

Learning Comprehensible Theories from Structured Data

Kee Siong Ng

A thesis submitted for the degree of
Doctor of Philosophy of
The Australian National University

April 2005
(Revised October 2005)

© Kee Siong Ng

Typeset in Palatino by T_EX and L^AT_EX 2_ε.

Except where otherwise indicated, this thesis is my own original work.

Kee Siong Ng
17 April 2005

To my parents
Lim Lian Hua and Ng Yong Swee
For their love, dedication, and courage.

Acknowledgements

People don't have to like or support you,
so you always have to say thank you.

Ruben Studdard

I would like to thank

my long-time supervisor John Lloyd for guidance and assistance of *every* kind going back five years. It's hard to imagine one can have a better supervisor.

my advisors Alex Smola and Gunnar Rätsch for exposing me to statistical machine learning through discussions and the weekly machine learning reading group.

my friend and ex-officemate Evan Greensmith for technical inputs of various kinds.

my colleagues (some of them ex-colleagues) Tony Bowers, Joshua Cole, Matthew Gray, Eric McCreath, Mathi Nagarajan, Vineet Nair and Xiaobing Wu for helpful discussions and valuable collaborations.

Michelle for making life so easy for everyone in the department.

my hosts during my travel to Europe Peter Flach, Stephen Muggleton, John Shawe-Taylor, James Cussens, and Luc De Raedt. I learnt a lot from that trip.

Dr Yin-tak Woo of the US Environmental Protection Agency and Dr Nick Dixon and Professor Rod Rickards of the Research School of Chemistry at the ANU for expert advice on predictive toxicology. Jean Jiayu Wen and Yu Di also helped in my (futile) endeavour to understand a little more of biology and chemistry.

The Australian National University and the Smart Internet Cooperative Research Centre for their generous scholarships.

my friends F.H. Huang (Dajie), Linda, Thararat, Yee Tuan, Agnes, Cheng Soon, Doug, Edward and Annie, Evan, Kerry, Kristy, Wongas (and Fiona), and Xiaobing for walks, talks, cakes, (Belgian) chocolates, fish, rice puddings in gula Melaka, sushi, bo-bo cha-cha, delicious curry, etc.

Mei for a very special relationship.

Ginny for being you.

my family for love and unconditional support over the last 27 years; my brother deserves special thanks.

It has been a most rewarding and exciting three years!

Abstract

This thesis is concerned with the problem of learning comprehensible theories from structured data and covers primarily classification and regression learning. The basic knowledge representation language is set around a polymorphically-typed, higher-order logic. The general setup is closely related to the learning from propositionalized knowledge and learning from interpretations settings in Inductive Logic Programming. Individuals (also called instances) are represented as terms in the logic. A grammar-like construct called a predicate rewrite system is used to define features in the form of predicates that individuals may or may not satisfy. For learning, decision-tree algorithms of various kinds are adopted.

The scope of the thesis spans both theory and practice. On the theoretical side, I study in this thesis

1. the representational power of different function classes and relationships between them;
2. the sample complexity of some commonly-used predicate classes, particularly those involving sets and multisets;
3. the computational complexity of various optimization problems associated with learning and algorithms for solving them; and
4. the (efficient) learnability of different function classes in the PAC and agnostic PAC models.

On the practical side, the usefulness of the learning system developed is demonstrated with applications in two important domains: bioinformatics and intelligent agents. Specifically, the following are covered in this thesis:

1. a solution to a benchmark multiple-instance learning problem and some useful lessons that can be drawn from it;
2. a successful attempt on a knowledge discovery problem in predictive toxicology, one that can serve as another proof-of-concept that *real* chemical knowledge can be obtained using symbolic learning;
3. a reworking of an exercise in relational reinforcement learning and some new insights and techniques we learned for this interesting problem; and
4. a general approach for personalizing user agents that takes full advantage of symbolic learning.

