University of Technology Sydney

Incorporating Prior Domain Knowledge into Inductive Machine Learning
Its implementation in contemporary capital markets

A dissertation submitted for the degree of

Doctor of Philosophy in Computing Sciences

by

Ting Yu

Sydney, Australia

2007

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as a part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

_____

Signature of Candidate

Table of Contents

List of Figures

List of Tables

Acknowledgments

This research is supported by the Institute for Information and Communication Technologies (IICT), the e-Markets Research Program in the University of Technology, Sydney, Capital Markets CRC Ltd., the International Institute of Forecasters (IIF) and the SAS Institute.

I would like to thank my advisors Dr Tony Jan, A/Professor Simeon Simoff, and Professor John Debenham in the Faculty of Information Technology and Professor Donald Stokes in the School of Accounting for their efforts. I would like to thank Dr Longbing Cao, Dr Maolin Huang, Professor Chengqi Zhang and A/Professor Tom Hintz in the Faculty of Information Technology, Prof Mark Tennant in the Graduate School, Dr Boris Choy, Dr Yakov Zinder, Dr Mark Craddock, Professor Alexander Novikov, Dr Ronald M. Sorli and Dr Narelle Smith in the Department of Mathematical Sciences, and A/Professor Patrick Wilson and Professor Tony Hall in the School of Finance and Economics, University of Technology, Sydney for their kind supports. I would like also to thank Deborah Turnbull for her precise proof-reading.

I would like to also thank my colleagues, Dr Debbie Zhang, Paul Bogg and other members in the e-Markets Research Group, my friends, Angie Ng and Caesar Tsai, and numerous reviewers.

Final and most thanks to my parents, who support me with their love for all my life.

Thank you.

Sydney, Australia, February 2007

Vita

2003 - 2007	Ph.D (Computing Science), The Faculty of Information Technology, The University of Technology, Sydney, Australia.
2001 - 2002	M.sc (Distributed Multimedia System), The School of Computing, The University of Leeds, Leeds, United Kingdom.
1993 - 1997	B.eng (Industrial Design), The College of Computer Science, The Zhejiang University, Hangzhou, P.R. China.

Publications

Ting Yu, Simeon Simoff and Donald Stokes. *Incorporating Prior Domain Knowledge into a Kernel Based Feature Selection*. The 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007), Nanjing, P.R. China, May 2007.

Ting Yu. *Incorporating Prior Domain Knowledge into Inductive Machine Learning*. Technique Report in the International Institute of Forecasters (IIF), The University of Massachusetts, Amherst, USA, October 2006.

Ting Yu, Tony Jan, John Debenham and Simeon Simoff. *Classify Unexpected News Impacts to Stock Price by Incorporating Time Series Analysis into Support Vector Machine*. The 2006 International Joint Conference on Neural Networks

(IJCNN 2006), 2006 IEEE World Congress on Computational Intelligence, 16-21 July, Vancouver, BC, Canada

Ting Yu, John Debenham, Tony Jan and Simeon Simoff, 2006, *Combine Vector Quantization and Support Vector Machine for Imbalanced Datasets*, in International Federation Information Processing, Volume 217, Artificial Intelligence in Theory and Practice, ed. M. Bramer, Boston: Springer, pp. 81-88.

Ting Yu, Tony Jan, John Debenham and Simeon Simoff. *Incorporate Domain Knowledge into Support Vector Machine to Classify Price Impacts of Unexpected News*. The Fourth Australasian Data Mining Conference, 5-6 December 2005, Sydney, Australia 2004

Ting Yu, Tony Jan, John Debenham and Simeon Simoff. *Incorporating Prior Domain Knowledge in Machine Learning: A Review*. 2004 International Conference on Advances in Intelligent Systems - Theory and Applications. November 2004, Luxembourg

Tony Jan, Ting Yu, John Debenham and Simeon Simoff. *Financial Prediction using Modified Probabilistic Learning Network with Embedded Local Linear Models*. CIMSA2004-IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, July 2004, Boston, MD, USA

Ting Yu and John Debenham. *Using an Associate Agent to Improve Human-Computer Collaboration in an E-Market System*. WCC2004-AIAI,The Symposium on Professional Practice in AI, August 2004, Toulouse, France

Ting Yu and Simeon J. Simoff. *Plan Recognition as an Aid in Virtual Worlds*. Australian Workshop on Interactive Entertainment 2004,13th February 2004, Sydney, Australia

Abstract

An ideal inductive machine learning algorithm produces a model best approximating an underlying target function by using reasonable computational cost. This requires the resultant model to be consistent with the training data, and generalize well over the unseen data. Regular inductive machine learning algorithms rely heavily on numerical data as well as general-purpose inductive bias. However certain environments contain rich domain knowledge prior to the learning task, but it is not easy for regular inductive learning algorithms to utilize prior domain knowledge. This thesis discusses and analyzes various methods of incorporating prior domain knowledge into inductive machine learning through three key issues: consistency, generalization and convergence. Additionally three new methods are proposed and tested over data sets collected from capital markets. These methods utilize financial knowledge collected from various sources, such as experts and research papers, to facilitate the learning process of kernel methods (emerging inductive learning algorithms). The test results are encouraging and demonstrate that prior domain knowledge is valuable to inductive learning machines.