

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

© Copyright by

Ting Yu

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

1	Introduction	1
1.1	Overview of Incorporating Prior Domain Knowledge into Inductive Machine Learning	2
1.2	Machine Learning and Prior Domain Knowledge	3
1.2.1	What is Machine Learning?	3
1.2.2	What is prior domain knowledge?	6
1.3	Motivation: Why is Domain Knowledge needed to enhance Inductive Machine Learning?	9
1.3.1	Open Areas	12
1.4	Proposal and Structure of the Thesis	13
2	Inductive Machine Learning and Prior Domain Knowledge . .	15
2.1	Overview of Inductive Machine Learning	15
2.1.1	Consistency and Inductive Bias	17
2.2	Statistical Learning Theory Overview	22
2.2.1	Maximal Margin Hyperplane	30
2.3	Linear Learning Machines and Kernel Methods	31
2.3.1	Support Vector Machines	33
2.3.2	ϵ -insensitive Support Vector Regression	37
2.4	Optimization	38
2.5	Prior Domain Knowledge	39

2.5.1	Functional or Representational Form	42
2.5.2	Relatedness Between Tasks and Domain Knowledge	42
2.6	Summary	46

3 Incorporating Prior Domain Knowledge into Inductive Machine

Learning	48
3.1	Overview of Incorporating Prior Domain Knowledge into Inductive Machine Learning	49
3.2	Consistency, Generalization and Convergence with Domain Knowl- edge	52
3.2.1	Consistency with Domain Knowledge	52
3.2.2	Generalization with Domain Knowledge	54
3.2.3	Convergence with Domain Knowledge	58
3.2.4	Summary	59
3.3	Using Prior Domain Knowledge to Preprocess Training Samples . .	62
3.4	Using Prior Domain Knowledge to Initiate the Hypothesis Space or the Hypothesis	65
3.5	Using Prior Domain Knowledge to Alter the Search Objective . .	69
3.5.1	Learning with Constraints	70
3.5.2	Learning with Weighted Examples:	76
3.5.3	Cost-sensitive Learning	77
3.6	Using Domain Knowledge to Augment Search	78
3.7	Semi-parametric Models and Hierarchical Models	79

3.8	Incorporating Financial Domain Knowledge into Inductive Machine Learning in Capital Markets	82
3.9	Summary	85
4	Methodology	87
4.1	Semi-parametric Hierarchical Modelling	88
4.1.1	Vector Quantization	90
4.1.2	VQSVM and Semi-parametric Hierarchical Modelling	92
4.1.3	Remarks of the Proposal Model	94
4.2	A Kernel Based Feature Selection via Analysis of Relevance and Redundancy	95
4.2.1	Feature Relevance and Feature Redundancy	96
4.2.2	Mutual Information Measurement	100
4.2.3	ML and MP Constraints	102
4.2.4	Remarks of the Proposal Method	103
4.3	Rule base and Inductive Machine Learning	103
5	Application I – Classifying Impacts from Unexpected News Events to Stock Price Movements	106
5.1	Introduction	106
5.2	Methodologies and System Design	109
5.3	Domain Knowledge Represented By Rule Bases	111
5.3.1	Domain Knowledge	111

5.3.2	Using Domain Knowledge to help the data-preparation of Machine Learning	112
5.3.3	Using Time Series Analysis to Discover Knowledge	115
5.4	Document Classification Using Support Vector Machine	117
5.5	Experiments	118
5.6	Summary	120
6	Application II – Measuring Audit Quality	121
6.1	VQSVM for Imbalanced Data	122
6.1.1	Algorithms	124
6.1.2	Experiments	128
6.1.3	Summary	132
6.2	Feature Selections	134
6.2.1	Algorithms	134
6.2.2	Experiments	137
6.2.3	Summary	141
7	Conclusion and Future Works	143
7.1	Brief Review and Contributions of this Thesis	143
7.2	Future Research	145
A	Abbreviation	151
B	Table of Symbols	153
	References	154

LIST OF FIGURES

1.1	Inductive Machine Learning System by the General Learning Process	5
1.2	An Example of Inductive Machine Learning without and with Domain Knowledge	11
1.3	Domain Knowledge and Inductive Machine Learning in the Given Domain	12
2.1	More than one functions that fits exactly the data	24
2.2	Risk, Empirical Risk vs. Function Class	25
2.3	Architecture of SV Machines	34
3.1	A spectrum of learning tasks [Mit97b]	49
3.2	A set of hypothesis spaces	55
3.3	8-2-1 Artificial Neural Network	56
3.4	Virtual Samples in SVM	63
3.5	Kernel Jittering	67
3.6	Catalytic Hints	74
4.1	A Simple Example of Two-dimensional LBG-VQ [GG92]	91
4.2	Hierarchical Modelling with SV Machines, which modifies the original SVM (see figure 2.3)	93
4.3	$\{f_1, f_2\}$ are relevant features, but the structures of influence are different	98
4.4	The domain knowledge may be contained by the optimal set of features (or approximated Markov Blanket) partially or entirely. .	103

5.1	Structure of the classifier	110
5.2	The half hourly Volume Weighted Average Prices and net-of-market return sequences of AMP	118
5.3	Shocks (large volatilities), Trend and Changing Points	119
6.1	VQSVM in case of a linear separable data set	125
6.2	In the audit data set, the observations (the bigger points) of the minority class is scattered with those (the smaller darker points) of the majority class	129
6.3	The measurements in the FCBF algorithm proposed by Lei Yu and Huan Liu [YL04]	135
6.4	The test accuracy of models constructed by a SVR while eliminat- ing one feature every iteration. The x-axis is the index of feature in the ascending list, and the y-axis is the R-square.	137
6.5	The test accuracy of SVR models while eliminating one feature every iteration. The x-axis is the index of feature in the ascending list, and the y-axis is the value of R-square.	140

LIST OF TABLES

1.1	An example of a data set	5
2.1	A list of empirical risk and true risk functions	19
3.1	Comparison of purely analytical and purely inductive learning [Mit97b]	50
6.1	Four UCI data sets and Audit data set with the numbers of local models.	128
6.2	Test Results of the VQSVM	131
6.3	The RBF SVM with the different values of sigma produces differ- ent lists of ranked features. The lists are ordered from the least important to the most important.	138

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*. CIMS2004-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.