Chenyang Shi

Website: https://olivia-shi.github.io
Phone: (+86) 18402881015
E-mail: olivia.c.shi@gmail.com

EDUCATION & ACADEMIC VISITING

Southwest Jiaotong University, Chengdu, China *09/2015-06/2019*

Bachelor of Science, Computer Science *GPA: 3.60/4.0(Major 3.72/4.0)*

COURSEWORK & SKILLS

Programming Languages: Python, R, Java, C, C++, Scala.
Software & Tools: Jupyter, Gephi, LaTeX, Spark GraphX.

RESEARCH EXPERIENCES & INTERSHIPS

Full-time Research Assistant (visting student) *12/2018-06/2019*

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

Traffic Prediction based on Tensor Factorization (in process)

Part-time Research Assistant *07/2018-09/2018*

Southwest Jiaotong University, Chengdu, China

Investigation on the Kaggle Competition Home Credit Default Risk based on Complex Modelling Analysis

■ **Exploratory Data Processing and Feature Capture**

- Checked missing values and visualized data distribution pattern; Performed imputation through replacing certain positive and negative INF values with NAN for Neutral Network (NN).
- Followed the former preprocessing to do data feature extraction; Facilitated the utilization of feature variables concerning one hot encoding.
- Applied Lightgbm model to refine the summarized features regarding their contributions to the model.

■ **Feature Engineering**

- Original Features: Obtained new features based on data computing integration.
- Statistical Features: Enhanced the numerical variables given the categorical variables to summarize statistics.
- Timing characteristics: Fixed time window and number of times to construct new features and combined summarized statistics.

■ **Single Model Optimization**

- Dielectric argumentation on applicable models:
  - The inborn missing data in the data prediction negatively affected the accuracy of Neutral Network (NN);
  - Catboost was promising for dataset containing multiple variables, but the running duration was too long.
  - Lightgbm and Xgboost could possibly accommodate the prediction due to they can automatically impute the missing values regarding the reduction on training loss during the learning process.

- Performed Bayesian Optimization to extend the optimal combination of hyper-parameters.

■ **Stacking Model Optimization**

- Performed 10 fold cross-validation to obtain out-of-fold prediction values as training and testing data on Logistic Regression model.
- Strengthened the data optimization based on the stacking model integrating Lightgbm and Xgboost model.

■ **Results:**

- The optimized results indicated that the Area Under Curve (AUC) evaluation on predicting performance reached 0.7979; Achieved a brown medal.
- The abilities of loan repayment depended on the previous applications requested to the credit bureau.

Part-time Research Assistant *05/2018*

Southwest Jiaotong University, Chengdu, China ***Advisor: Prof. Xiao Wu***

Research on Internet Search Engine via Two Dimensions: Automatic Downloading and Text Classification

■ **PART ONE: Automatic Downloading Process for English & Chinese Webpages**

- Applied Beautiful Soup Database in Python to extract and condense the content of HTML webpages on Jupyter and then selected the useful information.
- Deleted the Stop Words in English via Natural Language Toolkit (NLTK) to facilitate the Porter Stemming and saved the output of simplified TXT.
- Performed Porter Stemming for Stop Words in Chinese in the similar way as those in English and saved the output of simplified TXT.
- **Achievements:**
  - All the steps were successfully integrated into a complete automatic system.
  - Extended the chart for common Stop Words to generate adaptable chart for specific Stop Words.

■ **PART TWO: Text Classification Based on the Similarity of Different Texts**

- **Mechanism for Similarity Comparisons:**
  - Set the vector of the words frequency given the times of word appearance.
  - Calculated the similarity of words regarding the cosine distance of their vectors.
- **Mechanism for Text Classification:**
  - Transferred the text to matrix of word frequency through Term Frequency-Inverse Document Frequency (TF-IDF).
  - Performed class analysis via K-means algorithms and optimized the value of k based on Elbow Rule to divide the data into two categories.

Part-time Research Assistant *05/2017-04/2018*

Southwest Jiaotong University, Chengdu, China ***Advisor: Prof. Hongmei Chen***

Exploration on Overlapping Community Detection and Its Application on Graph Data Mining

- Preprocessed network dataset by extracting data information file into node and edge files.
- Applied small data clusters from Stanford Large Network Dataset (SNAP) to do visualized community structure analysis for Label-based Propagation algorithms (LPA), Clique Percolation Method algorithms (CPM) verification with Gephi.
- Optimized CPM with Weak Clique Percolation Method(WCPM) replacement to decrease the time complexity of the algorithm through transmitting exponential relationship to linear relationship.
- Improved a Label Propagation Algorithm for better mining performance and application on overlapping community.

HONORS & AWARDS

Students Scholarship for five times. *2016-2018*

Excellent volunteer in Life Mystery Museum at Chengdu city. *05/2016*

Third prize, China Undergraduate Mathematical Contest in Modeling. *05/2016*

EXTRACURRICULAR ACTIVITIES

A member of debating team in school of Life Science.

Volunteer in Life Mystery Museum at Chengdu city.

A member of Mathematical Modeling Institute in Southwest Jiaotong University.