. We propose an optimal location predictor which accepts partial information about customer locations and returns a location for the new facility. The predictor generates customer locations by using given partial information and it runs optimal location queries with generated location data. Extensive experiments with real data show that the predictor can find the optimal location when sufficient information is provided.

http://dl.acm.org/authorize?N33492

186. Supporting Employer Name Normalization at both Entity and Cluster Level

Qiaoling Liu (CareerBuilder LLC);Faizan Javed (Careerbuilder);Vachik Dave (Indiana University - Purdue University, Indianapolis);Ankita Joshi (University of Georgia)

In the recruitment domain, the employer name normalization task, which links employer names in job postings or resumes to entities in an employer knowledge base (KB), is important to many business applications. In previous work, we proposed the CompanyDepot system, which used machine learning techniques to address the problem. After applying it to several applications at CareerBuilder, we faced several new challenges: 1) how to avoid duplicate normalization results when the KB is noisy and contains many duplicate entities; 2) how to address the vocabulary gap between query names and entity names in the KB; and 3) how to use the context available in jobs and resumes to improve normalization quality. To address these challenges, in this paper we extend the previous CompanyDepot system to normalize employer names not only *at entity level*, but also *at cluster level* by mapping a query to a cluster in the KB that best matches the query. We also propose a new metric called *clustering risk* for evaluating the cluster-level normalization. Moreover, we perform query expansion based on five data sources to address the vocabulary gap challenge and leverage the url context for the employer names in many jobs and resumes to improve normalization quality. We show that the proposed CompanyDepot-V2 system outperforms the previous CompanyDepot system and several other baseline systems over multiple real-world datasets. We also demonstrate the large improvement on normalization quality from entity-level to cluster-level normalization.

http://dl.acm.org/authorize?N33478

187. Increasing Yield by Dropping Constraints in Ad Serving

Brendan Kitts (Lucid Commerce)

A persistent challenge with web-based ad-servers is that the requested advertiser requirements for targeting criteria and delivery may be infeasible. In order to account for this, we pivot the problem to one of error minimization. The system is deployed in one of the largest ad-servers in the United States, where we present results from test ads as well as live advertisers.

http://dl.acm.org/authorize?N33475

188. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster

Naeemul Hassan (University of Mississippi);Fatma Arslan (University of Texas at Arlington);Chengkai Li (University of Texas at Arlington);Mark Tremayne (University of Texas at Arlington)

In this paper, we describe the current state-of-the-art of fact-checking research and describe the approach we have taken with ClaimBuster. We create a novel, human-labeled dataset of check-worthy factual claims using the sentences of the U.S. presidential election general debate transcripts and use natural language processing and supervised learning techniques to develop a factual claim identification model which is one of the core components of the presented fact-checking platform, ClaimBuster. We describe various components of the ClaimBuster system architecture and outline our development plan. We showcase how ClaimBuster is used to live cover the 2016 U.S. presidential election debates and monitor social media platforms and Hansard for identifying check-worthy factual claims. The performance of ClaimBuster is compared with the professional journalists and fact-checking organizations.

http://dl.acm.org/authorize?N33460

189. Discovering Fine Grained Pollution Sources and Propagation Patterns in Urban Area

Yun Cheng (Air Scientific);Xiucheng Li (Nanyang Technological University);Gao Cong (Nanyang Technological University);Lisi Chen (Department of Computer Science, Hong Kong Baptist University)

Air quality is one of the most important environmental concerns in the world, and it has deteriorated substantially over the past years in many countries. For example, Chinese Academy of Social Sciences reports that the problem of haze and fog in China is hitting a record level, and China

is currently suffering from the worst air pollution. Among the various causalities of air quality, particulate matter with a diameter of 2.5 micrometers or less (i.e., PM2.5) is a very important factor; governments and people are increasingly concerned with the concentration of PM2.5. In many cities, stations for monitoring PM2.5 concentration have been built by governments or companies to monitor urban air quality. Apart from monitoring, there is a rising demand for finding pollution sources of PM2.5 and discovering the transmit of PM2.5 based on the data of PM2.5 monitoring stations. However, to the best of our knowledge, none of previous work proposes a solution to the problem of detecting pollution sources and mining pollution propagation patterns from such monitoring data. In this work, we propose the first solution for the problem, which comprises two steps. The first step is to extracting the uptrend intervals and calculating the causal strengths among spatially distributed sensors; The second step is to construct causality graphs and perform the frequent subgraphs mining on these causality graphs to find the pollution sources and propagation patterns. We use real-life monitoring data collected by a company in our experiments. Our experimental results demonstrate significant findings regarding the pollutant source and pollutant propagations in Beijing, which will be useful for government to make policy and govern pollution sources.

http://dl.acm.org/authorize?N33476

190. Real-Time Optimization Of Web Publisher RTB Revenues

Pedro Chahuara (XRCE);Jean-Michel Renders (Xerox Research Centre Europe);Nicolas Grislain (AlephD);Gregoire Jauvion (AlephD)

This paper describes an engine to optimize web publisher revenues from second-price auctions. These auctions are widely used to sell online ad spaces in a mechanism called real-time bidding (RTB). Optimization within these auctions is crucial for web publishers, because setting appropriate reserve prices can significantly increase revenue. We consider a practical real-world setting where the only available information before an auction occurs consists of a user identifier and an ad placement identifier. The real-world challenges we had to tackle consist mainly of tracking the dependencies on both the user and placement in an highly non-stationary environment and of dealing with censored bid observations. These challenges led us to make the following design choices: (i) we adopted a relatively simple non-parametric regression model of auction revenue based on an incremental time-weighted matrix factorization which implicitly builds adaptive users' and placements' profiles; (ii) we jointly used a non-parametric model to estimate the first and second bids' distribution when they are censored, based on an on-line extension of the Aalen's Additive model. Our engine is a component of a deployed system handling hundreds of web publishers across the world, serving billions of ads a day to hundreds of millions of visitors. The engine is able to predict, for each auction, an optimal reserve price in approximately one millisecond and yields a significant revenue increase for the web publishers.

