Kaixun Hua

Address
2360 East Mall
Vancouver, BC Canada V6T 1Z3
https://kingsley1989.github.io/

E-mail: kaixun.hua001@umb.edu
E-mail: huakaixun@gmail.com
Mobile: +1(607) 379-9785(US)
Office: +1(604) 559-9695(CA)

EDUCATION

PhD, Computer Science
University of Massachusetts Boston, Boston, MA 09/2013 – 05/2019
Research Field:
- Information Theory based ensemble clustering and its application
- Entropy-based imbalanced-class classification
- Ultrametric distance and its application on hierarchical clustering

Thesis: Clusterability, Model Selection and Evaluation (Outstanding Research Award)
Advisor: Dan A. Simovici

Master of Engineering, Systems Engineering
Cornell University, Ithaca, NY 09/2012 – 08/2013
Thesis: A Study on State Estimation of Power System with Uniform Distributed Residuals
Advisor: Hsiao-Dong Chiang

Bachelor of Science, Electrical Engineering
Shanghai Jiao Tong University, Shanghai, China 09/2008 – 08/2012
University of Michigan - Shanghai Jiao Tong University Joint Institute (JI) at SJTU

WORK & RESEARCH EXPERIENCE

Ultrametricity and Clusterability
Supervisor: Professor Dan A. Simovici
Boston, MA 04/2018 – Current
- Deciding whether it is worth to do clustering on a dataset.
- Improving the clustering result by twisting the distance space of dataset.
- Defining a novel matrix multiplication method to achieve clusterable distance matrix.
- Proposing a measure of clusterability that quantifies the degree of how much inherent cluster structure the data possess.

Research Internship at State Grid Energy Research Institute
Project Leader: Linlin Wang
Beijing, China 05/2017 – 08/2017
- Making prediction on daily cash flow of State Grid Corporation of China.
- Modifying Random Forest to satisfy the problem and create models for long-period(30 days) daily cash flow prediction, including daily income, outcome and deposit of the whole corporation.
- Achieving relatively high correlation(≥ 0.78) between predicted value and the real data on income and outcome and reasonable correlation(≥ 0.6) on deposit.
- Discovering important date and financial features for forecasting future daily cash flow.

Finding the Number of Clusters using Partition Entropic Metrics on Partitions
Supervisor: Professor Dan A. Simovici
Boston, MA 04/2016 – Current
- Using the metric space of partitions of a finite set in the context of ensemble clustering to identify the "natural" number of clusters in a dataset.
- Introducing two methods of evaluation on the choice of number of clusters. One is based on the sum of pairwise distances between partitions and the other one is focus on the distance between clustering

results from different cut levels of hierarchical clustering to k-means clustering.
- Highlighting the relationship between an acceleration of the clustering process and the "natural" number of clusters existent in a data set, by defining the speed of hierarchical clustering.

Long-lead Term Precipitation Forecasting by Hierarchical Clustering-based Bayesian Structural Vector Autoregression
Supervisor: Professor Dan A. Simovici
Boston, MA 10/2015 – 02/2016
- Provided a different path of predicting on heavy precipitation by performing regression analysis using the precipitation amounts at particular locations.
- Raised a novel process of combining Hierarchical Clustering (HC) and Bayesian-based Structural Vector AutoRegression (BSVAR) to forecast the long-term precipitation.
- Proposed the relationship between the cut level of clustering geographic locations and the regression model performance.

Data Science Internship at insuranceQuotes
Project Leader: Titi Alailima
Cambridge, MA 07/2015 – 08/2015
- Verified whether customer demographic data are useful or not on assigning customer to proper insurance carriers.
- Applied random survival forest to produce models for each insurance carrier by training demographic data from the customer of the company.
- Achieved a result that random survival forest model gives $1 uplift for the customer with additional purchased demographic data, comparing with $0.4 uplift of original data.

Long-term Patient Outcome in a Large Cohort of Renal Transplant Recipients with Protocol Biopsies
Supervisor: Professor Dan A. Simovici
Protocol Biopsy Program, Boston, MA 12/2014 – 06/2015
- Created models that permit reliable prediction of death and survival are established and will be used to identify patients on risk.
- Applied four data mining algorithms (Naive Bayesian, C5.0 Decision Tree, RPART and Random Forest) are applied to make survival prediction on 761 Tx-patients from Hannover Medical School in Germany.

