史晨阳

个人网站: http://olivia-shi.github.io

电话: (+86) 18402881015

E-mail: olivia.c.shi@gmail.com

教育经历&学术访问

西南交通大学，成都，中国 *09/2015-06/2019*

计算机科学与技术，工科学士 *GPA: 3.60/4.0*

中国科学院深圳先进技术研究院（SIAT） *01/2019-05/2019*

前瞻中心，无限能量与传输实验室，客座学生

主要技能/课程

编程语言（熟悉）：Python，C，C++，Java

编程语言（了解）：R，HTML, JavaScript，Scala，Matlab

工具/框架：NodeJS，Jupyter，Gephi，LaTeX，Spark

英语等级：CET 六级 558，托福 94，GRE 320

相关经历&实习

Udemy 网课：The Web Developer Bootcamp – YelpCamp 全栈项目 *02/2019-04/2019*

■ **技术栈**

- 前端：HTML，JS，CSS
- 后端：NodeJS，Express，MongoDB

全职科研项目 *01/2019-05/2019*

中国科学院深圳先进技术研究院（SIAT，CAS） *导师：**王晓东教授***

课题：基于 Tensor 模型的交通数据预测

■ **数据处理**

- 在纽约黄色出租车数据集，对原地点到目的地点之间的车辆需求量建立带有时间维度的 OD-Matrix
- 转换 Tensor 时间序列数据为 Matlab 格式；应用变换域的 MLDS 模型对时间序列数据建模

暑期科研项目 KAGGLE *07/2018-09/2018*

Home Credit Default Risk：通过数据挖掘和机器学习算法来估计客户的贷款违约概率.

■ **数据处理和特征工程**

- 检查数据缺失值并可视化数据分布；通过用 NAN 替换数据中正负 INF 值来预训练神经网络估算
- 在数据预处理之后进行数据特征提取;应用热编码促进了特征变量的利用
- 根据特征值对模型的贡献度，应用 LightGBM 模型提炼最终特征

■ **特征工程**

- 原始特征：运用简单的加减运算构造新的特征
- 统计特征：对数值变量应用聚合方法得到汇总特征
- 时序特征：采用固定时间窗口和次数构造得到新的时序特征

■ **单一模型优化**

- 可适用模型的论证：
  - ➢ 数据集本身的缺失值对神经网络预测准确率有负面影响；
  - ➢ Catboost 对于包含不同类型变量的数据集效果很好，但运行时间过长；
  - ➢ LightGBM 和 XGBoost 可以很好地对这个数据集预测，他们可以根据模型学习过程中的训练损失自动地填补缺失数据；
  - ➢ 应用贝叶斯优化选择模型超参数的最优组合；

■ **堆叠模型优化**

- 应用 10 次交叉验证得到每一次中预测值作为逻辑回归模型的训练测试的输入数据
- 基于 LightGBM 和 XGBoost 模型的堆叠模型，加强了数据优化

■ **结果：**

- ➢ 优化结果表明，曲线下面积（AUC）预测性能评价达到 0.7979；获得了一枚铜牌
- ➢ 偿还贷款的能力很大程度上取决于先前多少天向信贷局提出的申请

互联网搜索引擎项目： 自动下载和文本分类 *05/2018*

■ **第一部分: 中英文网页的自动下载过程**

- 应用 Python Beautiful Soup 在 Jupyter 上提取和压缩 HTML 网页的内容，选择有用的信息
- 通过自然语言工具包（NLTK）删除英语中的停用词以便于提取词干并保存简化 TXT 格式的输出
- 以与英文处理类似的方式执行中文停用词的 Porter Stemming 提取次干，并保存简化 TXT 的输出
- **成就：**
  - ➢ 所有步骤均已成功整合到一个完整的自动系统中
  - ➢ 扩展常用停用词的图表，以生成特定停用词的自适应图表

■ **第二部分：基于不同文本相似性的文本分类**

- **相似性比较机制：**
  - ➢ 根据单词出现次数设置单词频率的向量
  - ➢ 计算关于其向量的余弦距离的词的相似性
- **文本分类机制：**
  - ➢ 通过词频 - 逆文档频率（TF-IDF）将文本转换为词频矩阵
  - ➢ 通过 K-means 算法进行类别分析，并根据肘部法则（Elbow Rule）优化 k 值，将文本分为两类

在校科研项目训练（SRTP） *05/2017-04/2018*

西南交通大学，成都，中国

重叠社区发现检测及其在图数据挖掘中的应用探讨

- 通过将数据信息提取到节点和边文件中来预处理网络数据集
- 应用小型社交网络数据集，应用 Gephi 对基于标签的传播算法（LPA）和 Clique Percolation Method（CPM）算法进行可视化社区结构分析
- 改进了标签传播算法以提高挖掘性能并在重叠社区中的应用

奖励&荣誉

Kaggle 数据科学竞赛-奖金池类型 铜牌	*08/2018*
多次获西南交通大学奖学金	*2016-2018*
中国大学生数学建模比赛三等奖	*05/2016*

课内编程项目

基于 CC3200 开发板的嵌入式 WIFI 智能插座的设计	*07/2017*
Web 前端：企业员工年终总结与打分投票系统	*01/2018*
毕业设计：基于密度与标签传播的社区发现系统设计	01/2019-2019/06

Chenyang Shi

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

EDUCATION & ACADEMIC VISITING

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

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

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences *01/2019-05/2019*

Visiting Research Intern, Wireless Energy and Transportation Laboratory

COURSEWORK & SKILLS

Programming Languages: Python, R, CSS,HTML, JavaScript, Java, C, C++, Scala, Matlab.

Frame & Tools: NodeJS, Jupyter, Gephi, LaTeX, Spark.

Coursera: *Data Science, Machine Learning.*

Udemy：*The Web Developer Bootcamp* – **YelpCamp**

RESEARCH EXPERIENCES & INTERSHIPS

Full-time Research Assistant (visting student) *01/2019-05/2019*

Wireless Energy and Transportation Laboratory, SIAT, CAS ***Advisor: Prof. Xiaodong Wang***

Research on Traffic Prediction based on Tensor Methods (in process)

■ **Data Processing**

- Built OD-Matrix and added the time dimension by splitting the pick-up time to 15 minutes based on NYC yellow cabs dataset.
- Transformed the former tensor time series data into Matlab format; Applied Multilinear Dynamical System (MLDS) in a transform domain (eg. dft-MLDS) to model tensor time series.

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 0*5/2017-04/2018*

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

Web Project: Year-end Summary and Scoring System for Enterprise Employees	01/2018
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.