http://dl.acm.org/authorize?N33454

191. A Practical Algorithm for Solving the Incoherence Problem of Topic Models In Industrial Applications

Amr Ahmed (google);James Long (google);Dan Silva (google);Yuan Wang (google)

Topic models are often applied in industrial settings to discover user profiles from activity logs where documents corresponds to users and words to complex objects such as web sites and installed apps. Standard topic models ignore the content-based similarity structure between these objects largely because of the inability of the Dirichlet prior to capture such side information of word-word correlation. Several approaches were proposed to replace the Dirichlet prior with more expressive alternatives. However, this added expressivity comes with a heavy premium: inference becomes intractable and sparsity is lost which renders these alternatives not suitable for industrial scale applications. In this paper, we take a radically different approach to incorporating word-word correlation in topic models by applying this side information at the posterior level rather than at the prior level. We show that this choice preserves sparsity and results in a graph-based sampler

for LDA whose computational complexity is asymptotically on bar with state of the art Alias base samplers for LDA [6]. We illustrate the efficacy of our approach over real industrial datasets that span up to billions of users, tens of millions of words and thousands of topic. To the best of our knowledge, our approach provides the first practical and scalable solution to this important problem.

http://dl.acm.org/authorize?N33451

192. Collecting and Analyzing Millions of mHealth Data Streams

Thomas Quisel (Evidation Health);Luca Foschini (Evidation Health);Alessio Signorini (Evidation Health);David Kale (USC Information Sciences Institute)

Players across the health ecosystem are initiating studies of thousands, even millions, of participants to gather diverse types of data, including biomedical, behavioral, and lifestyle in order to advance medical research. These efforts to collect multi-modal data sets on large cohorts coincide with the rise of broad activity and behavior tracking across industries, particularly in healthcare and the growing field of mobile health (mHealth). Government and pharmaceutical sponsored, as well as patient-driven group studies in this arena leverage the ability of mobile technology to continuously track behaviors and environmental factors with minimal participant burden. However, the adoption of mHealth has been constrained by the lack of robust solutions for large-scale data collection in free-living conditions and concerns around data quality.

In this work, we describe the infrastructure Evidation Health has developed to collect mHealth data from millions of users through hundreds of different mobile devices and apps. Additionally, we provide evidence of the utility of the data for inferring individual traits pertaining to health, wellness, and behavior. To this end, we introduce and evaluate deep neural network models that achieve high prediction performance without requiring any feature engineering when trained directly on the densely sampled multivariate mHealth time series data.

We believe that the present work substantiates both the feasibility and the utility of creating a very large mHealth research cohort, as envisioned by the many large cohort studies currently underway across therapeutic areas and conditions.

http://dl.acm.org/authorize?N33487

193. Dynamic Attention Deep Model for Article Recommendation by Learning Human Editors Demonstration

Xuejian Wang (Shanghai Jiao Tong University);Lantao Yu (Shanghai Jiao Tong University);Kan Ren (Shanghai Jiao Tong University);Guanyu Tao (ULU Technologies Inc.);Weinan Zhang (Shanghai Jiao Tong University);Jun Wang (University College London);Yong Yu (Shanghai Jiao Tong University)

As aggregators, online news portals face great challenges in continuously selecting a pool of candidate articles to be shown to their users. Typically, those candidate articles are recommended manually by platform editors from a much larger pool of articles aggregated or submitted from multiple sources. Such a hand-pick process is labor intensive and time-consuming. In this paper, we study the editor article selection behavior and propose a learning by demonstration system to automatically select a subset of articles from the large pool. Our data analysis shows that (i) editors' selection criteria are non-explicit, which are less based only on the keywords or topics, but more depend on the quality and attractiveness of the writing from the candidate article, which is hard to capture based on traditional bag-of-words article representation. And (ii) editors' article selection behaviors are dynamic: articles with different data distribution come into the pool everyday and the editors' preference varies, which are driven by some underlying periodic or occasional patterns. To address such problems, we propose a meta-attention model across multiple deep neural nets to (i) automatically catch the editors' underlying selection criteria via the automatic representation learning of each article and its interaction with the meta data and (ii) adaptively capture the change of such criteria via a hybrid attention model. The attention model strategically incorporates multiple prediction models, which are trained in previous days. The system has been deployed in a commercial article feed platform. A 9-day A/B testing has demonstrated the consistent superiority of our proposed model over several strong baselines.

http://dl.acm.org/authorize?N33495

194. A Taxi Order Dispatch Model based On Combinatorial Optimization

Lingyu Zhang (Didi Chuxing);Tao Hu (Didi Chuxing);Yue Min (Didi Chuxing);Guobin Wu (Didi Chuxing);Junying Zhang (Didi Chuxing);Pengcheng Feng (Didi Chuxing);Pinghua Gong (Didi Chuxing);Jieping Ye (Didi Chuxing)

Taxi-booking apps have been very popular all over the world as they provide fast response time and convenience to the users. The key component of a taxi-booking app is the dispatch system which aims to provide optimal matches between drivers and riders. Traditional dispatch systems sequentially dispatch taxis to riders and aim to maximize the driver acceptance rate for each individual order. However, the traditional systems cannot guarantee the global success rate, which degrades the rider experience when using the app. In this paper, we propose a novel system that attempts to optimally dispatch taxis to serve multiple bookings. The proposed system aims to maximize the global success rate, and thus it optimizes the overall traffic efficiency, leading to enhanced user experience. To further enhance users' experience, we also propose a method to predict destinations of a user once the taxi-booking APP is started. The proposed method employs the Bayesian framework to model the distribution of a user's destination based on his/her travel histories. We use A/B tests to compare our new taxi dispatch method with state-of-the-art models using data collected in Beijing. Experimental results show that the proposed method is significantly better than other state-of-the art models in terms of global success rate (increased from 80

http://dl.acm.org/authorize?N33405

195. AESOP: Automatic Policy Learning for Predicting and Mitigating Network Service Impairments

Supratim Deb (AT&T Labs);Zihui Ge (AT&T Labs);Sastry Isukapalli (AT&T Labs);Sarat Puthenpura (AT&T Labs);Shobha Venkataraman (AT&T Labs);He Yan (AT&T Labs);Jennifer Yates (AT&T Labs)

