Session: Big Data Modeling and Applications

Poisson Graphical Models with Rich Dependence Structures
P. Ravikumar

Abstract: Undirected graphical models, such as Gaussian, Ising, and discrete/multinomial graphical models, are widely used in a variety of applications for modeling distributions over a large number of variables. These standard instances, however, are ill-suited to modeling count data, which are increasingly ubiquitous in big-data settings such as genomic sequencing data, user-ratings data, spatial incidence data, climate studies, and site visits. Existing proposals for distributions for multivariate count data have a crucial caveat: the dependence structures they model are largely restrictive, with solely negative or positive dependencies in some cases.

Can we devise multivariate distributions that can capture rich dependence structures between count-valued variables? We address this question via a series of multivariate extensions of the univariate Poisson distribution, providing a new class of Poisson graphical models. We also provide tractable schemes with guarantees for learning our class of Poisson graphical models from data, and demonstrate the performance of our methods by learning simulated networks as well as a network from microRNA-Sequencing data.

Joint work with Eunho Yang, Genevera Allen, Zhandong Liu, David Inouye, Inderjit Dhillon.

Computationally Efficient Nonparametric Testing
Guang Cheng

A recent trend of big data problems is to develop computationally efficient inferences that embed computational thinking into traditional uncertainty quantification methods. A particular focus of this talk involves two new classes of nonparametric testing that scales well with massive data. One class is based on randomized sketches which can be implemented in one computer, while another class requires parallel computing. Besides introducing these two new methods, our theoretical contribution is to characterize the minimal computational cost that is needed to achieve the minimax optimal testing power.

Big Data Applications in Petroleum Exploration and Production
Srikanta Mishra, Battelle Memorial Institute

Session: Big Data Computing

Decoupled System Architecture and Computation Model for Scientific Big Data Applications

Yong Chen, Texas Tech University

The data-driven scientific discovery presents a critical question to the research community - how to efficiently support these increasingly important scientific big data applications with high performance computing (HPC) systems that are traditionally designed for big compute applications? The conventional systems are computing-centric and designed for computation-intensive applications. Scientific big data applications have growlingly different characteristics compared to big compute applications. These scientific applications, however, will still largely rely on HPC systems to be solved. In this talk, I will introduce our research efforts in attempting to answer this question with studying a decoupled system architecture and a computation model for big data applications. In one project, we are extending the software stack to address the data-driven scientific discovery needs. In another project, we are deploying a system prototype called Data Intensive Scalable Computing Instrument focusing on big data problems. I will present our current findings and discuss open questions as well. More information about these research efforts can be found from: http://discl.cs.ttu.edu/.

Session: Big Data Industry

Big Data Industrial Analytics
Aldo Dagnino, ABB

In this talk, I will discuss a particular case that we have developed at ABB in the field of big data industrial analytics, the environment we utilized, and the lessons learned in the development process.

Big Data, Does it Deliver or Is It Hot Air?
Brian Hayes, EDP Renewables

In recent years, wind energy has quickly become a relevant part of the US power generation mix with expectations to be at 10% energy production by 2020. Many European countries have already exceeded a 20% wind generation threshold. With wind energy relevant and the tens of thousands of turbines operating around the world, managing this sizeable fleet has become of paramount importance to ensure wind power plants are performing as efficiently as possible.

Each wind turbine can have 400 or more operational data tags that spin off every second to 10 minutes plus additional data related to maintenance, weather conditions, technician efficiency, part usage, etc. Initiatives that combine and analyze all of this Big Data for wind energy have increased significantly over the last 5 years. Utilizing this data to create information that drives clear actions has the potential to yield benefits both in increased production and reduced operational cost. This discussion will provide a case study of EDP Renewables and their use and future plans with Big Data, so you can decide if the hype around Big Data is real or Hot Air.

Big Data Analysis of High Frequency SCADA Data and High Resolution Image Data in the Field of Wind Energy
Georgios Pechlivanoglou, Smart Blade, Germany

The talk will focus on the use of high frequency SCADA data with various statistical methods in order to evaluate wind turbine performance and identify O&M inefficiencies. Furthermore the use of large volumes of decomposed high resolution images for the evaluation of the aerodynamic performance of rotorblades will also be presented. The overall aim of the presentation is to demonstrate ways where Big Data can be effectively use in the field of wind energy.