INFORMS 13TH Data Mining & Decision Analytics Workshop

Technical Sessions

Saturday, 8:00AM - 9:00AM

■ Keynote 01

Hyatt, Cowboy Artists

Opening Remarks and Keynote 1 (Aaron Burciaga)

Keynote Session

Chair: Ali Dag, University of South Dakota, Vermillion, SD, 57069, United States

Presenter

Aaron Burciaga, Booz Allen Hamilton, Fairfax, VA, 22033, United States

Saturday, 9:00AM - 10:15AM

■ SA01

Hyatt, Borein A

Optimization in Data Analytics

General Session

Chair: Bart Paul Gerard Van Parys, MIT, Cambridge, MA, 02139, United States

1 - Bootstrap Robust Analytics

Bart Paul Gerard Van Parys, Cambridge, MA, 02139, United States, Dimitri Bertsimis

We address the problem of prescribing an optimal decision in a framework where its cost depends on uncertain problem parameters Y that need to be learned from data. Earlier work by Bertsimas and Kallus (2014) transforms classical machine learning methods that merely predict Y from supervised training data [(x1; y1); : : : ; (xn; yn)] into prescriptive methods taking optimal decisions specific to a particular covariate context X = x. Their prescriptive methods factor in additional observed contextual information on a potentially large number of covariates X = x to take context specifc actions z(x) which are superior to any static decision z. Any naive use of limited training data may, however, lead to gullible decisions over-calibrated to one particular data set. In this paper, we borrow ideas from distributionally robust optimization and the statistical bootstrap of Efron (1982) to propose a novel prescriptive method based on generalized nearest-neighbors learning which natively safeguards against over tting. Our robust prescriptive method reduces to a tractable convex optimization problem. We illustrate our data-driven decision-making framework and our novel robustness notion on a small news vendor problem.

2 - An Smooth Newton Method for SVM Type Models in Data Analysis

Hongxia Yin, Minnesota State University, Mankato, MN, 56001, United States, Weibing Chen

An smoothing Newton method for two support vector machine (SVM) models in data analysis are given by reformulate their dual problems to nonsmooth projection equation systems, and then is smoothed by the Chen-Harker-Kanzow-Smale smoothing function. We proved the global convergence and local quadratic convergence of the methods. Numerical tests on problems in UCI illustrate the efficiency and robustness of the algorithm comparing to the existing results in literature.

3 - Multiplicative Weights Update in Zero-Sum Game

James Patrick Bailey, Singapore University of Technology and Design, Singapore, Georgios Piliouras

We study the behavior of the Multiplicative Weights Update (MWU) algorithm, a classical and commonly used tool for online learning and decision making, in zero-sum games. The Nash equilibrium (NE) is the most universally agreed upon prediction for games; economists study markets under the assumption agents are at or near a NE, computer scientists report that NE are easy to compute, and optimization/decision theorists show no regret algorithms "converge" on average to a NE. We provide significant results in the opposition to this hypothesis, while at the same time making progress on a classic question in decision theory: What happens when two online optimization algorithms compete against each other? We show that MWU results in divergence from NE and convergence to the boundary of the strategy space. This comes in stark contrast with the standard interpretation of the behavior of MWU in zero-sum games, which is typically referred to as "converging to equilibrium". Our results show that the traditional study of markets, which hinges on the assumption that agents are at or near the Nash equilibrium, is a fundamentally flawed approach when agents are adaptively making decisions. Finally, we argue robustness of this non-equilibrating behavior by generalizing the results to include di erent or decaying learning rates and other Follow-the-Regularized-Leader algorithms, e.g., MWU versus Gradient Descent.

4 - A Better Linear Predictor Motivated by Robust Optimization

Long Zhao, UT Mccombs Bussiness School, Austin, TX, 78712-1277, United States

Regularization is widely used to address multicollinearity generated from overfitting. With the underlying specifications being correct, regularization has been proved to improve the estimation accuracy. However, how to be sure that the assumptions are correct and how good estimation leads to good prediction remain unknown. Here we focus directly on the prediction power of linear predictors instead of estimation accuracy. We leverage on a deep understanding of estimation errors to invent a combination of the classical solution and the robust optimization solution. The latter is chosen such that only the orthogonality information is used. For more than 50 di erent real-world scenarios, the method consistently outperforms other methods that ignore such orthogonality information and outperforms both ridge and lasso regression most of the time.

5 - Allocation Problems in Ridesharing Platforms: Online Matching with Offline Reusable Resources

Pan Xu, University of Maryland, College Park, MD, United States, Karthik A Sankararaman, John Dickerson, Aravind Srinivasan

Bipartite matching markets pair agents on one side of a market with agents, items, or contracts on the opposing side. Prior work addresses online bipartite matching markets, where agents arrive over time and are dynamically matched to a known set of disposable resources. In this paper, we propose a new model, Online Matching with (offline) Reusable Resources under Known Adversarial Distributions (OM-RR-KAD), in which resources on the offline side are reusable instead of disposable; that is, once matched, resources become available again at some point in the future. We show that our model is tractable by presenting an LP-based adaptive algorithm that achieves an online competitive ratio of $1 - E$ for any given $E > 0$. We also show that no non-adaptive algorithm can achieve a ratio of $1 + o(1)$ based on the same benchmark LP. Through a data-driven analysis on a massive openly-available dataset, we show our model is robust enough to capture the application in taxi dispatching services. We also present heuristics that perform well in practice.

6 - Many-server Queues with Autoregressive Inputs

Xu Sun, Columbia University, New York City, NY, 10027, United States

Recent studies have revealed the presence of significant autocorrelation and overdispersion in arrival data at large call centers. Motivated by these findings, we study a class of queueing systems where customers arrive according to a doubly stochastic Poisson point process whose intensities are driven by a time-dependent Cox Ingersoll-Ross (CIR) process. The nonnegativity and autoregressive feature of the CIR process makes it a good candidate for modeling temporary dips and surges in arrivals. We conduct performance analysis of such systems. In particular, we study asymptotic performances such as the queue length and customer delays. The results acknowledge the presence of autoregressive structure in arrivals and produce operational insights into staffing decisions.

■ SA02

Hyatt, Borein B

Statistical Data Analytics

General Session

Chair: Ashwin Venkataraman, New York University, New York, NY, 10011, United States

1 - A Conditional Gradient Approach for Nonparametric Estimation of Mixing Distributions

New York University, New York, NY, 10011, United States Srikanth Jagabathula, Lakshminarayanan Subramanian

A key challenge in estimating mixture models is that the mixing distribution is often unknown and imposing apriori parametric assumptions can lead to model misspecification issues. We propose a new methodology for nonparametric estimation of the mixing distribution. We formulate the likelihood-based estimation problem as a constrained convex program and our key contribution is applying the conditional gradient (aka Frank-Wolfe) algorithm to solve this convex program, showing that it iteratively generates the support of the mixing distribution. We show that our estimator is robust to different ground-truth mixing distributions and outperforms the EM benchmark in two case studies on real data.