Efficient management and control of modern (4G LTE) and next-gen (5G) cellular networks is of paramount importance as networks have to maintain highly reliable service quality to support rapid growth in traffic demand and new application services. Fast responses to network service degradation is going to be the key for networks to deliver promised benefits to end-users. Evolving network management towards data-driven automation would have dramatic impact on the quality and speed of mitigation. In this paper, we present AESOP, a data-driven intelligent system to facilitate automatic learning of policies and rules for triggering remedial troubleshooting actions in Networks. AESOP combines best practices of operations' intelligence with variety of measurement data to learn and validate operational policies to mitigate service problems in networks. AESOP design addresses key challenges like learning from high-dimensional noisy data, capturing multiple fault models, very high service-cost of false positives, evolving network infrastructure (leading to changing data distribution), etc. We present the design of our system and show results from our ongoing experiments to show the operational efficiency that can be had from such policy driven automation.

http://dl.acm.org/authorize?N33468

196. Finding Precursors to Anomalous Drop in Airspeed During a Flight's Take-off

Vijay Manikandan Janakiraman (USRA/ NASA AMES Research Center);Bryan Matthews (SGT/ NASA AMES Research Center);Nikunj Oza (NASA AMES Research Center)

Aerodynamic stall based loss of control in flight is a major cause of fatal flight accidents. In a typical take-off, a flight's airspeed continues to increase as it gains altitude. However, in some cases, the airspeed may drop immediately after take-off and when left uncorrected, the flight gets close to a stall condition which is extremely risky. The take-off is a high workload period for the flight crew involving frequent monitoring, control and communication with the ground control tower. Although there exists secondary safety systems and specialized recovery maneuvers, current technology is reactive; often based on simple threshold detection and does not provide the crew with sufficient lead time. Further, with increasing complexity of automation, the crew may not be aware of the true states of the automation to take corrective actions in time. At NASA, we aim to develop decision support tools by mining historic flight data to proactively identify and manage high risk situations encountered in flight. In this paper, we present our work on finding precursors to the anomalous drop-in-airspeed (ADA) event using the ADOPT (Automatic Discovery of Precursors in Time series) algorithm. ADOPT works by converting the precursor discovery problem into a search for sub-optimal decision making in the time series data, which is modeled using reinforcement learning. We give insights about the flight data, feature selection, ADOPT modeling and results on precursor discovery. Some improvements to ADOPT algorithm are implemented that reduces its computational complexity and enables forecasting of the adverse event. Using ADOPT analysis, we have identified some interesting precursor patterns that were validated to be operationally significant by subject matter experts. The performance of ADOPT is evaluated by using the precursor scores as features to perform classification of the time series data.

http://dl.acm.org/authorize?N33464

197. Customer Life Time Value (CLTV) Prediction Using Embeddings

Ben Chamberlain (Imperial College London);Angelo Cardoso (ASOS);Bryan Liu (ASOS);Marc Deisenroth (Imperial College London);Roberto Paglieri (ASOS)

We describe the Customer Life Time Value (CLTV) prediction system deployed at ASOS.com, a global online fashion retailer. CLTV prediction is an important problem in e-commerce where an accurate estimate of future value allows retailers to effectively allocate marketing spend, identify and nurture high value customers and mitigate exposure to losses. The system at ASOS provides daily estimates of the future value of every customer and is one of the cornerstones of the personalised shopping experience. The state of the art in this domain uses large numbers of handcrafted features and ensemble regressors to forecast value, predict churn and evaluate customer loyalty. We describe our system, which adopts this approach, and our ongoing efforts to further improve it. Recently, domains including language, vision and speech have shown dramatic advances by replacing hand-crafted features with features that are learned automatically from data. We show that learning feature representations is a promising extension to the state of the art in CLTV modeling. We propose a novel way to generate embeddings of customers, which addresses the issue of the ever changing product catalogue and obtain a significant improvement over an exhaustive set of handcrafted features.

http://dl.acm.org/authorize?N33465

198. Visual Search at eBay

Fan Yang (eBay Inc.);Ajinkya Kale (eBay Inc.);Yury Bubnov (eBay Inc.);Leon Stein (eBay Inc.);Qiaosong Wang (eBay Inc.);Hadi Kiapour (eBay Inc.);Robinson Piramuthu (eBay Inc.)

