2018 Summer Undergraduate Research Experience Project Showcase

University of California, Santa Barbara
August 16, 2018

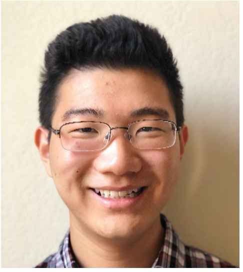
Funded by the National Science Foundation via an IGERT in Network Science (DGE-1258507)

The Moral Framing of Human Rights Reports: An Exploratory Data Analysis of the Human Rights Global Knowledge Graph

Aaron Barel, Statistics, UC Santa Barbara
Frederic Hopp, Rene Weber, Communication

The purpose of this study is to explore the Global Database of Events, Language, and Tone's (GDELT; Leetaru & Schrodt, 2014) Human Rights Global Knowledge Graph, a collection of over 100,000 computationally analyzed human rights reports from seven human rights organizations (e.g., Human Rights Watch, Amnesty International). The database contains numerous, automatically-extracted textual features, such as sentiment, themes (e.g., types of human rights violations), people, locations, and organizations. In addition to GDELT's provided features, we aim to scrape the full-text of each report to extract additional features pertaining to the framing of human rights violations, specifically moral language and the motivational relevance of this language. In turn, we aim to construct a supervised machine learning model that links variations in these language features to the attention these reports garner on social media websites, as well as the financial donations made to these organizations.

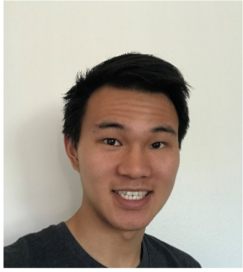
Deep Probabilistic Forecasting and Data Mining Granger Causality Networks in Time Series



Andy Jin, Computer Science, Stanford University
Xiaoyong Jin, Xifeng Yan, Computer Science

Probabilistic forecasting is a foundational technique for accelerating and optimizing decision-making processes in many industries. By learning from past observations—the conditioning range—a time series model is able to recognize seasonal fluctuations and prevalent patterns in the data, thereby enhancing forecasting accuracy for the prediction range. In this project, we utilize Autoregressive Recurrent Neural Networks (RNN) to achieve this goal. Specifically, we analyze the performance of an Autoregressive RNN in predicting hourly electricity consumption of Portuguese households. We conduct two rounds of analysis, one accounting for covariates that pertain to the hourly timestamp, and one that does not. Comparing our results to current state-of-the-art methods, we show that the autoregressive RNN produces more accurate forecasts, a phenomenon that can be attributed to the capacity for its Long Short-Term Memory (LSTM) cells to capture and retain past history. To deepen insights, we employ data mining on Granger Causality networks to analyze how one time series may be useful in forecasting another, hence allowing us to uncover interrelationships between multiple time series. In particular, we study the performances of the RNN and Multilayer Perceptron (MLP) that feature sparsity-inducing weight penalties in forecasting nonlinear time series. We test our methods on the DREAM3 dataset, which contains nonlinear time series networks for gene expression.

Database Query Prediction using Arrival Rate History



Lawrence Lim, Computer Science, UC Santa Barbara
Vivek Kulkarni, Divy Agrawal, William Wang, Computer Science

Forecasting future queries to a database provides substantial benefits, such as proper allocation of computing power and memory resources as well as faster responses to queries. Adequate forecasting is achievable because queries exhibit significant temporal patterns detectable by classical machine learning algorithms like recurrent neural networks (RNN), regression algorithms, and matrix factorization. Typical time series prediction problems forecast only one dependent variable. In database query prediction, however, where we are interested in forecasting a larger set of queries, both independent and dependent variables can be high-dimensional. Database query prediction occurs in many other common contexts, such as predicting the sales of multiple different products, or predicting the energy consumption of many households. We propose a method based on matrix factorization to forecast future queries from their arrival rate history. We comprehensively evaluate our method on three datasets (Electricity, E-commerce, and Google trends) and demonstrate that our method improves over other baseline models by up to 20%. We also discuss when it may be better to explore other options.

Molecular Fingerprinting on Small Datasets

Lucas Tong, Computer Science, UC Santa Barbara
Sourav Medya, Ambuj Singh, Computer Science

Precise molecular fingerprinting allows us to make accurate predictions regarding the properties of unknown molecules. Molecular Fingerprints encode important molecular structures in the form of fixed length integers where molecules with similar structures have similar fingerprints. Recent developments in the field have shown that learning fingerprints using modified CNN convolutions result in remarkable improvements in molecular prediction accuracy. This project explores the possibility of using svms and random forests to predict molecular properties in environments with constrained or unbalanced training datasets. Our implementation is evaluated against similar fingerprinting models that use neural networks as well as previously trained svm and regression implementations that don't use molecular fingerprinting. We compare our results against these models using the Directory of Useful Decoys (DUD) dataset - a dataset of molecules designed to contain a very high percentage of decoys that are difficult to distinguish from the active molecules. For each active molecule, 36 decoys with similar properties are added to the dataset. In our experiments, we found remarkable improvements when predicting against most targets using the random forest implementation. We hypothesize that our new svm and random forest model implementations are more accurate than the previous attempts due to reduced overfitting.