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)

Typeset in Palatino by TEX and LATEX 2ε.

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 demontrated 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