

Yan Zhou

Department of Computer Science
Washington University
Campus Box 1045
One Brookings Drive
St. Louis, Missouri 63130
(314) 935-4425

7860 Laclede Forest Dr.
St. Louis, Missouri 63143
(314) 644-9528

zy@cs.wustl.edu

RESEARCH INTERESTS

Machine Learning, robotics and related topics.

EDUCATION

Washington University, St. Louis, Missouri

D.Sc. in Computer Science, expected December 2001.
Dissertation title: Enhancing Supervised Learning with Unlabeled Data.
Research Advisor: Professor Sally Goldman.

The University of Mississippi, Oxford, Mississippi

M.S. in Computer Science, July 1998.
Project title: Neural Network Learning from Incomplete Data.
Research advisor: Professor Dawn Wilkins.

Xi'an Jiaotong University, Xi'an, China

Graduate study in Thermophysics Engineering, August 1994 – May 1996.
B.S. in Fluid Engineering, May 1994.
Project: 3D computing of a flow field with multigrid methods

EXPERIENCE

Washington University

1998 – present Research Assistant in Computer Science Department.
Summer 2001 Instructor for CS514 – Fundamentals of Computer Science.

The University of Mississippi

1996 – 1998 Research Assistant in National Center for Physical Acoustics.
1997 – 1998 Teaching Assistant in Computer Science Department.

Xi'an Jiaotong University, China

1994 – 1996 Research Assistant in Thermophysics Engineering.

PROFESSIONAL SERVICE AND HONORS

Professional Service

Refereeing for Annual Conference on Computational Learning Theory, 2000

Refereeing for Annual Conference on Computational Learning Theory, 2001

Refereeing for International Conference on Machine Learning, 2001

Honor

Boeing McDonnell Foundation Fellowship, 2001-2002

PUBLICATIONS

“A General Purpose Co-Learning Technique”, Technical Report, Department of Computer Science, Washington University, 2001

“Enhancing Supervised Learning with Unlabeled Data”, with Sally Goldman, *In the Proceedings of the Seventeenth International Conference of Machine Learning*, pp. 327-334, 2000

“Neural Network Control for A Fire-Fighting Robot,” with Dawn Wilkins and Robert P. Cook, *Software – Concepts and Tools* 19 3, pp. 146-152, 1999.

“Neural Network Learning from Incomplete Data,” Master’s thesis, TR1998-13, Department of Computer and Information Science, The University of Mississippi.

November 5, 2001

STATEMENT OF RESEARCH

My research interest is empirical machine learning study, especially supervised learning in situations in which the provided labeled data is incomplete and/or in short supply. My doctoral research concerns enhancing supervised learning for practical learning problems where there is a small amount of labeled data, along with a large pool of unlabeled data. Most standard supervised learning algorithms are designed to only take labeled examples as input, and usually do not perform well when provided with a small supply of labeled data. My work is to develop and study a new semi-supervised learning method called *co-learning*, for learning with both labeled and unlabeled data. Co-learning can leverage the difference between two or more different learning algorithms by allowing them to label data for each other. The idea is that several different learning algorithms may have different representations of the hypotheses and they are likely to detect different patterns in labeled data. A learning algorithm is allowed to label only those examples for which it has high confidence in its predictions.

I have implemented two different co-learning strategies, one is statistical co-learning, in which confidence of hypotheses is measured by using statistical approaches, and only two learning algorithms are used; the other one is democratic co-learning, where data is selected and labeled according to the majority vote of more than two learners. My empirical study shows statistical co-learning works well when two learning algorithms output quite different hypotheses, however, the performance is inconsistent for different input sizes. Statistical co-learning is prone to combining failure and *over co-learning* which cause the generalization error to increase due to serious distortion of data distribution. Democratic co-learning produce more consistent performance, since more than two learners are used to decide when and which data should be labeled according to the majority vote. Voting power from the majority group can dwindle the degradation caused by combining failure in many cases.

Co-learning highly relies on accurate confidence estimate of output hypotheses. This requirement is difficult to fulfill when the number of labeled examples is exceedingly small, therefore, more labeled examples must be provided. The most important issue is how to select a small set of the most valuable data for training. I proposed a priority sampling technique in use of active learning together with co-learning. My empirical study shows active learning can bootstrap and improve co-learning with small input sizes.

A very important part of my research is to explore the applicability of co-learning strategies in real-world applications. Part of my research is to test how the statistical co-learning strategy performs in text categorization. I obtained very promising preliminary results on Ken Lang's Newsgroups database.

I would like to continue this research on more real-world problems. Issues such as combining failure and over co-learning in statistical co-learning need to be resolved. Similar aspects need to be further investigated in the democratic co-learning technique.

My master's research involves the analysis and implementation of neural network algorithms for learning with incomplete data. Unlike a regular training example, an incomplete example has missing values for some of its attributes. Incomplete data may appear in many practical problems where training data may be collected from incomplete census database, recordings with unreliable sensors, or data with various number of meaningful attributes. This research has practical significance in this kind of domains.