Xiaojin Zhang

Mobile Phone:15807270905 | Email:xjzhang@cse.cuhk.edu.hk

Address: The Chinese University of Hong Kong, Shatin, N.T. Hong Kong

EDUCATION

09/2017 - Present **The Chinese University of Hong Kong, Hong Kong (PhD)**

- **Institute**: Computer Science and Engineering
- **Supervisor:** Shengyu Zhang (09/2017 - 06/2019) Siu On Chan (06/2019 - Present)
- **Major:** Computer Science

09/2014 - 07/2017 **University of Chinese Academy of Sciences, Beijing(Master)**

- **Institute**: Institute of Computing Technology
- **Supervisor**: Si-Min He
- **Major:** Computer Application Technology
- **Core Courses:** Advanced Algorithm Design (89/100), The Design and Analysis of Computer Algorithms (89/100), Machine Learning Methods for Computer Vision (95/100)

09/2010 - 07/2014 **Shandong University, Weihai (Bachelor)**

- **Institute:** School of Mathematics and Statistics, Elite Class
- **Dual Degree:** Statistics and Economics
- **Core Courses:** Math Analysis (99/100), Advanced Algebra (98/100), Functional Analysis (92/100), Applied Regression Analysis (99/100), Physics (96/100), Econometrics (96/100)

MAJOR AWARDS

06/2013 The 13th "Challenge Cup" National Undergraduate Curricular Academic Science and Technology Work Competition, Second Prize in College

04/2013 The Mathematical Contest in Modeling (MCM), Honorable Mentions

12/2012 Youth Entrepreneurship Competition, Third Prize in Shandong Province

10/2012 China Undergraduate Mathematical Contest of Modeling (CUMCM), National First Prize

09/2012 Summer Social Practice, First Prize at School

03/2012 Student Research Training Program, Second Prize at School\

11/2011 National Scholarship

10/2011 The First Prize Scholarship, Outstanding Student in College

07/2011 Essay Competition, Second Prize at School

05/2011 Competition of Body-building Exercises, Second Prize in College

ACADEMIC & COURSE PROJECTS

Student Research Training Program **03/2011 - 05/2012**

- Responsible for overall planning and allocated duty to team members based on their advantages.

- Applied my professional knowledge to establish logistic and SIR models for the analysis of micro-blog's potential value. Besides, a misconception was found and rectified through model analysis.
- Composed the final report, won approval and was invited to share my modeling experience with the participants in 2013.

External Sort - The First Program I Implemented **08/2014 - 09/2014**

- Given a file containing a large amount of double-precision floating-point numbers, e.g. two hundred and fifty million, sort them in ascending order and output into a new file.
- The loser-tree based merging approach was employed to refrain from the limitation of memory. To further improve the efficiency of the program, the radix sort for double-precision floating-point numbers was implemented.

Construct a Retrieval System **11/2015 - 12/2015**

- Responsible for overall planning and allocated duty to team members based on their advantages. It was impressive that several ideas proposed by me were found to have been introduced in the book, which made me joy and sorrow at the same time.
- Implemented single-link clustering and k-means clustering separately, analyzed the performances of these algorithms on various corpora. Reduced the complexity of these algorithms effectively.
- Constructed the query vector and document vector, and reduced the dimensions of these vectors by taking advantage of their sparsity. Implemented a text filtering algorithm based on the vector space model. Composed the final report.

Forecast Use of a City Bikeshare System **12/2014 - 01/2015**

- Combined historical usage patterns with weather data to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.
- Compared the effects of using different data preprocess methods such as PCA, log transform and zscore, different feature sets, different models including SVM combined with time series, GBRT, and the trained model with the smallest root mean squared logarithmic error on test data was selected.

Build Inverted Index **03/2015 - 05/2015**

- Simulated the fragmentation of the proteins into peptides according to the specified enzyme type, and built three index files to facilitate the search for peptide sequences satisfying the filtering criteria.
- Filtered out the redundant peptides based on loser tree merge and outputted the peptides in ascending order according to their mass value.
- The running time on a given dataset was reduced from almost 900 seconds to 1 second through optimization. As a result, this program was tested to have the highest efficiency as compared with my classmates.
- Composed the final report and won approval by my supervisor.

Create a Protein Search Engine **07/2015 - 08/2015**

- The MS/MS spectra were regarded as sentences, and the proteins were regarded as paragraphs, we need to figure out the matching relations between the sentences and the paragraphs.
- The initially matched peptides were imported based on the index files, and the modified peptides were generated based on the dynamic programming, then the peptides were further rated using the kernel-based spectral dot product and the one with the highest score was selected as the most confident identification.

- Wrote a script using python to compare the spectra identified by my engine and those identified by pFind 2.8, and composed the final report.

Identification of Glycopeptide with Machine Learning Model **11/2015 - 12/2015**

- Read hundreds of spectra and record the corresponding features. Through the combination of the conclusion of data analysis, literature research and the experiences accumulated in our group, more than 100 kinds of features were extracted.
- Tried distinct regression models including decision tree, adaboost, ridge, lasso, svr, and different classification models including random forest, gradient boosting, logistic regression, svm. The testing results showed that the assembled models like adaboost, gradient boosting and random forest performed better than the other models.