2 - Dynamic Aggregation of Consumer Ratings: A Latent Gaussian Process Model

Christof Naumzik, ETH Zurich, Zurich, 8006, Switzerland
Stefan Feuerriegel, Markus Weinmann

Online reviews serve as an important information source for consumer decision-making. When choosing upon a product or service, consumers frequently follow the average rating score, since it should reflect the true quality of a product or service. We argue that, despite its widespread use, the (weighted) average as a form of rating aggregation gives only a suboptimal indicator of quality. One reason is that quality is often subject to variation over time; however, the (weighted) average is slow in adapting to structural changes. As a remedy, we develop a latent Gaussian process model that uncovers the dynamic quality from a rating sequence. Accordingly, our model overcomes several of the inherent limitations of a (weighted) average as a form of rating aggregation: (1) our model caters for the time intervals between ratings, which allows it to dynamically adapt to changes in product quality; (2) the inclusion of latent dynamics reflect the stochastic relationship between ratings and quality, thereby making our model less sensitive to noise; and (3) other informative variables, such as textual sentiment and review helpfulness, are explicitly incorporated. Our rating aggregation is then evaluated using an extensive set of 28,309 restaurant reviews from Yelp, which demonstrates the superiority of our approach in estimating the perceived quality. In fact, our model reduces the corresponding mean absolute error over a moving average by 6.9%. These findings entail direct implications for review platforms: if implemented, our dynamic rating aggregation increases customer satisfaction and contributes to economic welfare.

3 - Person Name Disambiguation Based on Profession

Haimonti Dutta, Assistant Professor, University at Buffalo, Buffalo, NY, 14260, United States

Named Entity Disambiguation (NED) is the task of disambiguating entity mentions in natural language text. In this paper, we present a generative model based on Latent Dirichlet Allocation (LDA) to disambiguate person names occurring in noisy text. Empirical results presented on historical newspaper articles show that this generative model can obtain performance comparable to state-of-the-art techniques.

4 - Short-term Traffic Flow Forecasting using Approximate Bayesian Computation: Adapting to Perturbations

Hani S. Mahmassani, Northwestern University, Transportation Center, Evanston, IL, 60208-4055, United States
Lama Al Hajj Hassan, Ying Chen

Non-periodic perturbations in the traffic state result from changes in the demand structure; such randomness is not captured by typical predictive models, such as time-series models, when forecasting the traffic state. This paper presents the development, implementation, and evaluation of a methodology that accounts for demand driven perturbations and integrates them in traffic state predictive tools. We test and compare several forecasting models, ARIMA, Sample Moving Average, Standard Bayesian, and Approximate Bayesian Computation (ABC), implemented within a rolling horizon framework. The models are tested under typical conditions and under sudden perturbations using historical traffic data and assuming the availability of a continuous stream of demand triggers from external data sources when perturbations occur. The results indicate that ABC quickly adjusts to drastic changes in flow levels, and outperforms the other models when predicting for different horizons ahead. Under typical conditions ABC results in a 10-15% error when predicting up to three 15 minute intervals ahead and the error remains less than 30% when predicting up to two hours ahead. When unusual perturbations occur, ABC with Triggers results in a 15-25% error when predicting up to three 15 minute intervals ahead and the error remains less than 35% when predicting up to two hours ahead. In both scenarios, the errors resulting from ABC are less than those of the other tested models.

5 - Expected Utility Model for Optimizing Design for Resilience

Ramin Giahi, Iowa State University, Ames, IA, 50010, United States, Cameron Mackenzie, Chao Hu

Designers should try to design systems that are resilient to adverse conditions during a system's lifetime. The resilience of a system under time-dependent adverse conditions can be assessed by modeling the degradation and recovery of the system's components. Decision makers in a firm should attempt to find the optimal design to make the system resilient to the various adverse conditions. A risk-neutral firm maximizes the expected profit gained from fielding the system, but a risk-averse firm may sacrifice some profit in order to avoid failure from these adverse conditions. This research models the risk-averse firms with a concave utility function. This risk-averse decision-making method is applied to a design firm determining the resilience of a wind turbine system. Since the optimization model requires a complex Monte Carlo simulation to evaluate the objective function, we use Bayesian optimization to find the optimal design. The results show that, to make the system more resilient, risk-averse firms will pay more in design cost to prevent catastrophic costs of failure. In this case, the system is less likely to fail due to he high resilience of its physical components.

6 - Causation-based Monitoring and Diagnosis for Multivariate Categorical Processes with Ordinal Information

Xiaochen Xian, Madison, WI, 53705, United States
Jian Li, Kaibo Liu

In the monitoring and diagnosis of multivariate categorical processes (MCPs), there may exist causal relationships among multiple categorical variables, where the attribute level of a cause variable influences that of its effect variable. Furthermore, there usually exists natural order among the attribute levels of some categorical variables. In this paper, we leverage Bayesian networks to characterize MCPs with a causal structure, where the categorical variables can be either nominal, ordinal, or a combination of both. We develop one general control chart and one directional control chart, both of which fully exploit the causal relationships and the ordinal information.

■ SA03

Hyatt, Russel A

Data Mining Methods & Applications I

General Session

Chair: Abbas Ehsanfar, Stevens Institute, Hoboken, NJ, United States

1 - An Influence-based Clustering Model on Twitter

Abbas Ehsanfar, Stevens Institute, Hoboken, NJ, United States
Mo Mansouri

This paper introduces a temporal framework for detecting and clustering emergent and viral topics on social networks. Endogenous and exogenous inuence on developing viral content is explored using a clustering method based on the a user's behavior on social network and a dataset from Twitter API. Results are discussed by introducing metrics such as popularity, burstiness, and relevance score. The results show clear distinction in characteristics of developed content by the two classes of users.

2 - If You Can't Beat Them, Join Them: Collective Intelligence Outperforms Artificial Intelligence at Diagnosing Skin Cancer

Erik P. Duhaime, Massachusetts Institute of Technology, Cambridge, MA, 02142, United States

A growing chorus of academics, thought leaders, and politicians warn that artificial intelligence will cause unprecedented job loss, even among highly trained knowledge workers such as doctors. At the same time, advances in information technology have enabled entirely new ways of organizing work, and research has shown that crowds of people oftentimes outperform even their most highly skilled individual members. Here, I leverage a dataset of the diagnoses of 1 state-of-the-art artificial intelligence system and 21 board-certified dermatologists in order to examine what happens when artificial intelligence is combined with - rather than compared to - human intelligence.