In this paper, we propose a novel end-to-end approach for scalable visual search infrastructure. We discuss the challenges we faced for a massive volatile inventory like at eBay and present our solution to overcome those. We harness the availability of large image collection of eBay listings and state-of-the-art deep learning techniques to perform visual search at scale. Supervised approach for optimized search limited to top predicted categories and also for compact binary signature are key to scale up without compromising accuracy and precision. Both use a common deep neural network requiring only a single forward inference. The system architecture is presented with in-depth discussions of its basic components and optimizations for a trade-off between search relevance and latency. This solution is currently deployed in a distributed cloud infrastructure and fuels visual search in eBay ShopBot. We show benchmark on ImageNet dataset on which our approach is faster and more accurate than several unsupervised baselines. We share our learnings with the hope that visual search becomes a first class citizen for all large scale search engines rather than an afterthought.

http://dl.acm.org/authorize?N33490

199. Learning Tree-Structured Detection Cascades for Heterogeneous Networks of Embedded Devices

Hamid Dadkhahi (UMass Amherst);Benjamin Marlin (UMass Amherst)

In this paper, we present a new approach to learning cascaded classifiers for use in computing environments that involve networks of heterogeneous and resource-constrained, low-power embedded compute and sensing nodes. We present a generalization of the classical linear detection cascade to the case of tree-structured cascades where different branches of the tree execute on different physical compute nodes in the network. Different nodes have access to different features, as well as access to potentially different computation and energy resources. We concentrate on the problem of jointly learning the parameters for all of the classifiers in the cascade given a fixed cascade architecture and a known set of costs required to carry out the computation at each node.To accomplish the objective of joint learning of all detectors, we propose a novel approach to combining classifier outputs during training that better matches the hard cascade setting in which the learned system will be deployed. This work is motivated by research in the area of mobile health where energy efficient real time detectors integrating information from multiple wireless on-body sensors and a smart phone are needed for real-time monitoring and delivering just-in-time adaptive interventions. We apply our framework to two activity recognition datasets as well as the problem of cigarette smoking detection from a combination of wrist-worn actigraphy data and respiration chest band data.

http://dl.acm.org/authorize?N33467

200. DeepProbe: Information Directed Sequence Understanding and Chatbot Design via Recurrent Neural Networks

Zi Yin (Stanford University);Keng-Hao Chang (Microsoft);Ruofei Zhang (Microsoft)

Information extraction and user intention identification is a central topic in modern query understanding and recommendation systems. In this paper, we propose DeepProbe, a generic information-directed interaction framework which is built around an attention-based sequence to sequence (seq2seq) recurrent neural network. DeepProbe can rephrase, evaluate, and even actively ask questions, leveraging the generative ability and likelihood estimation made possible by seq2seq models. DeepProbe makes decisions based on a derived uncertainty (entropy) measure conditioned on user inputs, possibly with multiple rounds of interactions. Three applications, namely a rewriter, a relevance scorer and a chatbot for ad recommendation, were built around DeepProbe, with the first two serving as precursory building blocks for the third. We first use the seq2seq model in DeepProbe to rewrite a user query into one of standard query form, which is submitted to an ordinary recommendation system. Secondly, we evaluate the returned results by DeepProbe's seq2seq model-based relevance scoring. Finally, we build a chatbot prototype capable of making active user interactions, which can ask questions that maximize information gain, allowing for a more efficient user intention identification process. We evaluate first two applications by 1) comparing with baselines by BLEU and AUC, and 2) human judge evaluation. Both demonstrate significant improvements compared with current state-of-the-art systems, proving their values as useful tools on their own, and at the same time laying a good foundation for the ongoing chatbot application.

http://dl.acm.org/authorize?N33493

201. An Intelligent Customer Care Assistant System for Large-Scale Cellular Network Diagnosis

Lujia Pan (Noah Ark's Lab, Huawei Technologies);Jianfeng Zhang (Noah Ark's Lab, Huawei Technologies);Patrick P. C. Lee (The Chinese University of Hong Kong);Hong Cheng (The Chinese University of Hong Kong);Cheng He (Noah Ark's Lab, Huawei Technologies);Caifeng He (Noah Ark's Lab, Huawei Technologies);Keli Zhang (Noah Ark's Lab, Huawei Technologies)

With the advent of cellular network technologies, mobile Internet access becomes the norm in everyday life. In the meantime, the complaints made by subscribers about unsatisfactory cellular

network access also become increasingly frequent. From a network operator's perspective, achieving accurate and timely cellular network diagnosis about the causes of the complaints is critical for both improving subscriber-perceived experience and maintaining network robustness. We present the Intelligent Customer Care Assistant (ICCA), a distributed fault classification system that exploits a data-driven approach to perform large-scale cellular network diagnosis. ICCA takes massive network data as input, and realizes both offline model training and online feature computation to distinguish between user and network faults in real time. ICCA is currently deployed in a metropolitan LTE network in China that is serving around 50 million subscribers. We show via evaluation that ICCA achieves high classification accuracy (85.3

http://dl.acm.org/authorize?N33485

202. Train and Distribute: Managing Simplicity vs. Flexibility in High-Level Machine Learning Frameworks

Heng-Tze Cheng (Google);Lichan Hong (Google);Mustafa Ispir (Google);Clemens Mewald (Google);Zakaria Haque (Google);Illia Polosukhin (Google);Georgios Roumpos (Google);D Sculley (Google);Jamie Smith (Google);David Soergel (Google);Yuan Tang (Uptake Technologies);Philipp Tucker (Google);Martin Wicke (Google);Cassandra Xia (Google);Jianwei Xie (Google)

We present a framework for specifying, training, evaluating, and deploying machine learning models. Our focus is on simplifying cutting edge machine learning for practitioners in order to bring such technologies into production. Recognizing the fast evolution of the field of deep learning, we make no attempt to capture the design space of all possible model architectures in a domain-specific language (DSL) or similar configuration language. We allow users to write code to define their models, but provide abstractions that guide developers to write models in ways conducive to productionization. We also provide a unifying Estimator interface, making it possible to write downstream infrastructure (e.g. distributed training, hyperparameter tuning) independent of the model implementation. We balance the competing demands for flexibility and simplicity by offering APIs at different levels of abstraction, making common model architectures available "out of the box", while providing a library of utilities designed to speed up experimentation with model architectures. To make out of the box models flexible and usable across a wide range of problems, these canned Estimators are parameterized not only over traditional hyperparameters, but also using feature columns, a declarative specification describing how to interpret input data. We discuss our experience in using this framework in research and production environments, and show the impact on code health, maintainability, and development speed.

