

Research Interests

- Applied Machine Learning and Data Mining
- Information Retrieval and Text Mining
- Natural Language Processing
- Computational Biology

Education

- Ph.D. Candidate, Carnegie Mellon University** 2001 – Present
Language Technologies Institute, School of Computer Science
- Advisor: Prof. Yiming Yang
 - Dissertation Title: Structure learning with large sparse undirected graphs and its applications
 - Expected in June 2007
- Master of Science, Carnegie Mellon University** 2001 – 2004
Language Technologies Institute, School of Computer Science
- Master of Science, Tsinghua University** 1998 - 2001
Department of Computer Science and Technology
- Advisor: Prof. Yuchang Lu
 - Dissertation Title: New approaches in text categorization
- Bachelor of Science, Huazhong Univ. of Science and Technology** 1994- 1998
Department of Computer Science and Technology

Publication

Conference Papers and Technique Reports

- [1] **Fan Li**. **Structure learning with large sparse undirected graphs and its applications**, *Thesis Proposal, Carnegie Mellon University, 2006*
- [2] **Fan Li**, Yiming Yang, Eric P. Xing. **From Lasso regression to feature vector machine**, *Advances in Neural Information Processing Systems 18 (NIPS2005)*
- [3] **Fan Li**, Yiming Yang. **Use modified Lasso regressions to learn large undirected graphs in a probabilistic framework**, *20th National Conference on Artificial Intelligence (AAAI 2005)*
- [4] **Fan Li**, Yiming Yang. **An analysis of recursive feature elimination methods for statistical classification**, *27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 2005)*
- [5] **Fan Li**, Yiming Yang. **Using recursive classification to discover predictive features**, *20th Annual ACM Symposium on Applied Computing (ACM SAC 2005)*
- [6] **Fan Li**, Yiming Yang. **Learning the structure of linear Bayesian networks with a large number of variables**. *Search and Discovery in Bioinformatics workshop, 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 2004)*
- [7] **Fan Li**, Yiming Yang. **A loss function analysis for classification methods in text categorization**. *20th International Conference on Machine Learning (ICML 2003)*
- [8] Mingyu Lu, Qiang Zhou, **Fan Li**, Yuchang Lu, Lizhu Zhou. **Recommendation of web pages based on**

concept association. *WECWIS 2002*: 221-227

Journal Papers

[9] Fan Li, Yiming Yang. **Analysis of recursive gene selection approaches from micro-array data.** *Bioinformatics*, 2005 21(19):3741-3747

[10] Fan Li, Yiming Yang. **Recovering genetic regulatory networks from micro-array data and location analysis data,** *Journal Genome Informatics*, 2004, Vol. 15, No.2

[11] David D. Lewis, Yiming Yang, Tony Rose, Fan Li. **RCV1: A new text categorization test collection,** *Journal of Machine Learning Research* 5 (2004) 361-397

[12] Fan Li, Mingyu Lu, Yuchang Lu. **The research on some new feature selection approaches in text categorization,** *Tsinghua Academic Journal (Natural Science Version)*, Vol.41, No.7, pp 98-101, 2001.7

Ongoing Papers and Technique Reports

[13] Fan Li, Yiming Yang. **A new approach to classify structured data with arbitrary topologies,** *to be submitted*, 2007

[14] Fan Li, Cliff Brunk, Alexandrin Popescul and Tong Zhang. **Graph regularization with annotated edges and its application for webpage categorization,** *submitted*, 2007

[15] Fan Li, Yiming Yang and Eric P. Xing. **Inferring regulatory networks using a hierarchical Bayesian graphical Gaussian model,** *CMU-MLD Technical Report 06-117*, 2006

Research Experience

Information Retrieval Group, Language Technology Institute, Carnegie Mellon University

Research Assistant

2001 – Present

- **Structure Discovery From Relational Data for Text Mining and Bioinformatics.**
 - Structure discovery can help us to explore the hidden patterns in many real-world domains, which are richly structured, consisting of objects related to each other in complex ways. For example, unclassified textual documents may belong to multiple hidden topics correlated to each other. Learning such a topic network would help us to organize and search document collections in a much better way. Other examples include learning user networks from web data and gene networks from Bio-data. However, the problem becomes intractable for traditional statistical learning algorithms when the graph is large.
 - I have developed novel probabilistic generative models and inference/learning algorithms for structure discovery, relational link prediction and latent theme extraction. A global hierarchical Bayesian prior for the precision matrix of the Graphical Gaussian Model (GGM) is introduced to impose a bias toward sparse graph structures. Several main research problems are addressed in our framework and we show that the MAP estimation of the undirected graph can be finally obtained by solving a set of modified Lasso regressions. Our approach is scalable to very large datasets and we have successfully applied it on the real bio-medical data to learn genome regulatory networks (with about 20,000 nodes). We also applied it on the Reuters textual corpus to learn networks of class labels in order to help text categorization. My ongoing research in this line includes learning large latent semantic networks from textual data. (See [1], [3], [6], [15])
- **Sparse Regression within Non-linear Models.**

