Curriculum Vitae

XIAOXIAO (RICKY) SHI

Ph.D. in Computer Science (Machine Learning)
University of Illinois at Chicago
Office: 851 S. Morgan St., Rm 1336 SEO, Chicago, IL 60607
xshi9@uic.edu, xiao.x.shi@gmail.com (preferred)
Office +1-312-965-8103
http://www.cs.uic.edu/~xiaoxiao

EDUCATION

Ph.D. in Computer Science
Department of Computer Science, University of Illinois at Chicago, Chicago, IL, U.S.A. Advisor: Philip S. Yu (from Sep. 2009 to August 2013)

M.Sc. in Applied Math
Department of Mathematics, University of Illinois at Chicago, Chicago, IL, U.S.A (from Sep. 2010 to Dec. 2012)

Visiting Scholar Machine Learning and Data Mining, Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong. Advisor: Qiang Yang (from Sep. 2008 to Feb. 2009)

M.Sc. Department of Computer Science, Sun Yat-sen University, China, Jul. 2009.

B.Eng. Department of Computer Science, Sun Yat-sen University, China, Jul. 2007.

ACADEMIC RECORDS

- Computer Science PhD Graduate Program at UIC: GPA (4.0/4.0).
- Applied Math Graduate Program at UIC: GPA (4.0/4.0).
- Undergraduate at SYSU: Overall GPA (3.8/4.0), Major GPA (3.9/4.0), rank 1st among 62 students.

WORK EXPERIENCE

- Morgan Stanley, Equity Trading Labs (ETL, systematic trading), on deriving and developing scalable and robust machine learning models for large-capacity fundamental and statarb trading strategies.
- Eagle Seven Trading Firm, May 2012 to August 2012, on statistical arbitrage: (1) Inferring the probability of informed trading; (2) Deriving market marking strategy via cost-sensitive decision tree. (3) FX Post-trade analysis software.
- AT&T Labs, May 2011 to July 2011, on deriving and developing machine learning algorithms with heterogeneous sources (C++ and Java). The proposed algorithm optimizes the prediction models by using multiple data sources with heterogeneous structures (e.g., user profile, user behavior database, users' social netowrk, etc.). The corresponding paper is chosen as one of the 10 best papers in SIAM conference on data mining.

- Yahoo! Labs, May 2010 to August 2010, on deriving and developing large scale user profile pattern mining algorithms based on cloud computing system Hadoop (Java). A compression model is proposed to explore the most distinguishable patterns of the user groups by using the Dirichlet Process Clustering. The corresponding paper is published in SIGIR.

PUBLICATIONS

2013 **Xiaoxiao Shi**, Qi Liu, Wei Fan, and Philip S. Yu, "Transfer across Completely Different Feature Spaces via Spectral Embedding", *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2013, pp. 906-918.

2013 **Xiaoxiao Shi**, Jean-Francois Paiement, David Grangier, and Philip S. Yu, "GBC: Gradient Boosting Consensus Model for Heterogeneous Data", *Statistical Analysis and Data Mining*, 2013.

2013 Weixiang Shao, **Xiaoxiao Shi**, and Philip S. Yu, "Clustering on Multiple Incomplete Datasets via Collective Kernel Learning", IEEE International Conference on Data Mining (ICDM 2013).

2013 **Xiaoxiao Shi**, Wei Fan and Philip S. Yu, "Dynamic Shaker Detection from Evolving Entities", 2013 SIAM International Conference on Data Mining (SDM 2013).

2012 Qi Liu, Han Zhou, **Xiaoxiao Shi**, Wei Fan, Ruixin Zhu, Philip S. Yu and Zhiwei Cao, "In-silico Target-specific siRNA Design based on Domain Transfer in Heterogeneous Data", *PLOS Computational Biology* (**Impact factor: 5.2**).

2012 **Xiaoxiao Shi** and Philip S. Yu, "Dimensionality Reduction on Heterogeneous Feature Space", IEEE International Conference on Data Mining (ICDM 2012, acceptance rate 10.7%).

2012 Xian Wu, Wei Fan, Meilun Sheng, Li Zhang, **Xiaoxiao Shi**, Zhong Su, Yong Yu, "A framework to represent and mine knowledge evolution from Wikipedia revisions", the World Wide Web Conference (WWW 2012): 633-634.