3 - A New Signal-Noise Separation Approach for Stock Price Data Based on Empirical Mode Decomposition

Hongxia Yin, Minnesota State University, Mankato, MN, 56001, United States, Fu Qiao

It is common for data analysis in financial market to separate the noise from signal in stock price sequences. It is usually accomplished through the reconstruction of a group of price sequences derived from a data-driven adaptive filter called empirical mode decomposition (EMD). From the perspective of behavioural theory, in financial markets, the order of prices appeared in the time sequence is as important as the distributional properties of the price sequence for signal-noise separation. To this end we propose a new signal-noise separation approach based on Spearman correlation index rather than the statistical inference. Based on our approach, the order in which prices appear could be better characterized. According to our empirical analysis in Chinese stock index price data, two price sequences with similar distribution but different appearance order could be distinguished in signal-noise separation. Moreover, our approach is considerate to the relative importance of the sequences derived from the EMD, so the information loss and the unexpected distribution change caused by the signal-noise separation approach based on statistical inference will not occur anymore.

4 - A Nonlinear Programming & Convolution Smoothing Approach to Clustering

Syed Mujahid, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

In this work, a novel density based clustering algorithm that can identify non-convex separable clusters is presented. Firstly, a Gaussian density function that is smooth, locally concave, and robust is proposed. Secondly, a convolution smoothing based cluster search algorithm is designed that aptly exploits the above function properties. The proposed algorithm is accelerated through the second order function information during its optimization phase. Lastly, the performance of the proposed algorithm is depicted via head-to-head comparison with the well known clustering algorithms.

5 - Crowdsourcing Data Science via Open Contests

Wenjun Zhou, University of Tennessee, Knoxville, TN, 37996, United States

More and more organizations nowadays seek novel solutions to their data-intensive problems by crowdsourcing. Crowdsourced data science problems frequently takes the form of an open contest, in which contributors compete for awards offered by the solution seeker. Extant literature in open innovation showed the optimality of a winner-takes-all award scheme. However, in practice, a top-K scheme is frequently adopted in data science contests. We show that this gap between the literature and common practice is due to unique characteristics of seeker and contributor utilities in data science contests. In particular, seekers may aggregate multiple solutions as an ensemble to obtain higher collective utility, and contributors are allowed to make multiple submissions. Accounting for these characteristics, we formally show that a top-K award scheme is optimal for data science contests in many cases. Our findings are also supported by empirical evidence from Kaggle, one of the largest data science communities.

6 - Statistical Analysis of Efficiency and Profitability of Foreign Banks Operating in India, and of Indian Banks

Badri Toppur, Rajalakshmi School of Business, Chennai, 600124, India

How does the profitability of the scheduled commercial banks operating in India relate to the cost efficiency of the banks? This question is set in the context of the post-liberalization reforms. To address this question, we have analyzed 22 public sector banks, 21 private banks, and 44 foreign banks in a study using Statistical analysis and Data Envelopment Analysis, on the most recent data available (for financial year 2017) with the Reserve Bank of India. The correlation between profitability and efficiency is proven to be insignificant for all three categories of banks, though the efficiency of foreign banks is proven to be higher.

■ SA04

Hyatt, Russel BC

Network Analytics

General Session

Chair: Serhat Simsek, Auburn University, Auburn, AL, 36830, United States

1 - Do Analysts Mislead Medical Practitioners? A Comprehensive Analytics Technique to Better Detect Non-surviving Cancer Patients

Serhat Simsek, Auburn University, Auburn, AL, 36830, United States, Eyyub Kibis, Ali Dag

Analysis of survival times of cancer patients is crucial for medical practitioners to determine possible outcomes and make better future-plans for the patients. In the healthcare analytics literature, it is common to see employment of machine learning algorithms to predict survivability of the cancer patients. In this study, we detected the common misleading methodology that has been used in the literature in predicting the surviving and non-surviving cancer patients and propose a comprehensive modeling technique that overcomes the issue of misprediction of survival. In order to illustrate the issue and its solutions, we deploy Principle Component Analysis, K-Nearest Neighbors, Artificial Neural Networks, Support Vector Machine, Random Forest and Logistic Regression. The comprehensive model is applied on rectum cancer data and results are validated with breast cancer data from the SEER. The results will assist medical practitioners to make better decisions for their patients and thus appropriate interventions.

2 - Submodularity on Hypergraphs: From Sets to Sequences

Amin Karbasi, Yale University, New Haven, CT, 06520, United States

In a nutshell, submodular functions encode an intuitive notion of diminishing returns. As a result, submodularity appears in many important machine learning tasks such as feature selection and data summarization. Although there has been a large volume of work devoted to the study of submodular functions in recent years, the vast majority of this work has been focused on algorithms that output sets, not sequences. However, in many settings, the order in which we output items can be just as important as the items themselves.To extend the notion of submodularity to sequences, we use a directed graph on the items where the edges encode the additional value of selecting items in a particular order. Existing theory is limited to the case where this underlying graph is a directed acyclic graph. In this paper, we introduce two new algorithms that provably give constant factor approximations for general graphs and hypergraphs having bounded in or out degrees. Furthermore, we show the utility of our new algorithms on a real-world application in movie recommendation.

3 - Benchmarking Tabu Search Based Markov Blanket Attribute Selection and Classification

Daniel Gartner, Cardiff University, Cardiff, United Kingdom
John Threlfall, Rema Padman, Paul Harper

Datasets with many discrete variables and relatively few observations are generated in domains such as health care and electronic commerce. Learning which variables are relevant and nonredundant effectively and efficiently is a difficult task. Moreover, achieving high accuracies for making predictions is challenging. In this paper, we implement and evaluate a Java-based approach that combines a construction heuristic with a Tabu-search based improvement heuristic to learn a graphical Markov Blanket-based classifier from data. Computational results from benchmark data sets in different domains indicate that the method lead to competitive classification results. The graphical models which are generated can be interpreted by practicioners such as physicians. In addition, our results reveal that the graphical models learned from the data have substantially less predictor variables than in the full data set. Results of the approaches are analyzed and broken down by computation times, size of the graphs and classification accuracy.

4 - Block-Structure Based Time-Series Models for Graph Sequences

Mehrnaz Amjadi, UIC, Chicago, IL, United States
Theja Tulabandhula

Although the computational and statistical trade-off for modeling single graphs, for instance, using block models is relatively well understood, extending such results to sequences of graphs has proven to be difficult. In this work, we take a step in this direction by proposing a model for graph sequences that captures: (a) link persistence between nodes across time, and (b) community persistence of each node across time. We provide statistically and computationally efficient inference algorithms, whose unique feature is that they leverage community detection methods that work on single graphs. We also provide experimental results validating the suitability of our model and methods.

