Indiana University-Purdue University Indianapolis
Department of Mathematical Sciences

Statistics Seminar

12:15pm—1:15pm, Tuesday, Feb. 20, 2018
SL 137

Speaker: Dr. Mohammad Al Hasan (Associate Professor)
Department of Computer Science, IUPUI

Title: Online Name Disambiguation on Non-Exhaustiveness training data using Bayesian framework

Abstract:

All individuals are unique, but millions of people share names. How to distinguish among, or as it is technically known, disambiguate, people with the same names is an important real-life problem. Solution to this problem is needed in a wide range of environments, in case of law enforcement — which Mohammad Hasan is attempting to board an airplane flight? or in bibliographic — which W. Wang is the author of a research study? Effectively solving name disambiguation requires collecting features and then learning a classification model from training data. But, the training data for the classification model cannot include examples from each and every individual who the model needs to disambiguate at a later time, resulting poor prediction for individual who the model has not yet acquainted with.

In this talk, I will discuss some of our recent works for solving online name disambiguation by using a Bayesian classification framework. Our proposed method uses a Dirichlet process prior with a Normal x Normal x Inverse Wishart data model which enables identification of new ambiguous entities who have no records in the training data, thus addressing the problem of non-exhaustive classification scenarios. For probabilistic inference, we use simple method, such as, one-pass Gibbs sampling and also more advanced sequential Monte Carlo method, such as, particle filter. As a case study we consider bibliographic data in a temporal stream format and disambiguate authors by partitioning their papers into homogeneous groups. Our experimental results demonstrate that the proposed method is better than existing methods for performing online name disambiguation task.