2012 **Xiaoxiao Shi**, Jean-Francois Paiement, David Grangier, and Philip S. Yu, "Learning from Heterogeneous Sources via Gradient Boosting Consensus", 2012 SIAM International Conference on Data Mining (SDM'12).

2012 Guan Wang, Yuchen Zhao, **Xiaoxiao Shi**, and Philip S. Yu, "Magnent Community Identification on Social Networks", Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'12), 2012.

2012 **Xiaoxiao Shi**, Xiangnan Kong, and Philip S. Yu, "Transfer Significant Subgraphs across Graph Databases", 2012 SIAM International Conference on Data Mining (SDM'12).

2011 **Xiaoxiao Shi**, Yao Li, and Philip S. Yu, "Collective Classification with Latent Graphs", the 20th ACM Conference on Information and Knowledge Management (CIKM'11), Glasgow, UK, 2011 (acceptance rate: 15%).

2011 **Xiaoxiao Shi**, Wei Fan, Jianping Zhang, and Philip S. Yu, "Discovering Shaker from Evolving Entities via Cascading Graph Inference", Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'11), San Diego CA, 2011 (acceptance rate: 17.5%).

2011 **Xiaoxiao Shi** and Philip S. Yu, "Limitations of Matrix Completion via Trace Norm Minimization", *SIGKDD Explorations*, 2011, pp. 12(2):16-20.

2011 Xiangnan Kong, **Xiaoxiao Shi**, and Philip S. Yu, "Multi-label Collective Classification", 2011 SIAM International Conference on Data Mining (SDM'11), Mesa AZ, 2011 (acceptance rate: 25.07%).

2010 **Xiaoxiao Shi**, Wei Fan, and Philip S. Yu, "Efficient Semi-supervised Spectral Co-clustering with Constraints", 2010 IEEE International Conference on Data Mining (ICDM'10), 2010 (acceptance rate: 155/797=19.44%).

2010 **Xiaoxiao Shi**, Qi Liu, Wei Fan, Philip S. Yu, and Ruixin Zhu, "Transfer Learning on Heterogenous Feature Spaces via Spectral Transformation", 2010 IEEE International Conference on Data Mining (ICDM'10), 2010 (acceptance rate: 155/797=19.44%).

2010 **Xiaoxiao Shi**, Kevin Chang, Vijay K. Narayanan, Vanja Josifovski and Alex J. Smola, "A Compression Framework for Generating User Profiles", 2010 ACM SIGIR workshop on feature generation and selection for information retrieval, Geneva, Switzerland, July, 2010.

2010 **Xiaoxiao Shi**, Qi Liu, Wei Fan, Qiang Yang and Philip S. Yu, "Predictive Modeling with Heterogeneous Sources", 2010 SIAM International Conference on Data Mining (SDM 2010), Columbus, Ohio (acceptance rate: 82/351=23.36%).

2009 **Xiaoxiao Shi**, Wei Fan, Qiang Yang and Jiangtao Ren, "Relaxed Transfer of Different Classes via Spectral Partition", 2009 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2009) September 7-11, 2009, Bled, Slovenia (acceptance rate: 105/422=24.88%).

2008 **Xiaoxiao Shi**, Wei Fan, and Jiangtao Ren "Actively Transfer Domain Knowledge", 2008 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2008), Antwerp, Belgium (acceptance rate: 98/521=18.81%).

2008 Jiangtao Ren, **Xiaoxiao Shi**, Wei Fan, and Philip S. Yu "Type Independent Correction of Sample Selection Bias via Structural Discovery and Re-balancing", 2008 SIAM International Conference on Data Mining (SDM 2008), Atlanta, GA, Apr 2008 (acceptance rate: 77/282=27.30%).

PROFESSIONAL ACTIVITIES

- Co-organize the 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications in KDD 2013.
- Co-organize the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications in KDD 2012.
- Associate Editor of Transactions on Knowledge Discovery from Data (TKDD).
- Reviewer of IEEE Transactions on Knowledge and Data Engineering (TKDE).
- Reviewer of IEEE Intelligent Systems.

