Conference on Predictive Inference and Its Applications

May 7 and 8, 2018

Iowa State University

Poster Abstracts

Bayesian Regularisation from Stochastic Constraints

Joshua J Bon, Berwin Turlach, Kevin Murray, Christopher Drovandi
University of Western Australia, University of Western Australia, University of Western Australia, Queensland University of Technology

Regularisation in Bayesian modelling uses classes of priors which encourage shrinkage on posterior distributions. This property can be beneficial for sparse or underdetermined problems, and reduce overfitting (with benefits for prediction). In this paper, we propose a probabilistic interpretation for regularisation. We augment a given prior distribution with a stochastic constraint that probabilistically restricts the support of the prior, emitting a regularised prior distribution as a result. This introduces the notion of Bayesian regularisation as an operator that acts on a prior, rather than classes of priors which are considered to have desirable shrinkage properties. Among several advantages, regularisation from stochastic constraints allows multiple types of regularisation to be used and opens up new computational possibilities. The framework generalises some prominent priors in the literature including scale mixtures of normal distributions, such as the horseshoe prior [Carvalho et al., 2010]. We demonstrate the methodology on a forensic morphometric dataset.

A Novel Approach to Component-wise Gradient Boosting for High-Dimensional Linear Models

Brandon D. Butcher, Brian J. Smith
University of Iowa

Penalized regression methods, such as the lasso, are a robust statistical method for analyzing high-dimensional data that are commonplace in disciplines such as genetics and biomedical research. An alternative approach based on Freidman's gradient boosting machine, called component-wise gradient boosting, was developed by Buhlmann and Yu (2003) and Buhlmann (2006) for high-dimensional linear models. In general, gradient boosting with squared error loss is repeated fitting of least squares on the residuals. Component-wise gradient boosting adds an additional step where in each boosting iteration the predictor that provides that largest reduction in the residual sum of squares is selected prior to fitting the linear least squares base-learner on the current residual. As such, component-wise gradient boosting does variable selection and estimation in a similar manner as the LASSO, although step-wise coefficient paths are produced. We propose a modification to the component-wise gradient boosting algorithm which we refer to as iteratively re-estimated gradient boosting. Our modification uses all the currently selected predictors in fitting the linear least squares base-learner on the current residual, rather than only the selected predictor in the current boosting iteration. A high-dimensional simulation

study adapted from Hastie et al (2007) is conducted comparing the LASSO, component-wise gradient boosting, and iteratively re-estimated gradient boosting. Simulation results show that our modification provides better out of sample predictive performance and selects fewer spurious predictors. Additionally, our modification tended to require fewer boosting iterations than component-wise gradient boosting.

Bayes Bi-Clustering-Based Prediction

Abhishek Chakraborty, Steve Vardeman
Iowa State University

Ordinary clustering methods are used to form homogeneous groups in a set of objects. Bi-clustering methods are designed to take a rectangular dataset (rows for instances and columns for features) and simultaneously cluster rows and columns in such a way that responses/data values are homogeneous for all row-cluster by column-cluster groups. Potential applications abound. In marketing, it is desirable to simultaneously segment customers and products. In gene expression studies, it is desirable to simultaneously cluster genes and material samples. In agricultural yield studies, it is desirable to simultaneously group plant varieties and growing environments. In developing recommender systems, it is desirable to simultaneously group "raters" and "movies". We propose a Bayes methodology for bi-clustering that easily handles the commonly occurring situation where many of the entries in the data table are unobserved and provides important information beyond plausible sets of row-clusters and column-clusters. In particular, posterior probabilities that particular instances (or features) should be clustered together and predictors for unobserved values are available. Supposing that rows, i = 1,2,…,I, and columns, j = 1,2,…,J, of a data set are to be clustered into R groups of rows and C groups of columns, we suppose that r(i) belonging to {1,2,…,R} and c(j) belonging to {1,2,…,C} specify cluster assignments of data rows and columns respectively and for each (r,c) pair there is a mean data response μ-rc. The mean response for instance i and feature j is then μ-r(i), c(j). For Bayes modeling purposes, we use normal priors with zero mean and a fixed variance for the cluster means μ-rc. These normal priors are independent of the priors for the I-dimensional vector r = (r(1),r(2),…,r(I)) and the J-dimensional vector c = (c(1),c(2),…,c(J)), that have independent uniform marginals, on {1,2,…,R} and {1,2,…,C} respectively. To generate samples from the posterior and any parametric function of interest, we use Gibbs sampling for the cluster means, the vectors r and c and any of the unobserved values in the missing cells. Averages across iterations corresponding to the unobserved values give a plausible way to predict missing values in these kinds of data structures. Relative frequencies with which r(i) = r(i') give approximate posterior probabilities that instance i and instance i' belong in the same row-cluster (of course something exactly analogous is true for features and column-clusters).