http://dl.acm.org/authorize?N33466

203. Discovering Concepts Using Large Table Corpus

Keqian Li (University of California, Santa Barbara);Yeye He (Microsoft Research);Kris Ganjam (Microsoft Research)

Existing work on knowledge discovery mostly uses natural language techniques to extract entities and relationships from textual documents. However, today relational tables are abundant in quantities, often with clean and well-structured data values. So far these rich relational tables have been largely overlooked for the purpose of knowledge discovery. In this work, we study the problem of extracting concept hierarchies given a large table corpus. Our method first iteratively groups values in a table corpus based on co-occurrence statistics to produce a candidate hierarchical tree. The tree is then summarized by selecting nodes that can best "describe" the original corpus, in order to produce a small tree with desired concept hierarchies, and is easy for humans to understand and curate. We design our algorithms based on map-reduce to scale to large table corpus. Experiment evaluation on real enterprise table corpus shows that proposed approach can generate concepts with high quality.

http://dl.acm.org/authorize?N33477

204. Contextual Spatial Outlier Detection with Metric Learning

Guanjie Zheng (College of Information Sciences and Technology, Pennsylvania State University);Susan L. Brantley (Department of Geosciences, Pennsylvania State University);Zhenhui Li (College of Information Sciences and Technology, Pennsylvania State University)

Hydraulic fracturing (or "fracking'") is a revolutionary well stimulation technique for shale gas extraction, but has spawned controversy in environmental contamination. If methane from gas wells leaks extensively, this greenhouse gas can impact drinking water wells and enhance global warming. Our work is motivated by this heated debate on environmental issue and we propose data analytical techniques to detect anomalous water samples with potential leakages. We propose a spatial outlier detection method based on contextual neighbors. Different from existing work, our approach utilizes both spatial attributes and non-spatial contextual attributes to define neighbors. We use robust metric learning to combine different contextual attributes in order to find more precise neighbors. Our technique can be generalized to any spatial dataset. The extensive experimental results on six real-world datasets demonstrate the effectiveness of our proposed approach. We also show some interesting case studies, with one case linking to a gas well leakage.

http://dl.acm.org/authorize?N33406

205. Optimized Cost per Click in Taobao Display Advertising

Han Zhu (Alibaba Inc.);Junqi Jin (Alibaba Inc.);Chang Tan (Alibaba Inc.);Fei Pan (Alibaba Inc.);Yifan Zeng (Alibaba Inc.);Han Li (Alibaba Inc.);Kun Gai (Alibaba Inc.)

Taobao, as the largest online retail platform in the world, provides billions of online display advertising impressions for millions of advertisers every day. For commercial purposes, the advertisers bid for specific spots and target crowds to compete for business traffic. The platform chooses the most suitable ads to display in tens of milliseconds. Common pricing methods include cost per mille (CPM) and cost per click (CPC). Traditional advertising systems target certain traits of users and ad placements with fixed bids, essentially regarded as coarse-grained matching of bid and traffic quality. However, the fixed bids set by the advertisers competing for different quality requests cannot fully optimize the advertisers' key requirements. Moreover, the platform has to be responsible for the business revenue and user experience. Thus, we proposed a bid optimizing strategy called optimized cost per click (OCPC) which automatically adjusts the bid to achieve finer matching of bid and traffic quality of page view (PV) request granularity. Our approach optimizes advertisers' demands, platform business revenue and user experience and as a whole improves traffic allocation efficiency. We have validated our approach in Taobao display advertising system in production. The online A/B test shows our algorithm yields substantially better results than previous fixed bid manner.

http://dl.acm.org/authorize?N33409

206. BDT: Boosting Decision Tables for High Accuracy and Scoring Efficiency

Yin Lou (Airbnb);Mikhail Obukhov (LinkedIn)

In this paper we present gradient boosted decision tables (BDTs). A d-dimensional decision table maps a sequence of d boolean tests to a real value in $\mathbb{R}$. We propose novel algorithms to fit decision tables. Our thorough empirical study suggests that decision tables are better weak learners in the gradient boosting framework and can improve the accuracy of the boosted ensemble. In addition, we develop an efficient data structure to represent decision tables and propose a novel fast algorithm to improve the scoring efficiency of boosted ensemble of decision tables. Experiments on public classification and regression datasets demonstrate that our method is able to achieve 1.5x to 6x speedups over boosted regression trees baseline. We complement our experimental evaluation with a bias-variance analysis that explains how different weak models influence the predictive power of the boosted ensemble. Our experiments suggest gradient boosting with randomly backfitted decision tables distinguishes itself as the most accurate method on a number of classification and regression problems. We have deployed a BDT model to LinkedIn news feed system and achieved significant lift on key metrics.

http://dl.acm.org/authorize?N33479

207. RUSH! Targeted Time-limited Coupons via Purchase Forecasts

Emaad Manzoor (Carnegie Mellon University);Leman Akoglu (Carnegie Mellon University)

Time-limited promotions that exploit consumers' sense of urgency to boost sales account for billions of dollars in consumer spending each year. However, it is challenging to discover the right timing and duration of a promotion to increase its chances of being redeemed. In this work, we consider the problem of delivering time-limited discount coupons, where we partner with a large national bank functioning as a commission-based third-party coupon provider. Specifically, we use large-scale anonymized transaction records to model consumer spending and forecast future purchases, based on which we generate data-driven, personalized coupons. Our proposed model RUSH! (1) predicts both the time and category of the next event; (2) captures correlations between purchases in different categories (such as shopping triggering dining purchases); (3) incorporates temporal dynamics of purchase behavior (such as increased spending on weekends); (4) is composed of additive factors that are easily interpretable; and finally (5) scales linearly to millions of transactions. We design a cost-benefit framework that facilitates systematic evaluation in terms of our application, and show that RUSH! provides higher expected value than various baselines that do not jointly model time and category information.

