INTELLIGENT CLUSTERING WITH INSTANCE-LEVEL CONSTRAINTS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Kiri Lou Wagstaff

August 2002

INTELLIGENT CLUSTERING WITH INSTANCE-LEVEL CONSTRAINTS

Kiri Lou Wagstaff, Ph.D.
Cornell University 2002

One goal of research in artificial intelligence is to automate tasks that currently require human expertise; this automation is important because it saves time and brings problems that were previously too large to be solved into the feasible domain. Data analysis, or the ability to identify meaningful patterns and trends in large volumes of data, is an important task that falls into this category. Clustering algorithms are a particularly useful group of data analysis tools. These methods are used, for example, to analyze satellite images of the Earth to identify and categorize different land and foliage types or to analyze telescopic observations to determine what distinct types of astronomical bodies exist and to categorize each observation. However, most existing clustering methods apply general similarity techniques rather than making use of problem-specific information.

This dissertation first presents a novel method for converting existing clustering algorithms into *constrained clustering* algorithms. The resulting methods are able to accept domain-specific information in the form of constraints on the output clusters. At the most general level, each constraint is an instance-level statement about a pair of items in the data set that indicates a preference for being placed into the same cluster, or, alternatively, into different clusters. The constrained clustering algorithms developed and presented in this dissertation enforce each constraint according to the strength of that preference.

The second major contribution of this dissertation is the application of constrained clustering algorithms to diverse, significant, challenging real-world problems. We observe that the additional domain knowledge, when combined with the algorithms' ability to enforce that knowledge, produces improvements on a variety of tasks. The problem domains include automated map refinement, natural language processing, and automated data analysis of Hubble Space Telescope observations.

BIOGRAPHICAL SKETCH

Kiri Lou Wagstaff grew up in the small community of Castle Valley, Utah, amid the red rocks, desert sage, and wild freedom from the press of humanity. Her first computer was a Commodore Plus/4, a magical toy that utterly transformed her 11-year-old existence. After graduating from Grand County High School in 1993, Kiri spent four years in the Computer Science department at the University of Utah, earning her B.S. Along the way she developed an intense interest in artificial intelligence. The best of the AI topics, in her view, had to be machine learning - AI programs that could actually get better over time!

Kiri arrived at Cornell University in 1997 and continued in her pursuit of knowledge, getting heavily involved in natural language processing and ballroom dancing. In the summer of 1999, she worked as an intern at the Jet Propulsion Laboratory in Pasadena, California, and the following summer was spent at DaimlerChrysler's Research and Technology Center in Palo Alto, California. She spent her fourth and most productive year research-wise in absentia in Berkeley, California. In the spring of 2002, she put aside all of her extra-curricular interests and settled down to write the manuscript you now have before you. In August, she will start work as a researcher at the Johns Hopkins University Applied Physics Lab, in the Space Department. She looks forward eagerly to the chance to apply her research ideas to support space missions, spacecraft, and the analysis of data collected by their instruments.

ACKNOWLEDGEMENTS

First and foremost, I thank my advisor, Dr. Claire Cardie, for all of her guidance, patience, and instruction in the ways of the research world. Claire is a truly gifted technical writer, and I continue to strive to live up to her high standards. I would like to particularly thank my committee members, Dr. Jim Bell, Dr. Marie desJardins, and Dr. Bart Selman, for seeing me through this long journey and for offering good advice at critical points along the way. Several others contributed to this work. Sugato Basu generously contributed the graphs in Section 3.6 and several useful email exchanges. Dr. Seth Rogers and Dr. Stefan Schroedl collaborated with me on the road map refinement problem. Dr. Jim Bell worked closely with me on the analysis of STIS Mars observations, supplying the data set, offering advice and encouragement, and interpreting results, despite his hectic schedule. I would also like to thank Jascha Sohl-Dickstein, Jason Soderblom, and Eldar Dobrea for their assistance with the STIS data sets and general good cheer.

My informal advisors and mentors have been invaluable. Dr. Pat Langley has been a reliable collaborator and friend since I asked him for a summer job in the spring of 2000. Dr. Dennis DeCoste was an excellent supervisor at JPL and quite exhilarating to work with. Dr. Elaine Weyuker has been my mentor and frequent email correspondant since 1998 and repeatedly offered me priceless advice. I would also like to thank Dr. Chris Johnson at the University of Utah for being an inspiration and a role model. Even more significantly, I thank my parents, David and Lois, for their lifelong guidance and steadfast support in my pursuit of the Ph.D.

Many of my friends have contributed in significant ways to the completion of this work. Dr. Amy McGovern was my thesis buddy and offered a slew of critical insights on the initial drafts of each chapter. This dissertation would truly suffer in the absence of her input. Although she beat me to the title, she won't beat me to Mars! David Kempe shed brilliant light on the formal analysis of my algorithms and tactfully pointed out my errors. Lengthy discussions with Dan Klein about the nature of constrained clustering made me confront several hard questions and to emerge the better for it. Jeff Vinocur presented me with a very thorough proofreading of a nearly-final draft, and his suggestions helped add the final polish it needed. Andy Sfekas and Dr. Kim Cleland were the best friends I could hope for during my final year. They cheered me on in my job interviews and at my defense and didn't complain (too much) when I couldn't come out to play because I had to write. Westley Weimer, friend and aite, was there from the beginning. His critiques and suggestions added significantly to my research progress and the quality of the implementation. His friendship and sense of humor carried me through several difficult times.

There are several institutions that have contributed financially to the research that produced this dissertation. I would like to thank the National Science Foundation for three years of support as a Graduate Research Fellow. I would also like to thank NSF Grant IRI–9624639 and DARPA TIDES Grant N66001-00-C-8009. I also thank the Jet Propulsion Laboratory and DaimlerChrysler's Research and Technology Center for exceedingly productive summer internships.

Finally, I would like to thank astra, stardust, schneller, zinger, and the CFS linux cluster for their tireless dedication to helping me develop code and to running my clustering experiments. In addition, many thanks go to nova, astra's faithful printer sidekick.

ad astra

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES