ON THE INDUCTION OF DECISION TREES FOR MULTIPLE CONCEPT LEARNING

by

Usama M. Fayyad

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in The University of Michigan
1991

Doctoral Committee:

Professor Keki B. Irani, Chairman
Professor John H. Holland
Assistant Professor John E. Laird
Associate Professor Robert K. Lindsay
Dr. Ramasamy Uthurusamy, Research Scientist,
General Motors Research Laboratories

Author's current address:

Usama M. Fayyad
AI Group M/S 525-3660
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109

Fayyad@aig.jpl.nasa.gov

ABSTRACT

ON THE INDUCTION OF DECISION TREES FOR MULTIPLE CONCEPT LEARNING

by

Usama M. Fayyad

Chairman: Keki B. Irani

We focus on developing improvements to algorithms that generate decision trees from training data. This dissertation makes four contributions to the theory and practice of the top-down non-backtracking induction of decision trees for multiple concept learning. First, we provide formal results for determining how one generated tree is better than another. We consider several performance measures on decision trees and show that the most important measure to minimize is the number of leaves. Notably, we derive a probabilistic relation between the number of leaves of the decision tree and its expected error rate.

The second contribution deals with improving tree generation by avoiding problems inherent in the current popular approaches to tree induction. We formulate algorithms GID3 and GID3* that are capable of grouping irrelevant attribute values in subsets rather than branching on them individually. We empirically demonstrate that better trees are obtained. Thirdly, we present results applicable to the binary discretization of continuous-valued attributes using the information entropy minimization heuristic. The results serve to give a better understanding of the entropy

measure, to point out desirable properties that justify its usage in a formal sense, and to improve the efficiency of evaluating continuous-valued attributes for cut point selection. We then proceed to extend the binary discretization algorithm to derive multiple interval quantizations. We justify our criterion for deciding the intervals using decision-theoretic principles. Empirical results demonstrate improved efficiency and that the multiple interval discretization algorithm allows GID3* to find better trees.

Finally, we analyze the merits and limitations of using the entropy measure (and others from the family of impurity measures) for attribute selection. We argue that the currently used family of measures is not particularly well-suited for attribute selection. We motivate and formulate a new family of measures: C-SEP. The new algorithm, O-BTREE, that uses a selection measure from this family is empirically demonstrated to produce better trees.

Ample experimental results are provided to demonstrate the utility of the above contributions by applying them to synthetic and real-world problems. Some applications come from our involvement in the automation of semiconductor manufacturing techniques.

To My Parents

ACKNOWLEDGEMENTS

I often wondered whether individuals whose dissertations I came across tended to exaggerate in acknowledging "almost everyone on the planet." I now know better.

First and foremost, I wish to express my gratitude to Keki Irani, my thesis adviser, for making it all possible. We progressed from "informal" Friday afternoon meetings where we discussed life in general, into intense, sometimes prolonged, technical meetings as we both explored this field that's riddled with heuristics. Keki Irani taught me what research is all about: from what courses to take, to writing funding proposals and papers; he made my education as comprehensive as I could ever hope it to be. I thank him for his friendship, guidance, financial support, criticism, discipline, as well as encouragement throughout the years of my research. He also introduced me to "industry" through a great research opportunity at GM Research Labs where I finally converged on decision trees for a thesis topic (our shared, predawn, commutes to GMR were also very enjoyable). Perhaps his most lasting effect on me is due to his consistently steering me in the direction of mathematics and formal approaches to problems, and in demanding preciseness in whatever statements I make.

John Laird taught me "classical" AI and symbolic machine learning. He also opened up many professional opportunities for me, including my summer job at JPL. I thank him for his thorough, careful readings of this and many other drafts I passed his way. John Holland gave me a totally different perspective on the field. Besides introducing me to "nontraditional" machine learning, he really helped me put works in the field in their proper perspective. I have met no one who could do that as eloquently and globally. My thanks to Robert Lindsay and Ramasamy (Sam) Uthurusamy as well for serving on my doctoral committee. Sam was a good boss and is a great friend. I thank him for the "tons" of papers that he copied and passed on to me, and for the prolonged discussions on machine learning, pattern recognition, and my research problems. He always read my paper drafts carefully and diligently:

that did not change when it was time to read my dissertation.

I thank my office mates and good friends for many, many interesting and stimulating discussions. Jie Cheng and Zhaogang Qian were always ready with discussions and good pointers. The three of us worked together on coding the initial GID3 package: the SRC yearly reviews always made life hectic. Moncef Maiza provided excellent ideas when it came to mathematics. Discussions with him were always useful, at least to me. Kwang-Ryel Ryu and I shared the same office for the entire duration of our PhD work. He always gave sound advice and insightful discussions. Thanks to Omar Juma, Jaeho Lee, and Wafik Farag for many helpful Unix and computing hints. I thank Neal Rothleder for his sense of humor: it kept us all sane.

Others who were very helpful in this period include: Andrzej Ehrenfeucht for long detailed explanations of theoretical aspects of machine learning; Kevin Compton for help on approaching some theoretical problems; and Paul Scott for serving on my prelim exam committee. I would like to thank Richard Doyle at the Jet Propulsion Lab for his friendship and patience, for providing me with a job the past two years, and for funding many of my conference and workshop trips. Padhraic Smyth gave me great pointers and lots of reading material.

For my financial support, I am very thankful to: The EECS Department for assorted assistantships and fellowships; the Horace Rackham Graduate School for a Rackham Predoctoral Fellowship; Hughes Microelectronics Center for an unrestricted research grant; the DeVlieg Industrial Fellowship; the SRC Research Program for partial support under contract 87-MP-085; and AFOSR for initial support under contract F49620-82-C0089.

I leave to last, thanking those who are dearest to me. To my wife, Kristina, I am grateful for her patience, devotion, technical and nontechnical help, and many careful proofreadings. Life in this "best of all possible worlds" is always pleasant; somehow, she made it even more so. To my father, I am grateful for his patience, support over the years, and for being the ideal role model: he made my pursuing a PhD inevitable. To my mother and sisters: thanks for all the support, love, and good times. Last but not least, to my "brother" and my colleague since fifth grade, Sami Bayyuk: a true gentleman and a distinguished scholar. We have had *countless* technical discussions. His help on this dissertation and on day-to-day life affairs has indeed been very valuable.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

Appendix