http://dl.acm.org/authorize?N33472

208. Formative essay feedback using predictive scoring models

Bronwyn Woods (Turnitin);David Adamson (Turnitin);Shayne Miel (Turnitin);Elijah Mayfield (Turnitin)

A major component of secondary education is learning to write effectively, a skill which is bolstered by repeated practice with formative guidance. However, providing focused feedback to every student on multiple drafts of each essay throughout the school year is a challenge for even the most dedicated of teachers. This paper describes a new ordinal essay scoring model and its state of the art performance compared to recent results in the Automated Essay Scoring field. Extending this model, we describe a method for using prediction on realistic essay variants to select sentences for targeted feedback. This method is used in Revision Assistant, a deployed data-driven educational product that provides immediate, rubric-specific, sentence-level feedback to students to supplement teacher guidance. We present initial evaluations of this feedback generation, both offline and in deployment.

http://dl.acm.org/authorize?N33497

209. Resolving the Bias in Electronic Medical Records

Kaiping Zheng (National University of Singapore);Jinyang Gao (National University of Singapore);Kee Yuan Ngiam (National University Health System);Beng Chin Ooi (National University of Singapore);Wei Luen James Yip (National University Health System)

Electronic Medical Records (EMR) are the most fundamental resources used in healthcare data analytics. Since people visit hospital more frequently when they feel sick and doctors prescribe lab examinations when they feel necessary, we argue that there could be a strong bias in EMR observations compared with the hidden conditions of patients. Directly using such EMR for analytic tasks without considering the bias may lead to misinterpretation. To this end, we propose a general method to resolve the bias by transforming EMR to regular patient hidden condition series using a Hidden Markov Model (HMM) variant. Compared with the biased EMR series with irregular time stamps, the unbiased regular time series is much easier to be processed by most analytic models and yields better results. Extensive experimental results demonstrate that our bias resolving approach imputes missing values more accurately than baselines and improves the performance of the state-of-the-art methods on typical medical data analytics.

http://dl.acm.org/authorize?N33407

210. Optimization Beyond Prediction: Prescriptive Price Optimization

Shinji Ito (NEC coorporation);Ryohei Fujimaki (NEC)

This paper addresses a novel data science problem, prescriptive price optimization, which derives the optimal price strategy to maximize future profit/revenue on the basis of massive predictive formulas produced by machine learning. The prescriptive price optimization first builds sales forecast formulas of multiple products, on the basis of historical data, which reveal complex relationships between sales and prices, such as price elasticity of demand and cannibalization. Then, it constructs a mathematical optimization problem on the basis of those predictive formulas. We present that the optimization problem can be formulated as an instance of binary quadratic programming (BQP). Although BQP problems are NP-hard in general and computationally intractable, we propose a fast approximation algorithm using a semi-definite programming (SDP) relaxation. Our experiments on simulation and real retail datasets show that our prescriptive price optimization simultaneously derives the optimal prices of tens/hundreds products with practical computational time, that potentially improve approximately 30

http://dl.acm.org/authorize?N33463

211. Learning to Count Mosquitoes for the Sterile Insect Technique

Yaniv Ovadia (Google);Yoni Halpern (Google);Dilip Krishnan (Google);Josh Livni (Verily);Daniel Newburger (Verily);Ryan Poplin (Google, Inc.);Tiantian Zha (Verily (Google Life Sciences));D. Sculley (Google, Inc.)

Mosquito-borne illnesses such as dengue, chikungunya, and Zika are major global health problems, which are not yet addressable with vaccines and must be countered by reducing mosquito populations. The Sterile Insect Technique (SIT) is a promising alternative to pesticides; however, effective SIT relies on minimal releases of female insects. This paper describes a multi-objective convolutional neural net to significantly streamline the process of counting male and female mosquitoes released from a SIT factory and provides a statistical basis for verifying strict contamination rate limits from these counts despite measurement noise. These results are a promising indication that such methods may dramatically reduce the cost of effective SIT methods in practice.

http://dl.acm.org/authorize?N33474

212. Embedding-based News Recommendation for Millions of Users

Shumpei Okura (Yahoo! JAPAN);Yukihiro Tagami (Yahoo Japan Corporation);Shingo Ono (Yahoo Japan Corporation);Akira Tajima (Yahoo! Japan)

For effective news recommendation, it is necessary to understand content of articles and preferences of users. While ID-based methods such as collaborative filtering and low rank factorization are well-known approaches for recommendation, such methods are not suitable for news recommendation, because candidate articles expire quickly and replaced by new ones in a short span. Word-based approaches, often used in information retrieval settings, are good candidates in terms of system performance, but have some challenges such as coping with synonyms and orthographical variants and defining “queries” from users’ historical activities. In this paper,we propose an embedding-based approach to use distributed representations in an end-to-end manner: (i) start with distributed representations of articles based on a variant of denoising autoencoder, (ii) generate user representations by a recurrent neural network (RNN) with browsing histories as input sequences, and (iii) match and list articles for each user based on inner product operations in consideration of system performance. The proposed method showed good performance in the offline evaluation using past access data on Yahoo! JAPAN’s homepage. In response to the experimental result, we implemented it to the actual system and compared online performance with the word-based approach that had incorporated in the system traditionally. As a result, CTR and total duration improved by 23

http://dl.acm.org/authorize?N33473

213. Extremely Fast Decision Tree Mining for Evolving Data Streams

Albert Bifet (Telecom ParisTech);Jiajin Zhang (Noah's Ark Lab, Huawei);Wei Fan (Huawei Noah's Ark Lab);Cheng He (Noah's Ark Lab, Huawei);Jianfeng Zhang (Noah's Ark Lab, Huawei);Jianfeng Qian (Huawei Noah's Ark Lab);Geoffrey Holmes (University of Waikato);Bernhard Pfahringer (University of Waikato)

Nowadays real-time industrial applications are generating a huge amount of data continuously every day. To process these large data streams, we need fast and efficient methodologies and systems. A useful feature desired for data scientists and analysts is to have easy to visualize and understand machine learning models. Decision trees are preferred in many real-time applications for this reason, and also, because combined in an ensemble, they are one of the most powerful methods in machine learning.