I extended Lasso regression into a more general form isomorphic to SVM, which is referred as Feature Vector Machine (FVM). FVM can be easily generalized into non-linear versions by introducing kernels and soft margins defined on feature vectors. It generates sparse solutions in the feature space and it is much more tractable compared to traditional non-linear feature selection approaches such as feature scaling kernel machines. Based on FVM, our structure-learning framework can also be generalized to capture non-linear correlations among nodes in the graph.

(See [2])

- **Relational Classification for Structured Data.**
 - I proposed a novel model within a graph regularization framework for classification tasks on richly structured data (like web-pages).
 - Compared to traditional graph-based semi-supervised learning approaches, our model can learn the weights of edges in the graph based on observed labels, which makes it much more powerful in exploiting the structure information
 - Compared to discriminative relational models like conditional random fields (CRF), our model keeps the nice property that the optimal solution can be found efficiently with arbitrary graph topologies, while CRF takes exponential computational cost when the graph structure goes beyond chains and trees.
 - The experimental results on textual datasets are very encouraging compared to several baselines. (See [13])
- **Large-scale Text Categorization.**

Developed various large-scale text categorization systems and conducted experiments on several benchmark corpuses (including RCV1, OSHMED, Reuters 21578, etc). Investigated how to use the hierarchical structure of class labels to improve the categorization performance. Our experimental results on the RCV1 dataset have become a popular baseline in the field of automated text categorization. (See [7] ,[11])
- **Feature Selection from Genome Micro-array Data.**

Investigated different types of gene selection algorithms and developed a novel approach that has been applied successfully to extract disease-related genes from genome micro-array data. This approach has also been applied to improve the text categorization performance by eliminating redundant or irrelevant text features. (See [4], [5], [9])
- **Data Mining from Protein Mass Spectrums.**

Collaborated with faculties in Biochemistry Department at Vanderbilt University and developed novel protein identification and disease prediction systems based on mass spectrums.

Yahoo! Applied Research, Yahoo! INC.

Research Intern

Jul 2006 – Sep 2006

- Developed novel algorithms to classify web pages using hyperlinks and anchor text information within a graph regularization framework. (See [14])
- Improved the web-page classification accuracy significantly compared with state-of-art baseline approaches.

Center for Genomic Sciences, Allegheny General Hospital

Consultant

2005 – Present

- Analyzed micro-array data from patients and developed gene selection approaches to identify important genes related to certain diseases

Yibao Beixin Net. Co. Ltd, Beijing, China

Research Intern

2000 - 2001

- Chinese web-page classification and latent semantic extraction.
- Chinese phrase segmentation using statistical approaches.

State Key Lab of Intelligent Technology at Tsinghua University

Research Assistant

1998 – 2001

- Developed text-mining systems that extract association rules from text corpuses.
- Designed a text summarization system that automatically generates brief reports for the Largest Tobacco Industry in China. (See [8], [12])

Honors and Awards

- **Graduate Research Fellowship**, Carnegie Mellon University 2001 - Present
- **Graduate Research Fellowship**, Tsinghua University 1998 - 2001

- **Best undergraduate thesis**, Huazhong Univ. of Science and Technology 1998
 - Top 2.5%, 50 out of 2000 candidates
- **Undergraduate Fellowship**, Huazhong Univ. of Science and Technology 1994 - 1997

Professional Service

Reviewer for Journals:

- Bioinformatics
- IEEE/ACM Transactions on Computational Biology and Bioinformatics
- IEEE Intelligent Systems on Web Mining
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Journal of Machine Learning Research

Reviewer for Conferences:

- ACM SIGIR 2003,2004,2006
- ACM SIGKDD 2003

Student member of *Admissions Committee* at LTI/CMU (2005-2006)

References

Professor **Yiming Yang**

LTI/SCS, Carnegie Mellon University
 NSH 3612D, 5000 Forbes Ave, Pittsburgh, PA, 15213
 (412)268-1364
 yiming@cs.cmu.edu

Allen Newell Professor **Jaime Carbonell**

Director of Language Technologies Institute,
 LTI/SCS, Carnegie Mellon University
 NSH 4519, 5000 Forbes Ave, Pittsburgh, PA, 15213
 (412)268-7279
 jgc@cs.cmu.edu

Assistant Professor **Eric P. Xing**

CALD/SCS, Carnegie Mellon University
 4127 Wean Hall, 5000 Forbes Ave, Pittsburgh, PA, 15213
 (412)268-2559
 epxing@cs.cmu.edu