Ultrametricity of Dissimilarity Spaces and Its Significance for Data Mining
Supervisor: Professor Dan A. Simovici
Boston, MA 03/2014 – 12/2014

- Introduced a measure of ultrametricity for dissimilarity spaces and examine transformations of dissimilarities that impact this measure.
- Studied the influence of ultrametricity on the behavior of two classes of data mining algorithms (kNN classification and PAM clustering) that applied on dissimilarity spaces.
- Showed that there is an inverse variation between ultrametricity and performance of classifiers.

Design and Development of Smart Grid:
State Estimation of Power System with Uniformly distributed Residual
Supervisor: Professor Hsiao-Dong Chiang
Master of Enigneering Project, Ithaca, NY 09/2012 – 06/2013
- Designed the scheme of state estimation for power system, with weighted least square method (WLS) to get normal result.
- Modified a minimax approximation algorithm and embedded it into power system to achieve the uniformed error for final result.

TEACHING EXPERIENCE

Teaching Assistant: CS110 Introduction to Computing (Python)
University of Massachusetts Boston, Boston, MA 09/2015 – 06/2019

Teaching Assistant: CS420 Introduction to the Theory of Computation
University of Massachusetts Boston, Boston, MA 09/2014 – 6/2015

Teaching Assistant: CS240 Programming in C
University of Massachusetts Boston, Boston, MA 09/2013 – 6/2014

PUBLICATIONS

- Simovici, D. A. & **Hua, K.** (2019, February). Data Ultrametricity and Clusterability. In *2019 1st International Conference on Mathematical Models & Computational Techniques in Science & Engineering.*
- Xu, Z., Lian, J., Bin, L., **Hua, K.**, Xu, K., & Chan, H. Y. (2019). Water Price Prediction for Increasing Market Efficiency Using Random Forest Regression: A Case Study in the Western United States. *Water*, 11(2), 228. (**Co-Corresponding Author**).
- **Hua, K.**, & Simovici, D. A. (2018, September). Dual Criteria Determination of the Number of Clusters in Data. In *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)* (pp. 201-208). IEEE.
- Scheffner, I., **Hua, K.**, Simovici, D., Abeling, T., Haller, H., & Gwinner, W. (2016, June). Prediction of Patient Survival After Kidney Transplantation: Construction, Validation and Evaluation of Decision Models Using Data Mining Approaches. In *AMERICAN JOURNAL OF TRANSPLANTATION* (Vol. 16, pp. 572-573). 111 RIVER ST, HOBOKEN 07030-5774, NJ USA: WILEY-BLACKWELL.
- **Hua, K.**, & Simovici, D. A. (2016, April). Long-lead Term Precipitation Forecasting by Hierarchical Clustering-based Bayesian Structural Vector Autoregression. In *2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC)* (pp. 1-6). IEEE.
- Simovici, D. A., Vetro, R., & **Hua, K.** (2017). Ultrametricity of dissimilarity spaces and its significance for data mining. In *Advances in Knowledge Discovery and Management* (pp. 141-155). Springer, Cham.

INVITED TALKS

- "Clusterability and Model Selection", Bigwood System Inc., Cornell Business & Technology Park, Ithaca, NY, USA. August 2019.
- "Application of Unsupervised Learning (Clustering) and Artificial Intelligence in Power System Industry", State Grid Corporation of China, Shanghai, China. July 2018.

SERVICE AND AFFILIATIONS

- Program Committee Member: Conference on Information and Knowledge Management (CIKM) '19
- Journal Reviewer: Knowledge and Information System (KIS)

SKILLS & TECHNIQUES

- Computer/Statistic Language: R, Python, Matlab, SAS, Java, C/C++.
- Experience with PSAT matlab platform

REFERENCES

Simovici, Dan A.
Graduate Director (Advisor)
Department of Computer Science
College of Science and Mathematics
UMass Boston

Ding, Wei
Program Director
Information and Intelligent System
Computer, Information Science and Engineering
National Science Foundation

Chen, Ping
Associate Professor
Department of Engineering
College of Science and Mathematics
UMass Boston