Improving Cross-Validation for Penalized Cox Regression

Biyue Dai, Patrick Breheny
University of Iowa

Since its original proposal, penalized regression methods have been widely used for analyzing genetic or genomic data. Penalized Cox Regression Model allows researchers to build predictive models that link patients' survival outcomes with genetic profiles. While cross validation is a commonly used approach for selecting tuning parameters in penalized regression, little research has been done to study cross validation methods for penalized Cox regression. Due to its semi-parametric nature, conducting cross validation for Cox models has always been a methodological challenge and poorly understood in the

statistical community. We propose two new cross-validation methods for Cox Regression, and compare them to traditional information criteria as well as a cross-validated partial likelihood approach originally proposed by Verweij et al. Our simulation studies show that, in general, cross-validation tends to be conservative (i.e., select smaller models than the ideal choice of tuning parameters) for penalized Cox regression models. However, our proposed approach of cross-validating the linear predictors generally offers the best balance of stability and performance. We also illustrate these approaches on data from studies of gene expression and progression-free survival in cancer patients.

Optimal Calibration for Computer Model Prediction with Finite Samples

Xiaowu Dai, Peter Chien
University of Wisconsin-Madison

We consider a non-asymptotic frequentist framework for computer model prediction. This framework concerns two main issues: (1) many computer models are inadequate for physical systems and (2) only finite samples of physical observations are available for estimating model discrepancy and calibrating multivariate unknown parameters in computer models. We propose a method to achieve the optimal calibration and provide exact statistical guarantees in the sense that the predictive mean squared error is minimized with optimal calibration for any finite samples. We derive an equivalent formulation of optimal calibration which leads naturally to an iterative algorithm. The connection is built between the optimal calibration and the Bayesian calibration in Kennedy and O'Hagan [J. R. Stat. Soc. Ser. B. Stat. Methodol. 63 (2001) 425-464]. Numerical simulations and a real data example show that the proposed calibration outperforms the existing ones in terms of the prediction.

Improving Prediction Models Using Genomic and Image Data in Soybean

Reka Howard, Diego Jarquin, Alencar Xavier, Sruti Das Choudhury
University of Nebraska-Lincoln, University of Nebraska-Lincoln, Dow AgroSciences, University of Nebraska-Lincoln

Genomic prediction (GP) techniques became an important part of plant breeding programs due to their advantages compared to traditional phenotypic or pedigree based selections. GP is a technique where plants' genotypic and phenotypic information - called training set - are used to predict plant's phenotypic performance for which only marker information is available, also known as testing set. GP models can be extended to include high-throughput phenotypic information in the hope of increasing predictive ability. Herein we introduce two algorithms to predict a trait using marker and canopy information collected from 5600 recombinant inbred lines of a soybean nested association mapping panel. We introduce an algorithm that utilizes a hybrid matrix for the inclusion of marker and canopy information for the purpose of predicting the trait. We also extend the model to a multi-environmental case using an alternative of the reaction norm model that includes the interaction of the marker and canopy information through the utilization of the covariance structure.

Predictability of Hybrid Performance via Combining Ability models in interaction with environmental data

Diego Jarquin
University of Nebraska-Lincoln

The prediction of hybrid performance can enhance breeding programs by allowing the selection of the best set of parents for developing (potentially) the best progeny for a particular trait (yield, pest resistance, drought tolerance, etc.) in target environments. Genomic Selection is an emergent tool that aids the selection process by screening candidate hybrids based on the genetic profiles of the parents without having to plant the materials in fields. However, the assumption that the environmental conditions will not change from site to site or from year to another is not feasible in most of the cases. In general, the environmental conditions modulates the gene expression causing changes in the ranking of the performance of hybrids from one set of environmental conditions to another (presence of genotype-by-environment interaction GxE) complicating the labor of the breeders for selecting the parents for developing good hybrids. We propose a series of models that take advantage of the (i) general and specific combining abilities of the parents (inbreeds) and (ii) the GxE interaction via co-variance structures. Data from the genomes to field G2F project, which also includes weather information hourly recorded, was used for testing our proposed models. Four realistic scenarios that breeders face in fields were considered: CV2 prediction of incomplete field trials [tested hybrids in tested environments]; CV1 prediction of newly developed hybrids in tested environments; CV0 prediction of tested hybrids in new environments; CV00 prediction of new hybrids in new environments. The relative improvements of the proposed models compared with the traditional hybrid model in terms of predictive ability were: 75% [CV2], 95% [CV1], 57% [CV0] and 207% [CV00].

