Jamal Ching-Chuan Chen
陳慶全

Data Engineer / Data Analyst / R

CONTACT

No.23, Aly. 150, Tianbao St., Xitun Dist.
Taichung City, 407 Taiwan
(private) +886-966-676-326
(work) +886-963-855-707

zw12356@gmail.com

www.facebook.com/celestial0230

www.linkedin.com/in/celestial0230

github.com/ChingChuan-Chen

EDUCATION

2012.09 2014.09

National Cheng Kung University, Tainan, TW

Master

GPA: 4.0

Thesis:

A Classification Approach Based on Density Ratio Estimation with Subspace Projection

Advisor:

Ray-Bing Chen

Abstract:

For imbalanced data, the density ratio estimation (Kanamori et al. (2009)) is good solution to solve it. However, its performance shrink fast for sparse data. Therefore, we propose a projection method to perform the dimension reduction and make data more crowded to distinguish. Our result shows that the proposed method is better than the original method.

2008.09 2012.06

National Cheng Kung University, Tainan, TW

Bachelor

GPA: 3.5

ABOUT

My name is Jamal Chen. I am a data engineer and data analyst with 3+ years of experience in big data infrastructure, big data computation, data preprocessing, data visualization and modeling. I am an experienced R and MatLab programmer, also a experienced Linux user. Also, I am familiar with Spark in Python and Scala. In some scenarios, I also write C++ for performance. Except for programming languages, I have experiences on manipulating databases like Oracle, Hive and MongoDB. As for high performance computing, I am good at SLURM and MPI. Most importantly, I have a Master degree in statistics, so this is different to other data engineers. I am capable of linear regression with or without penalties and know the details of theorem. Therefore, I can realize from scratch. I also know about some clustering methods like K-means, Gaussian mixture model. For machine learning, I also study decision tree, random forest and gradient boosting tree model and I can apply them to do prediction in R. To sum up, I am an engineer who is capable of infrastructure, big data computing tolls and statistics.

WORK EXPERIENCES

Taiwan Semiconductor Manufacturing Company Limited

Taichung, Taiwan
July 2016 - Present

Junior Engineer, CIM Department

Data engineer and data analyst in semiconductor manufacturing data.

Highlights

- I am assigned to survey, construct and maintain the first big data solution for our department. I used Apache Hive as data storage and Apache Spark as a tool to pull data from Oracle.
- I developed an algorithm to detect the mean shift and variance shift on final WAT data. I also developed a fast computing procedure to get results in 1 hour for 100 million data via Hive and HDFS.
- I used R to analyze billions of wafer-processing data to identify the key factors of yields. For example, I used gradient boosting tree model to find out the differences of history between bad wafers and good wafers by.
- My colleague and me developed a correlation system to reveal the correlation between the thousands of measurements in time by SLURM, MPI and MongoDB. In this system, I took robust correlation instead of Pearson's one because there are many strange outliers which we cannot simply rule out in our data.
- My colleague and me got the third place in a TSMC defect detection competition. My colleague and me customized a neural network model with two inputs. This model is based on Xception and Swish.
- I manage infrastructure of R which based on CentOS 7 and responsible for writing packages including the frequently-used functions.

Academia Sinica

Taipei, Taiwan
September 2015 - June 2016

Research Assistant, Institute of Statistical Science

Functional data analysis of traffic data provided by Taiwan freeway bureau.

Highlights

- My main task is to improve the clustering and regression methods for functional data. It is based on functional principal component analysis.
- I wrote an R program to schedulely scrawl data from websites with R and parse XML to

LANGUAGES

- Chinese — Native speaker
- English — Conversant
- Japanese — Conversant

REFERENCES

Ray-Bing Chen
Professor
Department of Statistics
National Cheng Kung University
+886-6-275-7575 ext. 53645
rbchen@mail.ncku.edu.tw

Sheng-Mao Chang
Associate Professor
Department of Statistics
National Cheng Kung University
+886-6-275-7575 ext. 53632
smchang@mail.ncku.edu.tw

Jeng-Min Chiou
Research Fellow
Institute of Statistical Science
Academia Sinica
+886-2-2783-5611 ext 312
jmchiou@stat.sinica.edu.tw

store into MongoDB.

- I used R shiny to build an interactive data visualization GUI for the highway data. It shows the daily traffic situation.

JOURNALS

milr: Multiple-Instance Logistic Regression with Lasso Penalty

Ping-Yang Chen, Ching-Chuan Chen, Chun-Hao Yang, Sheng-Mao Chang and Kuo-Jung Lee
The R Journal (2017) 9 :1 , pages 446-457 .
https://journal.r-project.org/archive/2017/RJ-2017-013/index.html

AWARDS

December 2017

TSMC Kaggle Competition for the Defect Recognition

Third Place

A internal competition inside TSMC. The competition is to classify 4 types of defects from defect and referenced images. The score is accuracy rate on non-opened testing images. Our team used home-made neural network with 2 inputs for defect and referenced images, the model is based on Xception and Swish.

August 2014

Competition for Data Analysis with R in Taiwan

Honorable Mention

A national competition for university and master students in Taiwan. Its purpose is to let participants find their own topic on a given dataset and try to explain their topic by data. The whole analysis was limited to use R. The data is collected from a registering system created by Taiwan governmet, the system contains the actual selling prices of real estate. Our team chose to predict the price of house from a messy data.

PROJECTS

Automatically Generated Resume

https://github.com/ChingChuan-Chen/python-yaml-resume
A tool for automatically generated resume written in Python by YAML and Jinja2.

Highlights

- Easily maintained resume by modifying the YAML file.
- Themes are easily changed by using different Jinja templates.

R package RcppBlaze

https://github.com/ChingChuan-Chen/RcppBlaze
Blaze is an open-source, high-performance C++ math library for dense and sparse arithmetic. This package provides the header files for linking Blaze library in Rcpp.

Highlights

- Full API from R to Blaze under the RcppArmadillo-like framework.

R package milr

https://github.com/PingYangChen/milr

This package performs maximum likelihood estimation for multiple-instance logistic regression utilizing EM algorithm with LASSO penalty.

Highlights

- It is a first R package addressing the analysis of the multiple instance data.
- This package provides a MLE with EM algorithm under the framework of logistic regression.
- This package provides not only prediction, but also variable selection with L1 penalty.
- The performance issues are solved by RcppArmadillo.