In this paper, we present a new system called streamDM-C++, that implements decision trees for data streams in C++, and that has been used extensively at Huawei. Streaming decision trees adapt to changes on streams, a huge advantage since standard decision trees are built using a snapshot of data, and can not evolve over time. streamDM-C++ is easy to extend, and contains more powerful ensemble methods, and a more efficient and easy to use adaptive decision tree. We compare our new implementation with VFML, the current state of the art implementation in C, and show how our new system outperforms VFML in speed using less resources.

http://dl.acm.org/authorize?N33453

214. Local Algorithm for User Action Prediction Towards Display Ads

Hongxia Yang (Alibaba Group);Yada Zhu (IBM);Jingrui He (Arizona State University)

User behavior modeling is essential in computational adver- tisement, which builds users' profiles by tracking their online behaviors and then delivers the relevant ads according to each user's interests and needs. Accurate models will lead to higher targeting accuracy and thus improved advertising performance. Intuitively, similar users tend to have similar behaviors towards the displayed ads (e.g., impression, click, conversion). However, to the best of our knowledge, there is not much previous work that explicitly investigates such similarities of various types of user behaviors, and incorporates them into ad response targeting and prediction, largely due to the prohibitive scale of the problem. To bridge this gap, in this paper, we use bipartite graphs to represent historical user behaviors, which consist of both user nodes and advertiser campaign nodes, as well as edges reflecting various types of user-campaign interactions in the past. Based on this representation, we study random-walk- based local algorithms for user behavior modeling and action prediction, whose computational complexity depends only on the size of the output cluster, rather than the entire graph. Our goal is to improve action prediction by leveraging historical user-user, campaign-campaign, and user- campaign interactions. In particular, we propose the bipartite graphs AdvUserGraph accompanied with the ADNI algorithm. ADNI extends the NIBBLE algorithm to AdvUserGraph, and it is able to find the local cluster consisting of interested users towards a specific advertiser campaign. We also propose two extensions of ADNI with improved ef- ficiencies. The performance of the proposed algorithms is demonstrated on both synthetic data and a world leading Demand Side Platform (DSP), showing that they are able to discriminate extremely rare events in terms of their action propensity.

http://dl.acm.org/authorize?N33499

215. A Hybrid Framework for Text Modeling with Convolutional RNN

Chenglong Wang (Alibaba Group);Feijun Jiang (Alibaba Group);Hongxia Yang (Alibaba Group)

In this paper, we introduce a generic inference hybrid framework for Convolutional Recurrent Neural Network (conv-RNN) of semantic modeling of text, seamless integrating the merits on extracting different aspects of linguistic information from both convolutional and recurrent neural network structures and thus strengthening the semantic understanding power of the new framework. Besides, based on conv-RNN, we also propose a novel sentence classification model and an attention based answer selection model with strengthening power for the sentence matching and classification

respectively. We validate the proposed models on a very wide variety of data sets, including two challenging tasks of answer selection (AS) and five benchmark datasets for sentence classification (SC). To the best of our knowledge, it is by far the most complete comparison results in both AS and SC. We empirically show superior performances of conv-RNN in these different challenging tasks and benchmark datasets and also summarize insights on the performances of other state-of-the-arts methodologies.

http://dl.acm.org/authorize?N33496

216. Multi-view Learning over Retinal Thickness and Visual Sensitivity on Glaucomatous Eyes

Toshimitsu Uesaka (The University of Tokyo);Kai Morino (The University of Tokyo);Hiroki Sugiura (The University of Tokyo);Taichi Kiwaki (The University of Tokyo);Hiroshi Murata (The University of Tokyo);Ryo Asaoka (The University of Tokyo);Kenji Yamanishi (The University of Tokyo)

Dense measurements of visual-field, which is necessary to detect glaucoma, is known as very costly and labor intensive. Recently, measurement of retinal-thickness can be less costly than measurement of visual-field. Thus, it is sincerely desired that the retinal-thickness could be transformed into visual-sensitivity data somehow. In this paper, we propose two novel methods to estimate the sensitivity of the visual-field with SITA-Standard mode 10-2 resolution using retinal-thickness data measured with optical coherence tomography(OCT). The first method called Affine-Structured Nonnegative Matrix Factorization(ASNMF) which is able to cope with both the estimation of visual-field and the discovery of deep glaucoma knowledge. While, the second is based on Convolutional Neural Networks (CNNs) which demonstrates very high estimation performance. We experimentally tested the performance of our methods from several perspectives. We found that ASNMF method worked better for relatively small data size while CNN-based one did for relatively large data size. In addition, some clinical knowledge are discovered via ASNMF. To the best of our knowledge, this is the first paper to address the dense estimation of the visual-field based on the retinal-thickness data.

http:/dl.acm.org/authorize?N33484

217. Internet Device Graphs

Matthew Malloy (comScore);Paul Barford (comScore, University of Wisconsin);Enis Ceyhun Alp (University of Wisconsin);Jonathan Koller (comScore);Adria Jewel (comScore)