5 - Graphical Lasso and Thresholding: Equivalence and Closed-Form Solutions

Salar Fattahi, University of California, Berkeley, CA, 94702, United States, Somayeh Sojoudi

Graphical Lasso (GL) is a popular method for learning the structure of an undirected graphical model, which is based on an l1 regularization technique. The objective of this paper is to compare the computationallyheavy GL technique with a numerically-cheap heuristic method that is based on simply thresholding the sample covariance matrix. To this end, two notions of sign-consistent and inverse-consistent matrices are developed, and then it is shown that the thresholding and GL methods are equivalent if: (i) the thresholded sample covariance matrix is both sign-consistent and inverse-consistent, and (ii) the gap between the largest thresholded and the smallest un-thresholded entries of the sample covariance matrix is not too small. By building upon this result, it is proved that the GL method-as a conic optimization problem-has an explicit closed-form solution if the thresholded sample covariance matrix has an acyclic structure. This result is then generalized to arbitrary sparse support graphs, where a formula is found to obtain an approximate solution of GL. Furthermore, it is shown that the approximation error of the derived explicit formula decreases exponentially fast with respect to the length of the minimum-length cycle of the sparsity graph. The developed results are demonstrated on massive randomly generated datasets. We show that the proposed method can obtain an accurate approximation of the GL for instances with the sizes as large as 80,000×80,000 (more than 3.2 billion variables) in less than 30 minutes on a standard laptop computer running MATLAB, while other state-of-the-art methods do not converge within 4 hours.

6 - Are Extreme Value Estimation Method Useful for Network Data?

Tiandong Wang, Cornell University, Ithaca, NY, United States
Phyllis Wan, Richard A Davis, Sydney I Resnick

Preferential attachment is an appealing edge generating mechanism for modeling social networks. It provides both an intuitive description of network growth and an explanation for the observed power laws in degree distributions. However, there are often difficulties fitting parametric network models to data due to either model error or data corruption. In this paper, we consider semi-parametric estimation based on an extreme value approach that begins by estimating tail indices of the power laws of in- and out-degree for the nodes of the network using nodes with large in- and out-degree. This method uses tail behavior of both the marginal and joint degree distributions. We compare the extreme value method with the existing parametric approaches and demonstrate how it can provide more robust estimates of parameters associated with the network when the data are corrupted or when the model is misspecifed.

Saturday, 10:30AM - 11:45AM

■ Keynote 02

Hyatt, Cowboy Artists

Keynote Data Analytics and Decision Analytics Models across Different Domains

Keynote Session

Chair: Ali Dag, University of South Dakota, Vermillion, SD, 57069, United States

Data Analytics and Decision Analytics Models across Different Domains

Subodha Kumar, Mays Business School, Texas A&M University, College Station, TX, 77843, United States

We would discuss the applications of data analytics and decision analytics models across different domains, such as omni-channel retailing, social media, healthcare, and digital marketing. In the omni-channel retailing, customers often evaluate products at brick-and-mortar stores to identify their "best fit" product but buy it for a lower price at a competing online retailer. We begin with analyzing this free-riding behavior by customers (referred to as "showrooming") and show that this is detrimental to the profits of the brick-and-mortar stores. Next, we examine price matching as a short-term strategy and exclusivity of product assortments as a long-term strategy to counter showrooming. Next, we empirically examine the following two key questions in the context of multi-channel retailing: (i) What is the role of product returns in the evolution of new consumer's channel preference and channel choice? (ii) What is the impact of easier store access to customers on their online purchase behavior? For both of these studies, we use a unique panel dataset of consumers' purchase and return of products from different categories from a multi-channel department store. In the social media domain, we first empirically analyze the operational effects of social media content created by human brands (or influencers) on audience engagement. Then, we develop a data-driven optimization framework to help a firm successfully conduct an influencer marketing campaign. In the healthcare domain, based on our interactions with three different healthcare information exchange (HIE) providers, we first develop models to study the sustainability of HIEs and participation levels in these networks. Next, we empirically examine the impact of HIE use in emergency departments (EDs) on quality and efficiency of medical care. We focus on 30-day readmission rate to capture healthcare quality. From hospitals' efficiency perspective, we examine if HIE access reduces the length of stay (LOS) and the number of doctors who would be consulted for a treatment. Finally, in the digital marketing domain, we provide an approach to manage an ongoing Internet ad campaign that substantially improves the number of clicks and the revenue earned from clicks. The problem we study is faced by an Internet advertising firm (Chitika) that operates in the Boston area. We first develop a predictive model of a visitor clicking on a given ad. Using this prediction of the probability of a click, we develop a decision model that uses a threshold to decide whether or not to show an ad to the visitor.

Saturday, 1:30PM - 2:45PM

■ Keynote 03

Hyatt, Cowboy Artists

Keynote Big Data Analytics Research in Information Systems

Keynote Session

Chair: Ali Dag

University of South Dakota, Vermillion, SD, 57069, United States

1 - Big Data Analytics Research in Information Systems

Roger Chiang, University of Cincinnati, Cincinnati, OH, United States

Big data analytics (BDA) evolving from business intelligence and analytics (BI&A) has become a significant research topic for Information Systems (IS) discipline. From the technical perspective presented in a MISQ article (Chen et al., 2012), it ranges from BI&A 1.0, BI&A 2.0, to BI&A 3.0. Gradually, BDA attracts considerable attention from the whole information systems (IS) discipline, with several commentaries, editorials, and special issue introductions on the topic appearing in leading IS outlets. These articles present varying perspectives on promising BDA research topics and highlight some of the challenges that big data poses. In a JAIS editorial (Abbasi et al., 2016) authors offered a first step toward an inclusive big data research agenda for IS by focusing on the interplay between big data's characteristics, the information value chain encompassing people-process-technology, and the three dominant IS research traditions (behavioral, design, and economics of IS). Despite the publicity regarding big data and analytics, the success rate of these projects and strategic value created from them are unclear. In addition to the four V's (volume, velocity, variety, and veracity) characterization of big data, value has been considered the fifth key dimension in BDA. Analysis of data without generating strategic value offers no contribution to a business, regardless of whether data are big or small. In a recent JMIS article (Grover et al., 2018), authors offered a framing of BDA value by extending existing frameworks of information technology value. Their value creation framework includes BDA infrastructure, BDA capability, value creation mechanisms, value targets, and impacts. They also illustrated the framework through BDA applications in practice.

Saturday, 3:00PM - 4:15PM

■ SB01

Hyatt, Borein A

Neural Networks & Deep Learning

General Session

Chair: Arslan Ali Syed, BMW Group, BMW Group, Garching, 81375, Germany"

1 - Instance Based Meta-Heuristic Parameterization using Neural Network: With Application to Static Car-Passenger Matching Problem

Arslan Ali Syed, Bayerische Motoren Werke, Muenchen, 81375, Germany, Irina Gaponova, Klaus Bogenbeger

Majority of transportation problems include optimizing some sort of cost function. These optimization problems are often NP-Hard and have an exponential increase in computation time with the increase in model size. The problem of matching vehicles to passenger requests in On Demand Mobility (ODM) contexts typically fall into this category. Metaheuristics are often utilized for such problems with the aim of finding global optimal solution. But such algorithms usually include lots of parameters that need to be tuned for getting a good performance. Typically multiple simulations are run on diverse small size problems and the parameters values that perform the best on average are chosen for the subsequent larger simulations. Contrary to the above approach, we propose training a neural network to predict the parameter values that work the best for the given problem instance. We show that various features, based on the problem instance and Shareability graph's statistics, can be used to predict the solution quality of a matching problem in ODM services. Consequently, the values corresponding to the best predicted solution can be selected for the actual problem.We study the effectiveness of above approach for the static assignment of vehicles to passengers in ODM services. We utilize the DriveNow data from Bavarian Motor Works (BMW) for generating passenger requests inside Munich, and for the metaheuristic, we use Large Neighborhood Search (LNS) algorithm combined with Shareability graph.

