

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

© Usama M. Fayyad 1991
All Rights Reserved

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, pre-dawn, 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

DEDICATION	ii
ACKNOWLEDGEMENTS	iii
LIST OF FIGURES	ix
LIST OF TABLES	xii
LIST OF APPENDICES	xiii
CHAPTER	
I. INTRODUCTION	1
1.1 Motivation	5
1.1.1 Earlier Successes	7
1.2 Learning from Examples	8
1.2.1 Multiple Concept Learning	10
1.2.2 The Multiple Concept Learning Problem	13
1.2.3 Zero-Order Classification Rules	14
1.3 Inducing Decision Trees	17
1.4 Research Goals	18
1.5 Organization of this Document	19
II. BACKGROUND	22
2.1 The Bayesian Solution	24
2.1.1 The PAC Learning Model	27
2.2 Pattern Recognition Techniques	28
2.3 Decision Trees	30
2.3.1 Why the Decision Tree Approach?	30
2.3.2 Decision Tree Generation	31
2.3.3 Decision Trees Versus Classification Rules	35
2.3.4 Limitations of the Attribute-Value Pair Representa- tion Scheme	36
2.4 Perceptrons and Neural Networks	38

2.4.1	Problems with the Neural Network Approach	39
2.4.2	The Comprehensibility Issue	42
III.	PERFORMANCE EVALUATION: WHAT IS A BETTER DECISION TREE?	44
3.1	Introduction	45
3.1.1	Inducing Decision Trees From Examples	46
3.1.2	Assumptions	48
3.2	Performance Evaluation Criteria	48
3.3	The Fewer the Leaves, the Better the Tree	51
3.3.1	Binary Decision Trees	53
3.3.2	Counting Decision Trees	61
3.3.3	Relating Number of Leaves to Expected Error Rate	64
3.3.4	Preferring Trees with Fewer Leaves	67
3.4	Discussion	74
3.5	Conclusions	76
IV.	GENERALIZING THE ID3 ALGORITHM	78
4.1	Introduction	79
4.1.1	The ID3 Approach	79
4.2	Problems with the ID3 Approach	83
4.3	Experimenting with an Alternate Approach	87
4.3.1	The GID3 Algorithm	88
4.3.2	Empirical Evaluation	92
4.3.3	Discussion	95
4.4	The GID3* Algorithm	96
4.4.1	Filtering Attribute Values	96
4.4.2	Comments on GID3* Phantomization	101
4.5	Empirical Evaluation of GID3 and GID3*	103
4.5.1	Experiments on Synthetic RIE Data	104
4.5.2	Experiments on Random Synthetic Domains	105
4.5.3	Experiments on Real-World Discrete Data	108
4.6	Concluding Remarks	111
V.	BINARY DISCRETIZATION OF CONTINUOUS-VALUED ATTRIBUTES	114
5.1	Introduction	114
5.2	The Discretization Criterion for Continuous-Valued Attributes	116
5.3	Discussion of the Cut Point Selection Criterion	118
5.4	Cut Points Are Always on Boundaries	122
5.4.1	Support for the Entropy Minimization Heuristic	127
5.4.2	Improving the Efficiency of the Algorithm	128

5.5	Empirical Evaluation	129
5.6	Biased Cut Value Assignment	134
5.7	Concluding Remarks	138
VI. MULTIPLE INTERVAL DISCRETIZATION		139
6.1	Introduction	139
6.2	To Cut or not to Cut? That is the Question	142
6.2.1	The Minimum Description Length Principle	145
6.2.2	Connection to Bayesian Decision Strategy	147
6.3	Applying the MDLP: A Coding Problem	151
6.3.1	Coding the Null Theory <i>NT</i>	151
6.3.2	Coding the Partition <i>HT</i>	152
6.3.3	The Decision Criterion	155
6.4	Deriving Multiple Intervals	157
6.5	Empirical Evaluation	159
6.6	Cut Point Selection Revisited	163
6.7	Concluding Remarks	165
VII. THE ATTRIBUTE SELECTION PROBLEM		167
7.1	Introduction	168
7.1.1	Impurity Measures	168
7.2	On the Suitability of Information Entropy	170
7.3	Binary Decision Trees	177
7.3.1	Empirical Evidence	181
7.4	Properties Desirable In a Selection Measure	182
7.5	Formulating a Selection Measure	184
7.6	Empirical Evaluation	189
7.7	Concluding Remarks	190
VIII. CONCLUSIONS AND FUTURE WORK		193
8.1	Summary of Contributions	193
8.2	Recommendations	196
8.3	Future Work	197
8.3.1	Enriching the Node-Test Language	197
8.3.2	Attribute Costs, Domain Knowledge, and Error Costs	202
8.3.3	Attribute Selection and Establishment of Improvement	203
IX. RELATED MACHINE LEARNING LITERATURE		205
9.1	Induction and Knowledge Acquisition	207
9.1.1	Single Concept Learning	208
9.1.2	The INDUCE Package	210