Internet device graphs identify relationships between user-centric internet connected devices such as desktops, laptops, smartphones, tablets, gaming consoles, TV's, etc. The ability to create such graphs is compelling for online advertising, content customization, recommendation systems, security and operations. We begin by describing an algorithm for generating a device graph based on IP-colocation, and then apply the algorithm to a corpus of over 2.5 trillion internet events collected over the period of six weeks in the Unitied States. The resulting graph exhibits immense scale with greater than 7.3 billion edges (pair-wise relationships) between more than 1.2 billion nodes (devices), accounting for the vast majority of internet connected devices in the US 1 . Next, we apply community detection algorithms to the graph resulting in a partitioning of internet devices into 100 million small communities

http://dl.acm.org/authorize?N33471

218. STAR: A System for Ticket Analysis and Resolution

Wubai Zhou (Florida International University);Wei Xue (Florida International University);Tao Li (Florida International University);Chunqiu Zeng (Florida International University);Wang Qing (Florida International University);Larisa Shwartz (IBM Research);Genady Ya. Grabarnik (St. John's University)

In large scale and complex IT service environments, a problematic incident is logged as a ticket which contains the ticket summary (system status and problem description). The system administrators log the step-wise resolution description when such tickets are resolved. The repeating service events are most likely resolved by inferring the similar historical tickets. With the availability of reasonably large ticket datasets, we can have an automated system to recommend the best matching resolution for a given ticket summary. In this paper, we first identify the challenges in real-world ticket analysis and develop an integrated framework to efficiently handle those challenges. The framework first quantifies the quality of ticket resolutions using a regression model built on carefully designed features. The tickets along with their quality scores obtained from the resolution quality quantification are then used to train a deep neural network ranking model which outputs the matching scores of ticket summary and resolution pairs. This ranking model allows us to leverage the resolution quality in historical tickets when recommending resolutions for an incoming incident ticket. In addition, the feature vectors derived from the deep neural ranking model can be effectively used in other ticket analysis tasks, such as ticket classification and clustering. The proposed framework is extensively evaluated with a large real-world dataset.

http://dl.acm.org/authorize?N33408

219. Deep Design: Product Aesthetics for Heterogeneous Markets

Yanxin Pan (University of Michigan);Alexander Burnap (University of Michigan);Jeffrey Hartley (General Motors);Richard Gonzalez (University of Michigan);Panos Papalambros (University of Michigan)

Aesthetic appeal is a primary driver of customer consideration over product designs such as automobiles. Product designers must accordingly convey design attributes (e.g., 'Sportiness') that the customer will prefer, a challenging proposition given subjective perceptions of customers belonging to heterogeneous market segments. We introduce a scalable deep learning approach that aims to predict how customers across market segments perceive aesthetic designs, as well as visually interpret "why" the customer perceives as such. An experiment is conducted to test this approach, using a Siamese neural network architecture containing a pair of conditional generative adversarial networks, trained using large-scale product design and crowdsourced customer data. Our results show that we are able to predict how aesthetic design attributes are perceived by customers in heterogeneous market segments, as well visually interpret these aesthetic perceptions. This provides evidence that the proposed deep learning approach may provide an additional means of understanding customer aesthetic perceptions complementary to existing methods used in product design.

http://dl.acm.org/authorize?N33486

220. Using Machine Learning to Improve Emergency Medical Dispatch Decisions

Karen Lavi (Friedrich Miescher Institute, Part of Novartis Research Foundation);Ritvik Kharkar (UCLA);Mathew Kiang (Harvard Public Health Schoool);Christoph Hartmann (Boston Consulting Group);Paul Van Der Boor (McKinsey & Company);Adolfo De Unanue (Instituto Tecnologico Autonomo de Mexico);Leigh Tami (Office of Performance & Data Analytics, City of Cincinnati);Anson Turley (Cincinnati Fire Department);Cedric Robinson (Cincinnati Fire Department);Brandon Crowley (Office of Performance & Data Analytics, City of Cincinnati);Eric Potash (Center for Data Science & Public Policy, University of Chicago);Rayid Ghani (Center for Data Science & Public Policy, University of Chicago)

Emergency medical services (EMS) provide out-of-hospital acute medical care and transport to definitive care for those in need. For many medical incidents, minimizing out-of-hospital time is crucial and directly linked to patients' chance of survival. Therefore, ideally, a transport unit should be sent to every medical incident; But in reality resources are limited. Thus, it is immensely important that medical transport dispatches are as accurate as possible—- sending medical transport units as quickly as possible when necessary and not sending them unnecessarily.

In this paper, we describe our work in partnership with the City of Cincinnati in building a live dispatch system that predicts which incidents will result in hospital transport and require medical transport dispatch. In addition to using historical data on past incidents, our system incorporates weather, temporal, and spatial data. Compared to the current approach being used in Cincinnati, and while using the same available resources, we find that a prediction model uses this enriched data increased dispatch accuracy by 25%—getting faster to 3,000 patients that need hospital transport. Based on these results, this live predictive model is now being tested for deployment in the Fire Department of Cincinnati.

http://www.kdd.org/kdd2017/papers/request/using-machine-learning-to-improve-emergency-medical-dispatch-decisions

221. Automatic Application Identification from Billions of Files

Kyle Soska (CMU);Christopher Gates (Symantec);Kevin Roundy (Symantec);Nicolas Christin (CMU)

Understanding how to group a set of binary files into the piece of software they belong to is highly desirable (e.g., for software profiling, malware detection, or enterprise audits, among others). Unfortunately, it is also extremely challenging: there is absolutely no uniformity in the ways different applications rely on different files, in the ways binaries are signed, or in the versioning schemes used across different pieces of software. In this paper, we argue that, by combining information gleaned from a large number of endpoints (millions of hosts), we can accomplish large-scale application identification automatically and reliably. Our approach relies on collecting metadata on billions of files every day, summarizing it into much smaller "sketches," and performing approximate k-nearest neighbor clustering on non-metric space representations derived from these sketches. We design and implement our proposed system using Apache Spark, show that it manages to process billions of files in a matter of hours and could be used for daily processing, and further show our system manages to successfully identify which files belong to which application with very high precision, and adequate recall.

http://dl.acm.org/authorize?N33482