2 - Monitoring and Forecasting with Big Data for Enhancing Energy Supply Chain Management

Edward W Sun, KEDGE Business School, Talence, 33405, France
Yi-Ting Chen, Yi-Bing Lin

Based on the application of big data technology, we propose a novel quantitative method originating from learning the parallel neural networks (PNNs) for robust monitoring and forecasting power demand immediately to enhance servitization for a power manufacturing firm. We generalize the implementation by using real data from Australia. The overall empirical results confirm that our proposed method improves the robust performance significantly for dynamic monitoring and forecasting of power demand.

3 - A Genetic Algorithm Based Deep Learning Approach to Understand Genotype-Environment Interaction (G x E)

Guiping Hu, Iowa State University, Ames, IA, 50011, United States

The world population continues to increase which imposes rising demand in agriculture production. How to improve crop breeding to feed the growing population is a significant challenge. The traditional crop breeding is resource intensive and time limited. Predictive modeling on crop phenotype can speed up the process and make it resource efficient. Understanding the effects of the genotype and environment factors on the crop phenotypes is critical, which is also known as the Genotype by Environment (G x E) interaction problem. In this paper, we established a deep neural network for the prediction of yield difference using the genetic data and environment data. A genetic algorithm (GA) based solution method is proposed to evolve the weights of the neural networks. With the embedded batch strategy, the algorithm is also able to handle large datasets. The experimental results show that the new approach can improve the performance of the neural networks in prediction accuracy.

4 - Event Log Reconstruction Using Autoencoders

Jonghyeon Ko, UNIST, Ulsan, Korea, Republic of
Hoang Thi Cam Nguyen, Marco Comuzzi

Poor quality of process event logs prevents high quality business process analysis and improvement. Process event logs quality decreases because of missing attribute values or after incorrect or irrelevant attribute values are identified and removed. Reconstructing a correct value for these missing attributes is likely to increase the quality of event log-based process analyses. Traditional statistical reconstruction methods work poorly with event logs, because of the complex interrelations among attributes, events and cases. Machine learning approaches appear more suitable in this context, since they can learn complex models of event logs through training. This paper proposes a method for reconstructing missing attribute values in event logs based on the use of autoencoders. Autoencoders are a class of feed-forward neural networks that reconstruct their own input after having learnt a model of its latent distribution. They suit problems of unsupervised learning, such as the one considered in this paper. When reconstructing missing attribute values in an event log, in fact, one cannot assume that a training set with true labels is available for model training. The proposed method is evaluated on two real event logs against baseline methods commonly used in the literature for imputing missing values in large datasets.

5 - An Optimum Profit-driven Churn Decision Making Framework using an Innovative Artificial Neural Network in Telecommunications Industry

Roy Jafari, California Polytechnic State University, San Luis Obispo, CA, 93407, United States, Adnan Idris, Brian Smith, Josh Denton, Abbas Keramati

Churn refers to customers ceasing use of services provided by a firm and beginning use of the same service from one or more competitors. Knowledge-based churn prediction and decision making is invaluable for telecommunication companies due to their highly competitive market. The comprehensiveness and action- ability of a data-driven churn prediction system depends on an effective extraction of hidden patterns in the data. Generally, data analytics is employed to extrapolate the extracted patterns from the training dataset to the test set. In this work, one more step is taken; the improved prediction performance is attained by capturing the individuality of each customer while discovering the hidden pattern from the train-sent and before applying the discovered knowledge to the test set. The proposed churn prediction system is developed using artificial neural networks that takes advantage of both self-organizing and error driven learning approaches (ChP-SOEDNN). We are introducing a new dimension to the study of churn prediction phenomenon in Telecom industry which is a systematic profit-driven churn decision making framework. The comparison of the ChP-SOEDNN with other techniques shows its superiority regarding both accuracy and misclassification cost. Keywords: Churn prediction, Cost-sensitive classification, Profit-driven data analytics, Artificial neural networks (ANNs), Self-organizing map (SOM), Self-Organizing Error Drive (SOED)

6 - A Deep Learning Model for Traffic Flow State Classification Based on Smart Phone Sensor Data

Wenwen Tu, Southwest Jiaotong University, Chengdu, China
Feng Xiao, Liping Fu, Guangyuan Pan

This study proposes a Deep Belief Network model to classify traffic flow states. The model is capable of processing massive, high-density, and noise-contaminated data sets generated from smartphone sensors. The statistical features of vehicle acceleration, angular acceleration, and GPS speed data, recorded by smartphone software, are analyzed, and then used as input for traffic flow state classification. Data collected by a five-day experiment is used to train and test the proposed model. A total of 747,856 sets of data are generated and used for both traffic flow states classification and sensitivity analysis of input variables. When the parameters of the DBN model are optimized by the Differential Evolution Grey Wolf Optimizer algorithm, the classification accuracy is further improved. The results have demonstrated the effectiveness of using smartphone sensor data to estimate the traffic flow states and shown that the proposed Deep Belief Network model is superior to traditional machine learning methods in both classification accuracy and computational efficiency.

■ SB02

Hyatt, Borein B

Business Analytics

General Session

Chair: Murtaza Nasir, University of South Dakota, Vermillion, SD, 57069, United States

1 - Business Analytics in the Managerial Job Market

Murtaza Nasir, University of South Dakota, Vermillion, SD, 57069, United States, Ali Dag, William Young, Dursen Delen

In this analysis, we present a simple data-driven framework to understand the job market requirements and trends for business analytics in terms of the demand and value of different knowledge, skills and abilities (KSAs) based on region, career path and career stage. Analytics and big data have transformed the business landscape in the last decade, with Harvard Business Review calling big data the source of a “management revolution”. Capitalizing on these opportunities have allowed some businesses to gain massive competitive advantages by not only improving their core operations but also by creating completely new business models. But to effectively capitalize on the business opportunities presented by analytics and big data, companies first have to ensure they have the right management orientation and talent demanded by these opportunities. To understand the fast changing business analytics environment, business schools and educators have mostly looked at the industry to understand the requirements of employers, but in most cases, as existing literature shows, they have been lagging behind in terms of the analytical and technical skill sets and their integration with business skills being imparted to graduates. Our simple framework extends the understanding of this job market and in the process, also defines some simple ways for all participants, including educators, employers and professionals, to understand whether they are ahead of or behind the curve, as defined by the managerial job market’s aggregated views on analytics. The framework allows for easy interpretation and inference of actionable insights and is also generalizable to other fields.

