Shandian Zhe

Area: statistical machine learning & data mining

Personal Information

Email:szhe@purdue.edu **Home Page**: http://www.cs.purdue.edu/homes/szhe/

Phone: 219-629-1630 **Address**: 223 Arnold Drive, Apt. 10, West Lafayette, IN 47906

Education

Ph.D. candidate in Computer Science, Purdue University 01/2012 - present

M. Sc. in Computer Science, Chinese Academy of Sciences (CAS) 09/2007-03/2011

B. Eng. in Computer Science in Beijing Univ. of Aeronautics and Astronautics (BUAA) 09/2003-07/2007

Current GPA: 3.87

Related Courses

Statistical Machine Learning: A+	Data Mining: A	Algorithm: A	Numerical Analysis: A-
Optimization: A	Advanced Bayesian Learning: A	Randomized Algorithm: A	

Internship Experiences

Yahoo! Labs, Research Intern 05/2015-08/2015

- Distributed nonlinear tensor factorization on Spark, for Click-Through-Rate prediction.

NEC Laboratories America, INC, Research Intern 05/2014-08/2014

- Dynamic adaptive lasso for time series data analysis.

Research Experiences

Distributed flexible nonlinear tensor factorization on SPARK 09/2015－11/2015

- Proposed a new nonlinear tensor factorization model; the model places Gaussian process prior over the tensor entries and is flexible to use balanced tensor entries for training and overcomes the learning bias problem in existing nonlinear factorization method, InfTucker.
- Developed a distributed variational inference algorithm with SPARK: it optimizes a tight variational evidence lower bound on optimal variational posteriors, prevents inefficient EM procedure and avoid expensive data shuffling on MapReduce.
- Outperforms the state-of-the-art large scale tensor factorization algorithm, GigaTensor and DinTucker. Application on CTR prediction shows a 20% improvement over logistic regression and linear SVM.

Scalable nonparametric multiway data analysis 09/2014－11/2014

- Proposed a scalable, nonparametric Bayesian model for large scale multiway data analysis. The model integrates Dirichlet process and local Gaussian process to capture an undetermined number of latent clusters and potential nonlinear relationships between data entries.
- Proposed an efficient online VB-EM algorithm for model learning. Evaluated the proposed model on large datasets with billions of elements and it shows a significant better predictive performance than the state-of-the-art multiway data decomposition method, GigaTensor.

Distributed infinite Tucker decomposition 08/2013－12/2013

- Proposed DinTucker, a hierarchical nonparametric Bayesian model for tensor factorization.
- Devised a distributed online learning algorithm for model estimation (inference).
- Implemented and tested the algorithm with Python under Hadoop platform. The model obtained significantly better prediction accuracy than alternatives like TUCKER and PARAFAC. In large datasets (with billions of tensor entries), the proposed algorithm outperforms a distributed version of PARAFAC--GigaTensor in terms of both running time and prediction accuracy.

- Cooperated with IBM Thomas J. Watson Research Center to use the proposed model to analyze a large log from source code management system..

Large Scale Bayesian Sparse Learning 09/2013 – 04/2014

- Develop a large scale Bayesian spike and slab inference algorithm based on Laplace approximation, ensemble Nystrom and Gauss quadrature. The algorithm is successfully applied on Region-of-Interest (ROI) study on brain.

Multiview learning for association discovery 09/2012 – 05/2013

- Proposed a Bayesian model to combine multiple data views for prediction task and at the same time to extract associations between different data views; valuable biological knowledge, such as LD structure are encoded as a prior distribution in the model.
- Applied the proposed model on Alzheimer’s disease data and discovered meaningful associations between brain regions and gene fragments (SNPs). The model also obtained better prediction accuracy for Alzheimer’s disease status than alternative methods, e.g., lasso, elastic net and LapSVM.

Network and node selection 01/2012 – 08/2012

- Propose a hybrid Bayesian model to select useful networks and important nodes inside the selected networks. The model uses the structure of networks and selects useful networks (by conditional component) and nodes (by generative component), jointly.
- Applied the proposed model on biological microarray datasets and found critical regulatory pathways and genes related to some cancers. Those findings are supported by published biological literatures.

Publications

1. **Shandian Zhe**, Yuan Qi, Youngja Park, Ian Molloy and Suresh Chari, DinTucker: Scaling up Gaussian Process Models on Large Multidimensional Arrays, S. Zhe, Y. Qi, Y. Park, I.M. Molloy, and S. N Chari, AAAI-2016 (to appear).
2. **Shandian Zhe**, Pengyuan Wang, Kuang-chih Lee, Zenglin Xu, Jian Yang, Youngja Park and Yuan Qi, Distributed Flexible Nonlinear Tensor Factorization for Large Scale Multiway Data Analysis. NIPS 2015 Workshop on Networks in the Social Science and Information Sciences.
3. Syed A.Z. Naqvi, **Shandian Zhe**, Yuan Qi and Jieping Ye, Fast Laplace Approximation for Sparse Bayesian Spike and Slab Models. NIPS 2015 Workshop on Advances in Approximate Bayesian Inference.
4. **Shandian Zhe**, Zenglin Xu, Xinqi Chu, Yuan Qi and Youngja Park, Scalable Nonparametric Multiway Data Analysis. Artificial Intelligence and Statistics (AISTATS), 2015.
5. Changying Du, **Shandian Zhe**, Fuzhen Zhuang, Yuan Qi, Qing He, Zhongzhi Shi. Bayesian maximum margin PCA. The 29th AAAI Conference on Artificial Intelligence (AAAI-15), Austin, Texas, USA, January 25-29, 2015.
6. **Shandian Zhe,** Zenglin Xu, Yuan Qi and Peng Yu, Sparse Bayesian multiview learning for simultaneous association discovery and diagnosis of Alzheimer's Disease, AAAI 2015 (**Outstanding student paper honorable mention**).
7. **Shandian Zhe**, Zenglin Xu, Yuan Qi and Peng Yu, Joint association discovery and diagnosis for Alzheimer's Disease by supervised heterogeneous multiview learning, in Proceedings of the Pacific Symposium on Biocomputing (PSB), The Big Island of Hawaii, 2014.
8. **Shandian Zhe,** Syed A.Z. Naqvi, Yifan Yang, and Yuan Qi**,** Joint network and node selection for pathway-based genomic data analysis, Bioinformatics, Oxford, 2013.

Programming Skills

1. Excel in C/C++, Python, Matlab, C#, PHP, Javascript; proficient in Hadoop, Spark and Java; skilled in programming and debugging in Linux/Windows; familiar with SVN.
2. Mandarin Chinese (native), English (fluent)