A Functional Anova Approach to Detecting Changes in Soil Moisture and Temperature

Manju M. Johny, Petruta C. Caragea, Diane M. Debinski, Jill A. Sherwood
Iowa State University

Climate change poses significant challenges to the soil ecosystem, with profound implications for many aspects of life. In addition, much of the data collected in the field pose interesting statistical challenges due to the presence of temporal dependence. Due to inherent dependencies in time series, many of the classical statistical inferential methods are inadequate. To overcome this difficulty, a functional anova approach may be utilized to identify difference in patterns between groups of time series curves. We present an application in which climate change was experimentally simulated in the montane meadows, resulting in time series measurements of soil moisture and temperature. After smoothing the discrete measurements to obtain functional curves, an anova test for functional data is performed through a parametric bootstrap procedure to test equality of mean curves between treatment groups. Extending this procedure, we develop a method to test for interaction between treatments. We also develop and illustrate novel visualizations of the tests, which not only provide another facet in understanding the significance of the tests, but also allow for identification of when significant differences occur over time.

A Non-Parametric Bayesian Change-Point Detection Method in the Recurrent-Event Context

Qing Li, Feng Guo, Inyoung Kim
University of Wisconsin-Madison, Virginia Tech, Virginia Tech

This study proposes a non-parametric Bayesian method to detect when the intensity rates changes significantly and clusters the individuals or sampling unit by their change-points in the recurrent-event context. We assume that the event counts are non-homogeneous Poisson process with piecewise-constant intensity functions. We propose a Dirichlet process mixture model allowing change-points to vary among sampling units, while the change-points are assigned a Dirichlet process prior. The intensity rates are subject specific. A Markov chain Monte Carlo algorithm is developed to sample from the posterior distributions. The advantage of our approach is that automatic clustering is achieved based on the change-points without specifying the number of latent clusters or model selection. We apply the methodology to both simulated data and the Naturalist Teenage Driving Study data. We also compare the model performance with the Bayesian finite mixture model in the simulation study. The simulation study suggests that the method is robust and flexible. It outperforms the Bayesian finite mixture model mainly in detecting the correct number of clusters, in assigning individuals to the correct cluster and in efficiency. The method is practically useful in many fields like transportation, medicine, reliability of products, and human behaviour.

Sparse Learning for Image-on-Scalar Regression with Application to Imaging Genetics Studies

Xinyi Li, Li Wang, GuanNan Wang
Iowa State University, Iowa State University, College of William & Mary

Motivated by recent advances in technology for brain imaging and high-throughput genotyping, we consider an imaging genetics approach to discover relationships between the interplay of genetic variation and environmental factors and measurements from imaging phenotypes. We propose an image-on-scalar regression method, in which the spatial heterogeneity of gene-environment interactions on imaging responses is investigated via an ultra-high-dimensional spatially varying coefficient model (SVCM). Bivariate splines on triangulations are used to represent the coefficient functions over an irregular two-dimensional domain of interest. When using the image-on-scalar regression method, a natural question raised in practice is if the coefficient function is really varying over space. In this paper, we present a unified approach for simultaneous sparse learning and model structure identification (i.e., varying and constant coefficients separation). Our method can identify zero, nonzero constant and spatially varying components correctly and efficiently. The estimators of constant coefficients and varying coefficient functions are consistent and asymptotically normal. The performance of the method is evaluated by a few simulation examples and a brain mapping study based on the Alzheimer's Disease Neuroimaging Initiative data.

Predictive Regionalization of Multi-Scale Air Pollutants Based on Functional Principal Component Analysis

Decai Liang
Peking University, Beijing

Emission patterns of air pollutants may vary across species and regions. To make more effective environment policies, we need do predictive inference on spatio-temporal pattern of meteorological factors and regional emission features. Conventional methods to regionalize air pollutants, such as Empirical Orthogonal Functions (EOF) or Self Organizing Maps (SOM), have limitations in combing multiple air pollutants with their spatial correlation structures. We propose a new method based on Functional Principal Component Analysis (FPCA), in which the associations among different pollutants and monitor sites are treated as correlations between functional PC scores. The regionalization is realized by a likelihood-based clustering. We further extend our procedure to the heteroscedastic case where different clusters have different covariance structures and use a Monte Carlo EM algorithm for parameter estimation. The effectiveness of our method is illustrated by simulation studies. We apply our method to analyze 6 main pollutants over the North China Plain (NCP) from 2013 to 2016, and our findings indicate that different parts of NCP need different policies to control the air pollution.

