Bingyu Wang

(617) 750-6151 • rainicy@ccs.neu.edu • bingyuwang.net
36 Highcrest Ter., Boston, MA, 02131

Education

Northeastern University *Candidate for PhD of Computer Science, GPA: 3.85/4.00* *2015-Present*
Northeastern University *Master of Computer Science, GPA: 3.85/4.00* *2012 - 2014*
Northwest University *BE in Software Engineering, GPA: 3.30/4.00* *2008 - 2012*

Research Experience

Survival Analysis to Assess Association between Mortality and Air Pollution *2016-Present*

- Scalable implementation for survival linear models, such as Cox Proportional Hazards and Poisson Regression, to handle over 64 million US Medicare enrollees from 2000 to 2012, up to 2 billion enrollee-months of follow-up. The running time is around 10 minutes on one machine (Intel Xeon CPU E5-2680 with 56 logical cores).
- Efficient implementation for a non-linear Cox Proportional Hazards model with restricted cubic splines, in order to have a more flexible fitting between predictors and outcome.
- Conducted big data study to assess the association between long-term exposures, like $PM_{2.5}$, NO_2, O_3 etc., and cause-specific mortality, including cardiovascular, respiratory, cancer etc.

Extreme Multi-Label Classification(XCBM) *2017-Present*

- Developed a sparse CBM(XCBM) by exploring feature sparsity, label sparsity and label imbalance.
- Derived and implemented a Weighted Dual Coordinate Descent method to speed up training.
- XCBM achieved a comparable performance comparing with other extreme classifiers.

Regularizing Model and Label Structure for Multi-Label Classification *2016-2017*

- Regularized Multi-Label classifiers by ElasticNet to avoid overfitting and shrink model size.
- Combined General F-Measure Maximizer(GFM) with Support Inferences to obtain optimal instance-F1 prediction.
- Achieved better instance-F1 comparing with existing Multi-Label methods.

Conditional Bernoulli Mixtures(CBM) for Multi-label Classification *2015-2016*

- Derived and implemented a new Multi-label classifier using Mixtures of Bernoulli.
- Developed an efficient inference to make joint prediction by dynamic programming.
- CBM outperformed other state-of-the-art Multi-Label methods.

Topic-Factorized Ideal Point Estimation Model for Legislative Voting Network *2013-2014*

- Crawled Roll Call Votes data and built the dictionaries for Bill Text, Voting records and legislators.
- Implemented the topic models on bill texts, like Probabilistic latent semantic analysis (PLSA), latent Dirichlet allocation (LDA) for the baseline.
- Visualized the voters' ideological positions on website, using D3js.

Professional Experience

JD.Com Inc, Mountain View, CA *May-Aug 2018*
Data Scientist Intern (PyTorch)

- Proposed a scalable deep learning model for the extreme multi-label classification problems.
- Recognized the pattern of products data (descriptions and categories) from the electronic business platform, and improved the performance on predicting the product categories.
- Built a semi-automatic data annotation tool using the multi-label classifier, to enable more productive data labeling.

MassMutual Financial Group, Boston, MA *Jan-June 2014*
Data Analyst (Python)

- Recognized the pattern and performed analysis and predictions on the web log data of Oppenheimer Website using the Aster Express Tool from Teradata.
- Analyzed the MassMutual HR data and produced the predictions on the Quality of Hire.
- Performed twitter analysis for GeoAnalytics project to collect data from twitter using sentimental keywords and find out the areas where MassMutual can promote the sales.

Federal Home Loan Bank, Boston, MA *June-Aug 2013*
Information Technology Intern (Java)

- Developed a Test Automation Framework that can easily be used to test different web based projects using various technologies, such as Selenium, Open2Test and Eclipse.
- Delivered documentations, including test script based on SharePoint, test results covering test suite execution and screenshot, and user manual for non-computer staff.

Publications & Conferences

- **Wang**, Li, Sun, Qin, Li, Zhou, *"Ranking-based AutoEncoder for Extreme Multi-label Classification"*, *NAACL 2019*;
- Kazemiparkouhi, Eum, **Wang**, Manjourides, Suh, *"Long-Term Ozone Exposures and Cause-specific Mortality in a US Medicare Cohort"*, *JESEE 2019*;
- Eum, Kazemiparkouhi, **Wang**, Manjourides, Pun, Pavlu, Suh, *"Long-Term NO2 Exposures and Cause-specific Mortality in American Older Adults"*, *Environment International 2019*;
- **Wang**, Li, Pavlu, Aslam, *"A Pipeline for Optimizing F1-Measure in Multi-Label Text Classification"*, *ICMLA 2018*;
- Kazemiparkouhi, Eum, **Wang**, Manjourides, Suh, *"Effect of Confounding, Effect Modification, and Exposure Measures on the Association of Long-Term Ozone Exposure and Cause-Specific Mortality"*, *ISES-ISEE 2018*;
- **Wang**, Eum, Manjourides, Kazemiparkouhi, Pavlu, Suh, *"Effect Modification of the Association of Long-Term PM2.5 Exposure and Cause-Specific Mortality: An Analysis of 64 Million US Medicare Beneficiaries"*, *ISES-ISEE 2018*;
- Eum, **Wang**, Manjourides, Kazemiparkouhi, Pun, Pavlu, Suh, *"Effect of Confounding, Long-Term NO2 Exposures and Cause-specific Mortality in American Older Adults"*, *ISES-ISEE 2018*;
- Gu, Chen, Sun, **Wang**, *"Ideology Detection for Twitter Users via Link Analysis"*, *SBP-BRiMS 2017*;
- Li, **Wang**, Pavlu, Aslam, *"Conditional Bernoulli Mixtures for Multi-Label Classification"*, *ICML 2016*;
- Li, **Wang**, Pavlu, Aslam, *"An Empirical Study of Skip-gram Features and Regularization for Learning on Sentiment Analysis"*, *ECIR 2016*;
- Gu, Sun, Jiang, **Wang**, Chen, *"Topic-Factorized Ideal Point Estimation Model for Legislative Voting Network"*, *KDD 2014*.

Computer Knowledge

Machine Learning, Survival Analysis, Data Mining, Java, Python, PyTorch, R, MATLAB, LaTex, D3js