HONORS AND AWARDS

- The 2nd place of the first intern hackday, a competition where over 170 interns both from leading tech companies throughout the Bay Area and leading schools (Berkeley, Stanford, Waterloo, U Penn, and MIT among others) have participated. My team (3 members) ranks 2nd with a system to predict who is going to change jobs, and how it affects his/her social network. A video introduction of the competition can be found at `http://www.youtube.com/watch?v=BUKFJUttBOo`.
- The 6th place of the ACM-ICPC Programming Contest, Guangdong Provincial Contest, 2006 (rank 6th among 212 contest teams).
- Brainbench Master C++ Test: 4.84, higher than 96% of previous examinees; Brainbench Master Java Test: 4.12, higher than 97% of test takers; Brainbench Master C# Test: 4.47, higher than 99% of test takers (Transcript ID: 10851086).
- "Learning from Heterogeneous Sources via Gradient Boosting Consensus" was selected as one of the 10 best papers published in SDM'12.

- Hewlett Packard Fellowship, University of Illinois at Chicago, 2011 (only 1 student from Computer Science Department gets the award).
- University Fellowship, University of Illinois at Chicago, 2009 (only 1 student from Computer Science Department got the award).
- Departmental Fellowship, Computer Science Department, University of Illinois at Chicago, 2009 (only 1 student from Computer Science Department got the award).
- Wexler Award, University of Illinois at Chicago, Summer 2011 (3% of the students).
- Wexler Award, University of Illinois at Chicago, Fall 2011 (3% of the students).

MAIN RESEARCH AREAS AND INTERESTS

- Heterogeneous Learning: modeling when the data have different distributions, feature spaces, structures and output spaces. For example, how to aggregate social network, user profile, as well as users' historical activities to derive a better recommendation system? Related fields: transfer learning, sample selection bias, domain adaptation, multi-task learning, multi-view learning and relational learning.
- Semi-supervised learning: modeling with labeled and unlabeled data. Related fields: semi-supervised relational learning and semi-supervised co-clustering.
- Influence/Causality Inference: a field that combines time series and statistical inference. Its aim is to discover causality relationships given a set of evolving objects (time-series data). For example, given the dynamics in social network, how can we find out who affects whom?

MAIN APPLICATION AREAS AND INTERESTS

- Big Data and Cloud Computing: deriving machine learning algorithms that can work on the clusters with the "MAP-REDUCE" schema (Hadoop), in order to handle large scale dataset.
- Computational Advertising: on advertisement recommendation systems based on users historical behaviors.
- Knowledge Discovery from Social Networks: modeling trends, event sequences, user behavioral patterns in social network.
- Causal Relationship Analysis: given evolving objects (time series data), determine (1) who affects whom, and (2) how the effects are propagated, and (3) detect the most influential objects [published in KDD'11].

SYSTEM DEVELOPMENT EXPERIENCES

- LinkedOut. It is a system to predict who is going to change jobs in the late future, and how his/her move affects his/her social network. It is a work that I did in a global competition (LinkedIn Hackday). I mainly designed the prediction model, and implemented it in script language and a JAVA-based HCI (human computer interaction) framework called Gephi. I also implemented part of the HCI interface. For the HCI interface, I implemented an effect that simulates the elastic effect of a spring in physics. I wrote over 8,000 lines of codes in one day. The system finally wined the 2nd place over 170 competitors.
- Learning from multiple heterogeneous data sources via gradient boosting consensus. It is a work that I did when I was a summer intern in AT&T Labs in 2011. At that time, one challenging task for AT&T is to build a prediction model that can make good use of heterogeneous datasets (e.g., datasets about

user behavior and datasets about social network). No model has been proposed to solve the problem before. I defined the problem from scratch, and proposed an optimization framework to solve the problem. The model is derived from the gradient boosting decision tree, and it is further modified to solve the case when there are multiple heterogeneous datasets. There were mainly 3 research scientists working on the project with about 10 more research scientists providing help and discussion. I wrote about 30,000 lines of codes in Java and C++. The model was run on a cloud computing system called Condor. This work is published in SIAM conference on data mining, and is selected as one of the 10 best papers.

- Compression based user profile generation and advertisement recommendation. It is a work that I did when I was a summer intern in Yahoo! Labs in 2010. I mainly derived a map-reduce framework to explore an essential feature subspace of the users' historical behaviors based on the non-parametric clustering technique. The precision of the advertisement recommendation is improved in the reduced feature subspace. There were mainly 5 research scientists working on the project with 20 more research scientists providing help and discussion on the project. I wrote over 20,000 lines of codes in JAVA under HADOOP (a "MAP-REDUCE" framework) as well as thousands of lines of script languages. I contributed the main framework of the proposed model and set up extensive experiments on the large scale datasets. This work is published in SIGIR'10. Furthermore, the work is packed into a system used in Yahoo! Inc. and a related patent is submitted.