Matrix Completion with Covariate Information

Xiaojun Mao, Song Xi Chen, Raymond Wong
Iowa State University, Peking University, Texas A&M University

The poster abstract: This paper investigates the problem of matrix completion from corrupted data when additional covariates are available. Despite being seldom considered in the matrix completion literature, these covariates often provide valuable information for completing the unobserved entries of the high-dimensional target matrix A0. Given a covariate matrix X with its rows representing the row covariates of A0, we consider a column-space-decomposition model A0=X\beta_0+B0 where \beta_0 is a coefficient matrix and B0 is a low-rank matrix orthogonal to X in terms of column space. This model facilitates a clear separation between the interpretable covariate effects (X\beta_0) and the flexible hidden factor effects (B0). Besides, our work allows the probabilities of observation to depend on the covariate matrix, and hence a missing-at-random mechanism is permitted. We propose a novel penalized estimator for A0 by utilizing both Frobenius-norm and nuclear-norm regularizations with an efficient and scalable algorithm. Asymptotic convergence rates of the proposed estimators are studied. The empirical performance of the proposed methodology is illustrated via both numerical experiments and a real data application.

Order-of-Addition Modeling

Robert W. Mee
Nankai University and University of Tennessee

Since Van Nostrand (1995), the literature on order of addition experiments has generally relied on main effects models constructed from pair-wise ordering (PWO) factors. This article constructs models utilizing interactions of PWO factors to explain variation that is best accounted for by the ordering of sets of three or more components. Order-of-addition orthogonal arrays, as defined by Voelkel (2018), are optimal for fitting the main effects PWO model, but they differ in terms of their susceptibility to bias

due to model misspecification. A measure computed from the alias matrix is proposed to identify robust PWO designs and illustrated for cases with four and five components. Consideration of applications with constraints on the ordering, as well order of addition experiments having additional mixture proportion and factorial factors, are discussed briefly. Two drug sequence experiments obtained by private consulting are used to illustrate the usefulness of different PWO models.

RNA-seq Differential Expression Analysis Accounting for Relevant Covariates

Yet Nguyen, Dan Nettleton
Iowa State University

RNA-sequencing (RNA-seq) technology enables the detection of differentially expressed genes, i.e., genes whose mean transcript abundance levels vary across conditions. In practice, an RNA-seq dataset often contains some explanatory variables that will be included in analysis with certainty in addition to a set of covariates that are subject to selection. Some covariates may be relevant to gene expression levels, while others may be irrelevant. Either ignoring relevant covariates or attempting to adjust for the effect of irrelevant covariates can be detrimental to identifying differentially expressed genes. We address this issue by proposing a covariate selection method using pseudo-covariates to control the expected proportion of selected covariates that are irrelevant. We show that the proposed method can accurately choose the most relevant covariates while holding the false selection rate below a specified level. We also show that our method outperforms methods for detecting differentially expressed genes that do not take covariate selection into account, or methods that use surrogate variables instead of the available covariates.

Prediction-guided Inference in Statistical and Machine-Learning Models for Parkinson's Disease Progression

Ryan A. Peterson, Brandon Butcher, Stephanie Lussier, Brian Smith
University of Iowa

We demonstrate how prediction intervals based on cross-validation and interactive effect plots can be utilized to communicate the results of a robust set of both statistical and machine-learning models for disease progression in Parkinson's Disease (PKD). We also describe a simple two-stage modeling approach that can mediate the issue of correlated (and noisy) outcome data. PKD is a complicated disease, and the issue of properly defining and predicting progression in PKD is an open problem. As a part of a data challenge, we developed a series of regression and machine-learning models that aim to describe important factors (and predict) the progression of PKD. In the first stage, we measured progression as the slope of the outcome variable for each individual over the study's course. In the second stage, we built models to predict this slope using a wide array of predictive modeling frameworks, selecting a final model based on extra-sample predictive performance. Finally, we describe the model in a web application that provides a predictor and a prediction interval for a given set of covariate values. Users are able to input their own covariate values, and they are guided to start by inputting the covariates that have been judged to be most important in the model (allowing for imputation of the rest via K-nearest-neighbors). The web application also shows important model information, including interactive effect plots, relative variable importance measures, and model diagnostics. The prediction intervals assume only that the data are randomly distributed from the predictive population, and are therefore applicable to a broad class of modeling frameworks.