2 - Leveraging Comparables for New Product Sales Forecasting

Divya Singhvi, MIT, Cambridge, MA, 02139, United States
Lennart Baardman, Igor Levin, Georgia Perakis

Many firms regularly introduce new products, and before their launch, firms need to make various operational decisions guided by sales forecasts. The new product sales forecasting problem is challenging when compared to forecasting sales of existing products, because we lack historical sales data. We propose a novel sales forecasting approach using data from comparable past products. We formulate the problem of clustering products and fitting forecasting models to these clusters simultaneously as a quadratic mixed integer optimization problem. In estimation, we use regularization to attenuate the overfitting problem. This problem is computationally hard, and thus, we develop a scalable algorithm that produces a forecasting model with analytical guarantees on the prediction error.For this forecasting method, we prove a finite sample prediction error guarantee. Collaborating with Johnson & Johnson Consumer Companies Inc., a major consumer goods manufacturer, we show an average improvement of 40% in out-of-sample forecasting metrics such as MAPE and WMAPE over various product segments. For the consumer goods manufacturer, we develop a fast and easy-to-use Excel tool that aids managers with forecasting and decisions-making before a product launch.

3 - Data Mining to Enhance Customer Satisfaction within Interactive Voice Response Systems

Vinoth Kumar Raja, West Corporation, Omaha, NE, United States
Shruti Palasamudram Ramesh, Dmitriy Khots

Interactive Voice Response (IVR) systems play a pivotal role in today’s world of communication. It provides callers an automated solution or a quick route to an agent. Businesses love IVR systems because they take out hundreds of millions of dollars of call center costs in automation of routine tasks, while consumers hate IVRs as they often look at it as a barrier to overcome in order to talk to a real person. So it is important IVRs are managed in such a way that it saves money for business and at the same time minimizes consumer abrasion. While managing IVR systems is critical, listening to customer feedback is of paramount importance as survey scores help in determining the likes and dislikes about the automated system. The loop of listening to customers’ voice through survey scores and improving services based on analytics will lead to better business and more importantly enhancement in customer satisfaction. This paper discusses how data mining techniques can be utilized to improve customer satisfaction within IVR. The analysis is performed on one of West Corporation’s leading financial services client whose customers’ satisfaction scores, on a likert scale of 0 to 10, are based on their experience with IVR system. The scores along with their respective IVR call data are transformed into inputs and is used to build a statistical model. The model quantifies relationship to help us understand the call experience of customers inside IVR. Moreover, the same predictors are used to score new calls that were not part of the survey and helps us understand caller behavior. This paper illustrates how Customer Insights and Business Intelligence Group (CIBIG) at WEST prepared the data and built such model to improve customer experience.

4 - Data-Driven Portfolio Optimization with Drawdown Constraints Utilizing Machine Learning

Meng-Chen Hsieh, Rider University, Lawrence Township, NJ, 08648, United States

In practice, data-driven optimal portfolio decisions are derived based on the time series data of target asset returns. Such data-driven optimization decision rules are prone to inferior out-of-sample performance due to estimation errors of parameters plugged in the optimization setting. In the ’big data’ era, correlations between target asset returns and auxiliary variables are frequently observed. These auxiliary variables have the potential to provide valuable information on their association with the target asset returns and thus may be able to improve the out-of-sample performance of the constructed optimal portfolio. In this work, we consider a portfolio optimization problem with drawdown constraints. We apply machine learning methods to leverage the association between target asset returns and auxiliary variables to derive optimal portfolio decisions. A comparison study on the out-of-sample performance of the constructed portfolio with and without utilizing machine learning methods shows the improvement of implementing machine learning in optimal portfolio decisions.

5 - A Data-Driven Forecasting Approach for Newly Launched Seasonal Products

Tugba Efendigil, MIT, Cambridge, MA, United States
Vicky Wing Kei Chan, Majd Kharfan

Demand forecasting is becoming a very complicated process in fashion industry due to the short product lifecycles, the obsolescence of the retail calendar, and the lack of information for newly launched seasonal items. Therefore, this study focuses on demand prediction with a data-driven perspective both leveraging machine-learning techniques and identifying significant predictor variables to help fashion industry including apparel and footwear retailers achieve better forecast accuracy.

6 - Improving Demand Forecast Accuracy Through Forecasting Individual ARMA or Partially Aggregated ARMA Processes

Vladimir Kovtun, Yeshiva University Sy Syms School of Business, New York, NY, 10016, United States

Modern-day technologies not only permit firms to accurately track their point of sales data and lost sales data but also gather more granular data. These data streams deluge firms with information which either can be aggregated for planning purposes or considered in its entirety or follow some where in-between approach. In this paper we analyze a model in which a retailer is faced with exactly the same choices and provide guidelines for combining the data for the purpose of forecasting demand.

■ SB03

Hyatt, Russel A

Data Mining Methods & Applications II

General Session

Chair: Zahra Sedighi-Maman, Auburn University, Auburn, AL, 36830, United States

1 - A Data Analytic Framework for Physical Fatigue Management using Wearable Sensors

Zahra Sedighi-Maman, Adelphi University, Garden City, NY, 11530, United States, Ying-Ju Chen, Amir Baghdadi, Seamus Lombardo, Lora Cavuoto, Fadel Megahed

This paper lays out a framework for the use of wearable sensors to detect and manage worker fatigue in manufacturing environments. The proposed framework includes four main phases: fatigue detection, identification, diagnosis, and recovery. Based on input data, the framework establishes criteria for feature and machine learning algorithm selection for detection. A specific application case of the framework, for one manufacturing-related task, is presented to illustrate the specific considerations for data processing and results interpretation.

2 - Data-driven Location Planning with Latent Spatial Modeling

Daniel Tschernutter, ETH Zurich, Zurich, 8092, Switzerland
Stefan Feuerriegel

Location planning for business sites commonly involves discretization and subsequent optimization with respect to distances or maximal covering, rather than the expected utility. Conversely, we propose a data-driven approach to spatial decision-making that explicitly infers the latent spatial utility and then optimizes over it. For this purpose, the past use of business sites is modeled based on business-specific characteristics that explain the between-business heterogeneity and spatial interactions between neighboring businesses. Both components can be parametrized as, e.g., linear models or even neural networks in order to incorporate non-linear effects, while additional spatial heterogeneity is handled through a Gaussian process. We show that the between-business interactions can be effectively computed by formulating this problem as a multi-dimensional fixed-point theorem. An optimization procedure over the latent dynamics is proposed and it is further extended to multi-location planning. The effectiveness of our approach is demonstrated through computational experiments with simulated data and actual check-ins from over 6,000 restaurants.