- Co-develop the "PJ8 Finance Management System" for the Guangzhou Financial Bureau of China. "PJ8" is a project worthy of $5,000,000 which is now the core system for the finance budget and balance of the whole Guangzhou city. The objective of the system is to integrate data from 8 existing systems of different local financial bureaus of Guangzhou, and evaluate the finance budget and balance of the previous year. It also schedules financial resources and predicts the balance of the current year. There were 34 people in the project. I joined the development of the system during the summer internship in a Chinese IT company called ZSOFT in 2007 and mainly served as a requirement analyst and programmer. As a requirement analyst, I mainly contributed the requirement reports from 3 local financial bureaus, and wrote 2 chapters (for 2 function modules) of the design manual. As a programmer, my contribution was mainly on the database design and implementation, user interface design using the AJAX technique. I wrote about 9,000 lines of codes in JAVA and in a framework language called "ZK".

- Spam Mining. It was a course project to discover the faked reviews in resellerrating.com[1]. There are over 20,000 stores and over 200,000 reviews. I mainly contributed a concept-based model to find suspicious reviews with strange concepts related to the stores. There were two people working in this project and I took the lead. The whole system has about 7,000 lines of codes, and I wrote 3,000 lines in JAVA and 1,000 lines in Matlab.

- Optimal Market Marking Strategy via Cost-sensitive Decision Tree. I derive a machine learning model to determine how to submit an order in the market marking process. For example, one can submit a market order, or submit a limit order competing with the best prices, or submit a limit order one or two ticks deviating from the best prices. In this task, I basically design 5 class labels (2 ticks higher, 2 ticks lower, 1 tick higher, 1 tick lower, trade at market), and convert the market order book data into a summarized feature vector with various indication values (e.g., bid/offer spread, bid/offer ratio, etc.). I then build a cost-sensitive decision tree to learn the best policy for market making. Furthermore, I also apply an artificial neural network system called "LAMSTAR" (A Very Large Scale Memory Neural Network) to derive the optimal strategy. This work is selected to appear in the 3rd edition of the book "Principles of Artificial Neural Networks" by Daniel Graupe. The models are written in R, Java and Matlab.

- MPI Parallel Computing. It was a course project to design different logical schema (e.g., ring structure, hypercube structure, etc.) in parallel computing. It was written in pure C, and I wrote about 3,000 lines of C codes in the project.

[1] http://www.resellerratings.com/rlist-s1-n2.html

PROGRAMMING SKILLS

- Extensive experience with C++ and Java. Familiar with MATLAB/R/PYTHON/C# and used them extensively in implementing various algorithms and systems. Brainbench Master C++ Test: 4.84, higher than 96% of previous examinees; Brainbench Master Java Test: 4.12, higher than 97% of test takers; Brainbench Master C# Test: 4.47, higher than 99% of test takers.
- Quite familiar with cloud computing, especially the "MAP-REDUCE" framework such as "Hadoop".
- Frequent use of different kinds of softwares/script languages/environments, e.g., LATEX, GNUPLOT, CYGWIN (Unix virtual machine), PROLOG, WEKA and the like.
- Familiar with HCI based language such as Gephi, Processing.

MATHEMATICS SKILLS

- Matrix computation, optimization and linear inference problems, especially deriving spectral theorems and applying spectral models such as spectral clustering, spectral classification, spectral embedding, and the like to capture hidden correlations among different statistics.
- Information theory, probability inference and their applications on data mining.
- Familiar with ODE, PDE and other applied maths.

REFERENCES

Philip S. Yu Professor in Computer Science Department, University of Illinois at Chicago.
Address: 851 S. Morgan St., Rm 1138 SEO, Chicago, IL 60607.
Email: psyu@cs.uic.edu

Wei Fan Associate Director, Leader of DMML Group, Noah's Ark Lab.
Address: Units 525-530, Core Building 2, Hong Kong Science Park Shatin, Hong Kong.
Email: wei.fan@gmail.com

Charles Knessl Professor in Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago.
Address: 851 S. Morgan St., Rm 722 SEO, Chicago, IL 60607.
Email: knessl@uic.edu