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, the performance of density ratio is poor when data is sparse in the high dimension. Therefore, we propose using projection to perform dimension reduction. 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. I have experiences on Oracle SQL, Hive SQL and MongoDB. I am also familiar with SLURM and MPI. I can combine the multiple tools to develop a stable and fast system to provide analyzed results from billions of data for users. Therefore, user can response for the changing in processing in time and not to worry about that the data size crashes their computer and can not response in time.

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

- Survey, construct and maintain the first big data solution for our department. Data storage is on Apache Hive and Apache Spark is used in pulling data from Oracle.
- Develop a fast-responding correlation analyzing system to reveal the correlation between the thousands of measurements by SLURM, MPI and MongoDB.
- Develop an algorithm to detect the mean shift and variance shift on WAT data for wafer processing.
- Using R to analyzing billions of wafer processing data to find the key factors of yields. For example, to find the differences on the history of bad wafer and good wafer.
- Manage R infrastructure and package 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

- Schedulely scrawl data from websites with R and parse XML-formatted data into MongoDB.
- Build an interactive data visualization GUI by R shiny for the highway data.
- Construct statistical models with MatLab and R. Focus on developing and improving clustering and regression methods for functional data.

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

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

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 maintain resume by modifying the YAML file.
- Simply changing Jinja template for different themes.

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

- A first R package address 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 panalty.
- The performance issues are addressed by RcppArmadillo.

SKILLS

R — Master

- Skilled at vectorizing programming and parallel programming.
- Mastering data.table for data manipulation in organizing billions of data.
- Good at massive data processing (100 billions of data) in MPI on SLURM cluster.
- Mastering lattice, ggplot2, plotly and shiny for data visualization.
- Model building for statistical models and machine learning.
- Linking with other programming languages (C/C++, Java) for improving performance.

MatLab — Master

- Skilled at vectorizing programming and parallel programming.
- Good at data manipulation and data visualization.
- Ability to link C++ for accelerating programs.

SQL — Advanced

- Familiar with Oracle SQL, MySQL SQL and Hive SQL.

Statistics / Machine Learning — Advanced

- Familiar with thoeries and good at explaining meaning for results.
- Experienced in hypothesis testing, statistical models, change point detection, clustering and dimension reduction.
- Experienced in generalized linear models with/without mixed effect, generalized estimating equation and generalized additive model.
- Experienced in functional data analysis including functional PCA, functional regression and functional clustering - Experienced in supervised learning and unsupervised learning.
- Experienced in decision tree, random forest and gradient boosting decision tree.

Python — High-Intermediate

- Familiar with numpy and pandas.
- Familiar with PySpark.

C++ — Intermediate

- Not so familiar with OOP, but good at using Armadillo and Eigen to accelerate program in R, MatLab.