Density Estimation with Small Berkson Errors

Ramchandra Rimal, Marianna Pensky
University of Central Florida

Consider a sample of independent and identically distributed observations of variables Y=X+E where measurements of X and E are unavailable, the probability density function of X is unknown but the probability density function of E is known. The objective is to estimate the probability density function of W = X+Z, where Z has a known probability distribution, on the basis of the observations of Y.

The problem is known as density estimation with Berkson errors and has applications in econometrics, astronomy, biometrics, medical statistics, and image reconstruction. It is known that presence of the Berkson errors improves the accuracy of estimation of the desired density function. The focus of our work is an investigation of the case when Berkson errors have variances which may be very small. We construct kernel estimators of the density function of W and obtain the upper bounds of the risk. We further conclude in which cases the kernel estimator is necessary for improvement of the error bounds.

A Robust Residual-Based Approach to Random Forest Regression

Andrew Sage, Ulrike Genschel, Dan Nettleton
Iowa State University

We introduce a novel robust approach for random forest regression that is useful when the conditional distribution of the response variable, given predictor values, is contaminated. Residual analysis is used to identify unusual response values in training data, and the contributions of these values are down-weighted accordingly. This approach is motivated by the robust fitting procedure first proposed in the context of locally weighted polynomial regression and scatterplot smoothing (Cleveland, 1979). We further demonstrate that tuning the parameter in the robustness algorithm using a weighted cross-validation approach is advantageous when contamination is suspected in training data. We conduct extensive simulations, comparing our method to existing robust approaches, some of which have not been compared to one another in prior studies. Our approach outperforms existing techniques on noisy training datasets with response contamination. While no approach is uniformly optimal, ours is consistently competitive with the best existing approaches for robust random forest regression.

Multi-level variable screening and selection for survival data

Natasha Sahr, Kwang Woo Ahn, Soyoung Kim
Medical College of Wisconsin

Variable selection methods for the marginal proportional hazards model is a relatively understudied research area in biostatistics. The limited available methods focus on the selection of non-zero individual variables. However, variable selection in the presence of grouped covariates is often required. Some methods are available for the selection of non-zero group and within-group variables for the univariate proportional hazards model. There are no available methods to perform group variable selection in the clustered multivariate survival setting. In this context, the hierarchical adaptive group bridge penalty is proposed to select non-zero group and within-group variables for the marginal proportional hazards model with independent or clustered multivariate failure time data. The simultaneous selection of non-zero group and within-group variables for multivariate modeling is defined as multi-level selection.

Simulation studies show the hierarchical adaptive group bridge method has superior performance compared to the extension of the adaptive group bridge in terms of variable selection accuracy. Sometimes, survival data suffers from high-dimensional group variables. Most existing screening methods address the sure screening property for individual variable selection. The sure group joint variable screening method is proposed to screen independent and clustered multivariate survival data in the presence of group variables. Simulation studies show the sure group joint variable screening method performs better than existing screening procedures extended to the multivariate survival setting. The hierarchical adaptive group bridge and sure group joint variable screening methods can be effective tools, used in a two-step process, in identifying non-zero group and within-group variables for high-dimensional multivariate survival data.

Prediction of Warranty Returns Based on Modeling Seasonal Recurrent Event Data

Qianqian Shan, William Meeker
Iowa State University

Warranty return data from repairable systems, such as vehicles, result in recurrent event data. The non-homogeneous Poisson process (NHPP) model is used widely to describe such data. Seasonality, however, complicates the modeling of recurrent-event data. This paper provides a general approach for the application of NHPP models to predict warranty returns. A hierarchical clustering method is used to stratify the population into groups that are more homogeneous than the overall population. The stratification facilitates modeling the recurrent-event data with both time-varying and time-constant covariates. We demonstrate and validate the models using vehicle warranty claims data. The results show that our approach provides significant improvements in predictive power.