3 - A Machine Learning Approach to Shipping Box Design

Guang Yang, Jet.com/Walmart Labs, Hoboken, NJ, United States
Cun (Matthew) Mu

Having the right assortment of shipping boxes in the fulfillment warehouse to pack and ship customer's online orders is an indispensable and integral part of nowadays eCommerce business, as it will not only help maintain a profitable business but also create great experiences for customers. However, it is an extremely challenging operations task to strategically select the best combination of tens of box sizes from thousands of feasible ones to be responsible for hundreds of thousands of orders daily placed on millions of inventory products. In this paper, we present a machine learning approach to tackle the task by formulating the box design problem prescriptively as a generalized version of weighted k-medoids clustering problem, where the parameters are estimated through a variety of descriptive analytics. We test this machine learning approach on fulfillment data collected from Walmart U.S. eCommerce, and our approach is shown to be capable of improving the box utilization rate by more than 10%.

4 - Optimal Placement of Actuators for Composite Fuselage Shape Control

Juan Du, Peking University, Beijing, 100871, China
Xiaowei Yue, Jeffrey H Hunt, Jianju Shi

Actuator placement is critical and challenging for shape control due to dimensional variabilities of composite fuselages. This paper proposes an optimal actuator placement methodology by developing a sparse learning model and an ADMM-based parameter estimation algorithm. A case study shows that our method reduces dimensional deviations and actuator forces effectively.

5 - Exploring Student Matriculation and Admissions Melt in Graduate Business Programs

Liye Sun, Purdue University, West Lafayette, IN, 47906, United States, Lorena V. Bustamante, Matthew A Lanham Kimberly J. Lanham

With increasing demand of analytics talents in companies, both the demand and supply of master programs in Business Analytics and Data Science has increased rapidly in recent years. In this study, we aim to build a predictive model that can accurately predict whom would matriculate in an analytical graduate program, if provided an admission offer. For this, we analyzed the application and matriculation rates over the past two years to develop a predictive model of whom is most likely to attend if provided an offer. From a planning perspective, being able to estimate the matriculation rate could provide valuable decision-support for program preparation.

6 - Incremental Learning for Nonstationary Traffic Control in Automated Vehicle Systems

Sang Min Lee, Korea University, Seoul, Korea, Republic of

A large-scale automated vehicle system (AVSs), designed to transfer materials through thousands of vehicles, are popularly utilized in wafer transfer tasks in semiconductor manufacturing. With the necessity of controlling traffic during the transfer, especially in preventing congestion of the vehicles, prediction models have been proposed to reflect traffic patterns of routes. However, in real-life traffic situation, there exists a nonstationary traffic environment, which renders change in relations of traffic patterns over time, which hinders the prediction model from detecting route congestion. To overcome this problem, we present an incremental learning framework for adaptive traffic control in AVSs. We propose a change-aware learning method that combines a change detector with incremental learning algorithms. To demonstrate the effectiveness and efficiency of the proposed method, we conducted an experimental study to evaluate the predictive performance of incremental learning algorithms. The experimental results confirm that AVSs with the proposed method demonstrate outperformance by updating itself whenever salient changes of traffic patterns occur.

7 - Uncovering Service Quality's Impact upon Ecommerce through the Mining Electronic Word of Mouth Reviews

Benjamin George, University of South Dakota, Vermillion, SD, United States, Bartlomiej Hanus

User generated content is an essential element of the modern day online purchasing decision. Though widely recognized as influential, little research has been investigating the presence and impact of both service and product quality dimension in online customer reviews. These dimensions have the potential to affect online purchasing behaviors. Such potential might mislead the online retailers in evaluating the customer e-feedbacks about their product. Isolating such effect of service quality on online customer ratings would provide better insights to both online retailers and online consumers. The goal of the present study is to explore the contents of online reviews and examine how service quality elements influence product ratings. More specifically, this study investigates the following questions: How prevalent are service quality dimensions within online product reviews? How does the presence of service quality dimensions affect the overall (star) product rating? Does service quality affect the product quality evaluation? To investigate the above research questions, we scraped product reviews from Amazon.com and examined the data using latent semantic analysis and hierarchical clustering to confirm the presence of service quality component and its impact on product ratings in these reviews. Our results indicate that service quality is a significant factor in predicting product ratings in e-commerce setting. Hence, it should be accounted for when evaluating online customer reviews, both from the buyer and the seller perspectives.

Saturday, 4:30PM - 5:45PM

■ SC01

Hyatt, Borein A

Student Finalist

General Session

Chair: Ramin Moghaddas, University of Miami, Palmetto Bay, FL, 33176, United States

1 - Maximum a Posteriori Estimate of the Max-flow Problem for Defect Localization in Noisy Industrial Images

Asif S. Iquebal, Texas A&M University, College Station, TX, 77843, United States, Satish Bukkapatnam

We present a novel unsupervised max-flow formulation by iteratively estimating the label cofiguration and the flow capacities. More specifically, we show that the maximum a posteriori estimate of image segmentation can be formulated as a capacitated max-flow problem over a continuous domain given the ow capacities are known. The maximum likelihood estimate of the flow capacity, in turn, is then obtained by considering a Markov random field prior over the neighborhood structure in the image. Experimental results on industrial applications suggest the that proposed method is able to segment various defects (pores as well as balling effect) from the additive manufactured surfaces. Comparative results with five state-of-the-art methods suggest that the proposed method is able to exclude noise and selectively segment the defects. From a process standpoint, segmentation results indicate that the defect concentration in a specific additive manufacturing process reduces by more than 75% by optimally controlling the laser power and laser scan speed.

2 - Balanced Random Survival Forests for Mortality Prediction from Extremely Unbalanced, Right Censored Data

Kahkashan Afrin, Texas A&M University, College Station, TX, 77840, United States, Gurudev Illangovan, Sanjay S. Srivatsa T.S. Bukkapatnam

Cardiovascular diseases (CVD) are the leading cause of death worldwide. Critical decisions related to the emergency services and treatment options for CVDs are made based on the prediction of life expectancy and risk using survival models. However, accuracies of survival models for such critical-care applications are significantly compromised due to the sparsity of samples and extreme imbalance between the survival (usually, the majority) and mortality class sizes. Imbalanced datasets result in an underestimation (overestimation) of the hazard of the mortality (survival) classes. A balanced random survival forests (BRSF) model, based on training the RSF model with data generated from a synthetic minority sampling scheme is presented to address this gap. Theoretical results on the effect of balancing on prediction accuracies in BRSF are reported. Benchmarking studies were conducted using five datasets with different levels of class imbalance from public repositories and an imbalanced dataset of 267 acute cardiac patients, collected at the Heart, Artery, and Vein Center of Fresno, CA. Investigations suggest an improved discriminatory strength between the survival and the mortality classes using BRSF. It outperformed both optimized Cox (the current gold standard) and RSF with an average reduction of 55% in the prediction error over the next best alternative.