Contents

Acknowledgements	vii
Abstract	ix
1 Introduction	1
1.1 The General Problem	1
1.2 A Symbolic Approach	2
1.3 Two Scientific Questions	2
1.4 Thesis Organization	3
1.5 The Alkemy Software	3
2 Knowledge Representation	5
2.1 Introduction	5
2.2 Representation of Individuals	6
2.3 Representation of Features	8
2.4 Predicate Construction	11
2.4.1 Predicate Enumeration	13
2.4.2 Structuring the Search Space	17
2.4.3 Operations on Predicate Rewrite Systems	18
2.5 Data Types and Transformations	23
2.6 Related Work	25
3 Classification	27
3.1 Introduction	27
3.2 Learning Algorithms	28
3.2.1 Learning Stumps	28
3.2.2 Learning Trees	33
3.2.3 Learning Lists	38
3.2.4 Others	40
3.3 Function Classes and Their Relationships	41
3.3.1 Basic Setup	42
3.3.2 Basic Class Definitions	43
3.3.3 Relationships	45
3.4 Generalization Bounds	47
3.4.1 Classical Bounds	50
3.4.2 Data-Dependent Bounds	51
3.4.3 Some Tools for Calculating VC Dimensions	53

3.4.4	Five Illustrations	62
3.4.5	Tighter Bounds	66
3.5	Optimization Issues	67
3.5.1	The Stump Algorithm	67
3.5.2	The Top-Down Tree-Induction Algorithm	70
3.5.3	The Covering Algorithm	75
3.6	PAC Learnability	78
3.6.1	PAC Learning	79
3.6.2	Generating Efficiently-Computable Predicates	79
3.6.3	PAC Learnability of Stumps, Lists, and Trees	82
3.6.4	Agnostic PAC Learning	83
3.6.5	Learning In Practice	86
3.7	Related Work	87
4	Regression	89
4.1	Introduction	89
4.2	Learning Algorithms	89
4.2.1	Learning Stumps	89
4.2.2	Learning Trees	96
4.2.3	Others	98
4.3	Generalization Bounds	99
4.4	Related Work	100
5	Incremental Induction	101
5.1	Introduction	101
5.2	Regression	102
5.2.1	Properties of the Batch Algorithm	102
5.2.2	The Incremental Algorithm	103
5.3	Classification	111
5.3.1	Properties of the Batch Algorithm	112
5.3.2	The Incremental Algorithm	112
5.4	Discussion	113
5.5	Related Work	116
6	Applications	119
6.1	Introduction	119
6.2	Multiple-Instance Learning — Musk	120
6.2.1	Representation of Individuals	120
6.2.2	An Early Effort	121
6.2.3	Doing it without Bunyip	122
6.2.4	A Pitfall in Learning with ALKEMY	124
6.2.5	An Observation	125
6.3	Knowledge Discovery — Predictive Toxicology	126
6.3.1	The 2000-1 Predictive Toxicology Challenge (PTC)	126

6.3.2	Representation of Individuals	127
6.3.3	A First Experiment	128
6.3.4	A Second Experiment	132
6.3.5	Other Features	135
6.3.6	Evaluation	135
6.3.7	A Predicate Rewrite System for PTC	138
6.4	Relational Reinforcement Learning — Blocks World	142
6.4.1	The Basic Framework	143
6.4.2	Blocks World	145
6.4.3	Experiments and Results	150
6.4.4	Discussions	154
6.5	Personalization — An Infotainment Agent	154
6.5.1	An Infotainment Agent	155
6.5.2	Adaptation	157
6.5.3	Personalization of the TV Recommender	157
6.5.4	Experiments and Results	161
6.5.5	Conclusions and Future Work	164
6.6	Other Applications	164
7	Comparison and Evaluation	167
7.1	First-order vs Higher-order Learning: Practical Differences	167
7.1.1	Programming Language	167
7.1.2	Representation of Individuals	169
7.1.3	Construction of Hypothesis Languages	170
7.2	Quantitative Performance Comparisons	172
7.2.1	ALKEMY's Performance on Attribute Value Data	172
7.2.2	ALKEMY's Performance on Structured Data	174
8	Conclusion	183
8.1	Thesis Contributions	183
8.2	Future Work	184
	Bibliography	187
	Index	202