A new method for constructing the population-specific gene regulatory networks based on the meta-LASSO regression model

Jiang Shu, Bruno Vieira Resende e Silva, Tian Gao, Zheng Xu, Juan Cui
University of Nebraska-Lincoln

MicroRNA is responsible for the fine-tuning of fundamental cellular activities and human disease development. The altered availability of microRNAs, target mRNAs, and other types of endogenous RNAs competing for microRNA interactions reflects the dynamic and conditional property of microRNA-mediated gene regulation that remains under-investigated. Here we propose a new integrative method to study this dynamic process by considering both competing and cooperative mechanisms and identifying functional modules where different microRNAs co-regulate the same functional process. Specifically, a new pipeline was built based on a meta-Lasso regression model and the proof-of-concept study was performed using a large-scale genomic dataset from ~4,200 patients with 9 cancer types. In the analysis, 10,726 microRNA-mRNA interactions were identified to be associated with a specific stage and/or type of cancer, which demonstrated the dynamic and conditional miRNA regulation during cancer progression. On the other hands, we detected 4,134 regulatory modules that exhibit high fidelity of microRNA function through selective microRNA-mRNA binding and modulation. For example, miR-18a-3p, −320a, −193b-3p, and −92b-3p co-regulate the glycolysis/gluconeogenesis and focal adhesion in cancers of kidney, liver, lung, and uterus. Furthermore, several new insights into dynamic microRNA regulation in cancers have been discovered in this study.

Machine Learning Approaches for High-Throughput Plant Phenotyping Based on Hyper-Spectral Images

Zheng Xu, Zheng Xu, Piyush Pandey, Yufeng Ge
University of Nebraska-Lincoln

Image-based high-throughput plant phenotyping has relieved the bottleneck of traditional phenotyping which limits the downstream studies integrating genotypes, treatments and phenotypes. RGB-imaging and its corresponding phenotyping have shown success in extracting morphological traits and tracking plant growth. Hyper-spectral imaging and its corresponding phenotyping reveal traits beyond plant morphology and growth. One utility of hyperspectral imaging is its ability to quantify chemical properties of maize and soybean plants in vivo. Statistical approaches using principal component analysis (PCA) and partial least square (PLS) have found success. This study aims to found improved approaches over the current PCA+PLS approach (Pandey et al 2017). Hyperspectral images were collected from 60 maize and 60 soybean. Plants with the concentrations of elements including nitrogen (N), phosphorus (P) and potassium (K) have been experimentally measured. We found improvement in prediction accuracy using the combination of approaches including PCA, LASSO and PLS.

Generalized Additive Geo-Spatial Models

Shan Yu, Lily Wang, GuanNan Wang
Iowa State University, Iowa State University, College of William & Mary

In many application areas, data are collected on a count or binary response with spatial covariate information. In this paper, we introduce a new class of generalized additive models (GAMs) for spatial data distributed over complex domains. Through a link function, the proposed GAM assumes that the mean of the discrete response variable depends on additive univariate functions of explanatory variables and a bivariate function to adjust for the spatial effect. We propose a two-stage approach for estimating and making the inference of the components in the GAM. In the first stage, we approximate each of the univariate additive components in the model via univariate polynomial splines. The bivariate component is approximated using bivariate penalized splines over triangulation, which is proved to be efficient to deal with spatial data distributed on irregular domains with complicated boundaries. In the second stage, local polynomial smoothing is then applied to the cleaned univariate data average out the variation of the first-stage estimators. We investigate the consistency of the proposed estimators of the component functions and the asymptotic normality of the univariate components. We also establish the simultaneous confidence bands of the univariate components. The performance of the method is evaluated by simulation studies. We apply the proposed method to the crash counts data in the Tampa-St. Petersburg urbanized area in Florida.

Random Forest Prediction Intervals

Haozhe Zhang, Dan Nettleton, Dan Nordman
Iowa State University

Random forest methodology (Breiman, 2001) is one of the most popular machine learning techniques for prediction problems. An important but often overlooked challenge is the determination of prediction intervals that contain unobserved response values with specified probabilities. In this article, we proposes a method for constructing prediction intervals from a single random forest and its byproducts. Under certain regularity conditions, we prove that the proposed intervals have asymptotically correct

coverage rates. The finite-sample properties of the proposed intervals are compared via simulation with two existing approaches: quantile regression forests (Meinshausen, 2006) and split conformal intervals (Lei et al., 2017). The effects of tuning parameters on prediction interval performance are also explored in the simulation study. In addition, we analyzed 67 datasets from the UCI machine learning repository and Chipman et al. (2010). The numeric results demonstrate that intervals constructed with our proposed method have smaller interval widths than split conformal intervals, while both our intervals and split conformal intervals have more accurate and more robust marginal coverage rates than quantile regression forest intervals.

The Conference on Predictive Inference and Its Applications is supported by the National Science Foundation under Grant No. 1810945. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.