



Jamal Ching-Chuan Chen 陳慶全

Data Engineer / Data Analyst / R

CONTACT

📍	No.114, Guangfeng St., Pingzhen Dist. Taoyuan City, 32452 Taiwan (private) +886-966-676-326 (work) +886-963-855-707
✉	zw12356@gmail.com
f	www.facebook.com/celestial0230
in	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 and I am a data engineer and data analyst with 3+ years of experience in big data infrastructure, data preprocessing and modeling. I am an experienced R programmer in data preprocessing and modeling, also a experienced Linux maintainer in automated process and system service management. I am familiar with packaging codes and reusing for efficient and fast development of applications. Also, I can give insights from data and provides picture for making decisions.

WORK EXPERIENCES

Taiwan Semiconductor Manufacturing Company Limited

Taichung, Taiwan
July 2016 - Present

Junior Engineer, CIM Department

Data engineer and data analyst on semiconductor manufacturing data.

Highlights

- ✔ Constructing a big data solution for analyzing manufacturing data. I used Apache Hive for data warehouse and Apache Spark for data streaming from Oracle to Apache Hive.
- ✔ Building up a development and production environment for R.
- ✔ Collecting usually used R functions and packing to R packages. For example, function to connect Oracle, Apache Hive and MongoDB, function of building SQL and often used graphical functions.
- ✔ Organizing R workshop to teach colleagues using R professionally and efficiently. Sometimes, I also introduce new R packages and skills of performing machine learning and data analysis.
- ✔ Analyzing the semiconductor manufacturing data like offline/inline measurement, tool sensors or yield related issues.
- ✔ Building up an R web service for other colleagues easily using R to plot or perform machine learning.
- ✔ Using statistical method to detect the shift of location or variance on the measurements.

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

- ✔ Automatically downloading open data from websites with R and parsing data in XML format for saving data to MongoDB.
- ✔ Constructing the statistical method with MatLab and R.
- ✔ Preprocessing data and analyzing the relationship between flow, speed and occupancy rate.
- ✔ Building an interactive data visualization for the highway data with shiny in R.

LANGUAGES

Chinese

Native speaker

English

Conversant

Japanese

Basic Knowledge

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

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.
- Providing not only prediction, but also variable selection with L1 penalty.
- The performance issues are addressed by using RcppArmadillo.

AWARDS

December
2017

TSMC Kaggle Competition for the Defect Recognition

🏆 Third Place

A internal competition in TSMC. Its purpose is to make classification of defects able to judge automatically by machine for lessening human cost. They provides 3000 pictures of 4 types of defects and let employees fit a deep learning model to classify. Then send the model to the platform for get the accuracy rate of testing set (1200 pictures.).

August
2014

Competition for Data Analysis with R in Taiwan

🏆 Honorable Mention

A national competition in Taiwan. Its purpose is to let participants find their own topic in given data and try to explain by data. The whole analysis need to be done by R. The data is collected from a registering system created by Taiwan governmnet of the actual selling price of real estate. Our team chose to predict the price of house from a messy data. Each team had the times of a day to finish their report.

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>

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.
- ✔ 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.
- ✔ Package development for reusing code and team development.

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 and MySQL SQL.

Statistics

Advanced

- ✔ Familiar with theories 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

Machine Learning

High-Intermediate

- ✔ 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 2 deep learning frameworks, Keras and mxnet.
- ✓ Familiar with PySpark.

Shell Script

Intermediate

- ✓ Building automated process and automated deployment of applications.

C++

Intermediate

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