Glycan Database Construction Based on the Topological Structures **09/2015 - 12/2015**

- Took over the codes written by my senior brother. Aimed at building a rational as well as comprehensive glycan database, thereby making it more beneficial for the identification of glycopeptide.
- Improve the running time of the code from 650 seconds to 2 seconds.
- Detected the mirrored glycan structures by comparing the structures through a depth-first traversal, and two novel hashing methods were successively proposed to prune the comparison operations efficiently.
- The canonical representations for the glycans were generated. The glycans would be concluded as isomorphic if and only if they correspond to identical canons. As a result, the glycan isomorphism problem could be equivalently recast as the problem of identifying the duplicate numbers.
- Read lots of papers about data mining, and implemented different canonical representation schemes of glycans. In addition, distinct implementation methods were utilized for the same canonical representation scheme in order to select the most efficient one. The application of canonical representation not only ensured high efficiency in terms of application, but it also guaranteed the elegance of the algorithm in view of theory.

Glycan Database Construction Based on the Canonical Representations **02/2016 - 04/2016**

- Designed and implemented the algorithms including enumerating the glycans based on the canonical representations, generating the subglycans of each glycan and calculating the degree of each glyco according to the canonical representation of the glycan.
- Consequently, the memory required had been reduced and the efficiency of the program had been improved to a great extent. Additionally, the conclusion of literature research showed that no better algorithm had ever been put forward.

PUBLICATIONS (Authors are ordered alphabetically unless labeled by (*))

- Pinyan Lu, Chao Tao, Xiaojin Zhang. *Variance-Dependent Best Arm Identification.* **In Proceedings of the Conference on Uncertainty in Artificial Intelligence.** (UAI 2021) (To appear)
- Chungwei Lee, Haipeng Luo, Chenyu Wei, Mengxiao Zhang, Xiaojin Zhang. *Achieving Near Instance-Optimality and Minimax-Optimality in Stochastic and Adversarial Linear Bandits Simultaneously.* **In Proceedings of the 36th International Conference on Machine Learning.** (ICML 2021) (To appear)

- (*)Xiaojin Zhang, Honglei Zhuang, Shengyu Zhang, Yuan Zhou. *Adaptive Double-Exploration Tradeoff for Outlier Detection.* **The Thirty-Fourth AAAI Conference on Artificial Intelligence.** (AAAI 2020)
- (*)Ming-Qi Liu, Wen-Feng Zeng, Pan Fang, Wei-Qian Cao, Chao Liu, Guo-Quan Yan, Yang Zhang, Chao Peng, Jian-Qiang Wu, Xiao-Jin Zhang, Hui-Jun Tu, Hao Chi, Rui-Xiang Sun, Yong Cao, Meng-Qiu Dong, Bi-Yun Jiang, Jiang-Ming Huang, Hua-Li Shen, Catherine C. L. Wong, Si-Min He, Peng-Yuan Yang. *pGlyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification.* **Nature Communications**, 8, 438, 2017.
- (*)Wen-Feng Zeng, Yang Zhang, Ming-Qi Liu, Jian-Qiang Wu, Xiao-Jin Zhang, Hao Yang, Chao Liu, Hao Chi, Kun Zhang, Rui-Xiang Sun, Peng-Yuan Yang, Si-Min He. *Trends in Mass Spectrometry-Based Large-Scale N-Glycopeptides Analysis[J].* **Progress in Biochemistry and Biophysics**, 2016, 43(6):550-562.

MANUSCRIPT

- (*)Xiao-Jin Zhang, Wen-Feng Zeng, Jian-Qiang Wu, Yang Zhang, Ming-Qi Liu, Hao Yang, Rui-Xiang Sun, Peng-Yuan Yang, Si-Min He. *Construction of N-glycan databases based on a linear canonical representation of N-glycans.*
- (*)Xiaojin Zhang, Shuai Li, Weiwen Liu. *Contextual Combinatorial Conservative Bandits.*
- Xiaojin Zhang. *Automatic Ensemble Learning for Online Influence Maximization.*
- Xiaojin Zhang. *Near-Optimal Algorithm for Distribution-Free Junta Testing.* (Blurb on *Property Testing Review*. Cited by *https://eccc.weizmann.ac.il/report/2021/004/*.)
- Xiaojin Zhang. *Improved Algorithm for Testing Permutations.*

STUDY EXPERIENCE

Shanghai Qizhi Institute — Shanghai, China
Research Intern — 09/2020-01/2021

Shanghai University of Finance and Economics — Shanghai, China
Summer Course — 07/2018-08/2018

Tencent — Shenzhen, China
Research Intern — 06/2018-07/2018

Shanghai Jiao Tong University — Shanghai, China
Summer Course — 07/2017-08/2017

TEACHING EXPERIENCE

- Guest Lecture Assistant, Game Theory (Yao Class), Tsinghua University, May 10, 2021
- TA, Fundamentals for Embedded Systems (CENG 2030), CUHK, Spring 2021
- TA, Techniques for Data Mining (ENGG 5103), CUHK, Fall 2020
- TA, Probability and Statistics for Engineers (ENGG 2430), CUHK, Spring 2019
- TA, Design and Analysis of Algorithms (CSCI 3160), CUHK, Fall 2018
- TA, Approximation Algorithms (CSCI 5160), CUHK, Spring 2018
- TA, Design and Analysis of Algorithms (CSCI 3160), CUHK, Fall 2017

EXTRACURRICULUM EXPERIENCE

- Interested in playing piano, dancing and doing yoga.
- Served as a volunteer in HOBIE, and attended the etiquette training which was really beneficial.
- Took the duty of broadcasting Image-Text, invited by the communities as a media reporter.
- Succeeded in getting support from some relevant departments during the Summer Social Practice by virtue of perseverance.
- Organized and participated in various dance shows with my classmates.
- Cooperated with four classmates to translate Philip Guo's memoir--The Ph.D. Grind--into Chinese.

SKILLS

- CET4 (556), CET6 (511), GRE (152+165+3.5)
- Have experience with C++, Matlab, Lingo, Eviews, SPSS, R and Python.