9.2	Incremental Learning Algorithms	211
9.2.1	Conceptual Clustering	214
9.3	Formal Modelling of Learning from Examples	215
9.4	Learning from Examples and Knowledge Representation . . .	216
9.4.1	Extended Problems, Extended Languages	218
9.4.2	Structured Domains and Use of Domain Knowledge	220
9.5	Decision Tree Pruning	221
APPENDICES		223
BIBLIOGRAPHY		236

LIST OF FIGURES

<u>Figure</u>		
1.1	A Sample Decision Tree	18
2.1	Flowchart of a Skeleton Algorithm for Decision Tree Generation. . .	33
2.2	Two Approximating Decision Trees.	36
3.1	An Example Binarization Operation.	54
3.2	A Decision Tree nT and its Binary Equivalent nT'	56
3.3	A Single Step in the Binarization Algorithm.	58
3.4	Combining Two Trees to Obtain a Tree with n Leaves.	62
3.5	The Space of All Binary Decision Trees.	65
3.6	Number of Training Examples as a Function of the Number of Leaves.	67
3.7	Tree Extension Operation: The Two Trees Are Logically Equivalent.	68
3.8	The P_1 - P_2 Plane Under Uniform Distribution.	73
4.1	An Example Data Set with the "Correct" Classification Rules.	85
4.2	Possible Trees Consistent with Reduced Data (Disregarding attribute E).	85
4.3	The Only Tree Consistent with the Unreduced Training Set.	86
4.4	The Decision Tree Generated by ID3.	87
4.5	Flowchart of the GID3 Attribute Selection Process.	88
4.6	The GID3 Tree Generated for the Sample Data Set.	92

4.7	Two Possible Binary Partitions of a Set of Examples.	98
4.8	Flowchart of Attribute Phantomization in GID3*.	102
4.9	Results Along Measures M2-M4 for the RIE Domain.	104
4.10	Results for Performance Measure M1 for the RIE Domain.	105
4.11	Average Ratios and SD for Measures M2 and M3 in Random Domains.107	
4.12	Average Ratios and SD for Error Rates in Random Domains.	107
4.13	Average Ratios and SD for Measures M4 and M5 in Random Domains.108	
4.14	Results along Measures M1-M4 for the HARR90 Data Set.	110
4.15	Results along Measures M1-M4 for the SOYBEAN Data Set.	111
4.16	Results along Measures M1-M4 for the AUTO Data Set.	112
4.17	Results for Measures M1 and M2 for the MUSHRM Data Set.	113
5.1	A Potential Cut Point that Separates Examples from the Same Class.119	
5.2	Modelling a Possible Partition at $A = T$	124
5.3	Savings in the Actual Number of Evaluations Performed	132
5.4	Actual Speedup Ratios for Empirical Results	133
5.5	Actual Run Times for the Various Data Sets	133
5.6	Biased Versus Unbiased Cut Value Assignment.	135
6.1	A Set of Examples of Four Classes Sorted by A -values.	140
6.2	Recursive Multi-Interval Discretization.	141
6.3	Decision Trees with and without Multiple Interval Discretization.	142
6.4	Ratios of Number of Leaves for the Cut Strategies.	162
6.5	Ratios of Error Rates for the Cut Strategies.	163

7.1	Two Possible Binary Partitions of a Set of Examples from Three Classes.	171
7.2	Two Possible Binary Partitions of a Set of Examples from Five Classes.	172
7.3	Illustration of Insensitivity of the Entropy Measure to Within-Class Fragmentation.	176
7.4	ID3 vs. ID3-BIN in Random RIE Domain (Relative to GID3* performance)	181
7.5	ID3 vs. ID3-BIN for Several Domains (Relative to GID3* performance).	182
7.6	Performance of Various Algorithm (Relative to GID3*) in RIE Domain.	189
7.7	Performance of Various Algorithms (Relative to GID3*) over Several Domains.	190
8.1	A Decision Tree which Branches on Subsets of Values.	199

LIST OF TABLES

Table

1.1	A Simple Training Set of Examples.	10
4.1	GID3 Versus ID3 in the RIE Synthesized Domain	94
4.2	Sample Average Results for One of the Random Domains.	106
4.3	Details of the Data Sets Used for Empirical Evaluation.	109
5.1	Details of the Data Sets Used for Empirical Evaluation.	131
5.2	Error Rates for Biased and Unbiased Cut Value Assignment.	137
6.1	Results of Different Decision Criteria.	161

LIST OF APPENDICES

Appendix

A.	HANDLING PARTIALLY DEFINED ATTRIBUTES	224
A.1	Introduction	224
A.2	Possible Approaches	225
A.2.1	Attribute Evaluation	226
A.2.2	Partitioning Data when Branching	227
A.2.3	Discussion of Various Methods	227
A.3	The Adopted Approach	228
A.3.1	Requirements of the Method	229
B.	ENTROPY AND REFINED PARTITIONS	231