3 - Spectral Algorithms for Computing Fair Support Vector Machines

Mahbod Olfat, University of California-Berkeley, CA, 94702, United States, Anil Aswani

Classifiers and rating scores are prone to implicitly codifying biases, which may be present in the training data, against protected classes(i.e., age, gender, orrace). So it is important to understand how to design classifiers and scores that prevent discrimination in predictions. This paper develops computationally tractable algorithms for designing accurate but fair support vector machines (SVM's). Our approach imposes a constraint on the covariance matrices conditioned on each protected class, which leads to a nonconvex quadratic constraint in the SVM formulation. We develop iterative algorithms to compute fair SVM's, which solve a sequence of relaxations constructed using a spectral decomposition of the nonconvex constraint. Its effectiveness in achieving high prediction accuracy while ensuring fairness is shown through numerical experiments on several data sets.

4 - Elm-som: A Continuous Self-organizing Map for Visualization

Renjie Hu, University of Iowa/ University of Houston, Houston, TX, 77054, United States

This paper presents a novel dimensionality reduction technique: ELM-SOM. This technique preserves the intrinsic quality of Self-Organizing Maps (SOM): it is nonlinear and suitable for big data. It also brings continuity to the projection using two Extreme Learning Machine (ELM) models, the first one to perform the dimensionality reduction and the second one to perform the reconstruction. ELM-SOM is tested successfully on six diverse datasets. Regarding reconstruction error, ELM-SOM is comparable to SOM while bringing continuity.

■ SC02

Hyatt, Borein B

Non-Student Finalist

General Session

Chair: Shouyi Wang

University of Texas-Arlington, Arlington, TX, 76019, United States

1 - Selection of Hierarchical Features via Sparse Group Regularization

Kin Ming Puk, University of Texas at Arlington, Arlington, TX, 76013, United States, Jay M Rosenberger, Shouyi Wang

This research presents a framework of hierarchical sparse group lasso (HSGL), which selects features arranged in a hierarchical manner with sparse group regularization. HSGL is proposed to better address cases in which features can be naturally grouped, especially in bioinformatics and medical imaging. HSGL generalizes from sparse group lasso (SGL) and hierarchical lasso (HL) and is particularly successful in identifying a handful but yet important features from a large group of irrelevant features. In addition, HSGL is easy to interpret and implement, converges quickly and has better learning performance than other baseline learners when there are more features than observations.

2 - A Low Rank Model for Estimation of Response Functions in Multi-subject fMRI Data

Minh Pham, Rochester Institute of Technology, Rochester, NY, United Statesy, Tingting Zxhang, Jianhui Sun, Guofen Yan Ruizhon Miao, Huazhang Li, Sara Medina-Devilliers, James Coan

Functional magnetic resonance imaging (fMRI) data analysis faces several challenges, including extensive computation and di culty in obtaining statistically e cient estimates of the brain responses. We propose a new statistical model and computational algorithm to address these challenges. Specifically, we develop a new multi-subject, low-rank model within the general linear model framework for stimulus-evoked fMRI data. The new model assumes that the brain responses of di erent brain regions and subjects fall into a low-rank structure and can be represented by a few principal functional shapes. As such, the new model enables borrowing information across subjects and regions and increasing the ensuing estimation e ciency of brain responses, while accommodating the variation of brain activities across subjects, stimulus types, and regions. We propose two di erent optimization problems and a new fast-to-compute algorithm to address two research questions of broad interest in psychology studies: evaluating brain responses to di erent stimuli and identifying brain regions with di erent responses. Through both simulation and real data analysis, we show that the new method can outperform the existing methods by providing more e cient estimates of brain responses to designed stimuli.

3 - Collaborative Model-Agnostic Linear Hybrid Model

Tong Wang, University of Iowa, Iowa City, IA, 52245, United States, Qihang Lin

Interpretable machine learning models and black-box models are close competitors when it comes to deciding which approach to adopt. Interpretable models, such as linear models, have the benefit of being easytounderstandandexplain,whileblack-boxmodelssuchasdeepneuralnetworksoftenachievebetter predictive performance on complicated tasks. These two types of models have been distinctive choices for practitioners. In this work, we propose a Collaborative Model-Agnostic Linear Hybrid (COMALY) model that combines a linear model with any black-box model to make joint decisions. The COMALY model connects black-boxes and transparent-boxes into a continuum of hybrid models where the linear model determines part of the data while the black-box model determines the rest, creating a partition of the data space. As the partition changes, COMALY can smoothly transit from one extreme to the other, striking a balance between predictive accuracy and the explainability of the decision making process. We formulate the training of a COMALY model as a convex optimization, where predictive accuracy and explainability are balanced through objective function, and solve it with the accelerated proximal gradient method. We apply COMALY to structured data, text and image classification. Experiments show that COMALY can effectively trade prediction accuracy for explainability and provide an efficient frontier that spans the entire spectrum of explainability for users to choose.

4 - Knowledge Learning of Insurance Risks Using Dependence Models

Zifeng Zhao, Assistant Professor, University of Notre Dame, Notre Dame, IN, 46556, United States

In this paper, we explore the idea of using dependence models to learn the hidden risk of policyholders in property insurance. Specifically, we build a copula model to accommodate the dependence over time and over space among spatially clustered property risks. To tackle the computational challenge due to the discreteness feature of large-scale insurance data, we propose an efficient multilevel composite likelihood approach for parameter estimation. Provided that latent risk induces correlation, the proposed method allows one to borrow strength from related risks in prediction and further to study the relative importance of the multiple sources of unobserved heterogeneity in updating policyholders' risk profile.

Saturday, 5:45PM - 6:45PM

■ SD01

Hyatt, Cowboy Artists

Editors' Panel: How to Publish in Top Journals?

Panel Session

Chair: Ali Dag, University of South Dakota, Vermillion, SD, 57069, United States

1 - Editors' Panel: How to Publish in Top Journals?
Ali Dag, University of South Dakota, Vermillion, SD, 57069, United States

Top-tier premier outlets of the leading journals in the field of data mining and decision analytics would be represented by their respective editors. After a short presentation of the editors (5 minutes each), there will be a Q&A session, in which the audience would have chance to ask questions to the editors2

2 – Production and Operations Management
Subodha Kumar, Mays Business School, Texas A&M University, College Station, TX, 77843, United States

3 - Productions and Operations Management
Saravanan Kesavan, University of North Carolina-Chapel Hill, Chapel Hill, NC, 27599-3490, United States

4 - Management Science
Manel Baucells, University of Virginia, Charlottesville, VA, 22903, United States

5 - Journal of Operations Management
Suzanne de Treville, University of Lausanne, Lausanne-Dorigny, CH1015, Switzerland