OMEGA: ON-LINE MEMORY-BASED GENERAL PURPOSE SYSTEM CLASSIFIER

Kan Deng

CMU-RI-TR-98-33

November 1998

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Thesis committee:

Andrew W. Moore (Chair)
Dean Pomerleau
Scott E. Fahlman
Christopher G. Atkeson, Georgia Institute of Technology

ABSTRACT

A *time series* is a sequence of data points in which the order of the data points is important. In many cases, each data point consists of both inputs and outputs. The reason that the time order of such a time series is important may be that at a certain time instant, the outputs are determined not only by the current inputs, but also by some of the more recent inputs and outputs. If we extend the input vector to include those previous inputs and outputs in addition to the current inputs, then the outputs are fully determined by the expanded input vector. Thus, we can transform a time series into a set of data points where the time order is no longer important.

Given a time series, a system classifier's purpose is to determine to which category the underlying system belongs, among a set of pre-defined candidate categories. To do so, our system classification algorithm transforms the time series into a set of expanded data points. It then employs a memory-based classifier to calculate a sequence of probabilities that measure how likely these expanded data points are to belong to each of the categories. Finally, it uses likelihood analysis and hypothesis testing to summarize these classification results. Our method can also handle the classification of non-time series.

Our contributions include: (1) the methodology that decomposes time series classification into the likelihood analysis of a sequence of classifications; (2) a new memory-based classifier that has many desirable properties; (3) re-organization of the memory in the form of a cached kd-tree that greatly improves the computational efficiency of information retrieval and memory-based learning algorithms; and (4) fast feature selection based on intensive cross-validation and greedy searching.

Compared with other methods, our new system classifier is simple to understand, easy to implement, robust for various types of systems, and adaptive to datasets with different densities and/or noise levels. It is capable of distinguishing the various categories of the underlying system without requiring any predefined thresholds. It is efficient not only because it can perform classification quickly, but also because it can focus on the promising categories while ignoring the others after only a few iterations. Based on our empirical evaluations, our method tends to be more accurate than other methods.

DEDICATION

To my family.

To my advisor.

ACKNOWLEDGMENTS

When I write this acknowledgments, I realize that my formal education will end very soon. Reviewing the six years and three months of my graduate study at the Robotics Institute of Carnegie Mellon University, not only has CMU given me a very solid education in computer science and robotics, but also it helped me to build my confidence to pursue career goals after my graduation. However, my experience in CMU is like a medicine: it is good for my health but sometimes it tastes bad.

First, I sincerely thank my advisor, Andrew Moore for his friendship and his patient and intellectual guidance. I feel very lucky to be his student; otherwise, I cannot imagine how I could have finished my Ph.D. study. My other three committee members, Christopher Atkeson, Dean Pomerleau and Scott Falman, were also very helpful. I especially appreciate their tough questions, which pushed me to make my work more solid.

My labmates and classmates, Leeman Baird, Justin Boyan, Steve Chen, Fabio Cozman, Scott Davis, Remi Munos, Michael Nechyba, Andrew Ng, and Jeff Schneider, contributed immeasurably to my research by arguing about the research ideas. I appreciate Michael Nechyba and Andrew Ng for their selfless help which introduced me to the culture of computer science research. Besides, I thank all my friends in CMU's machine learning community, especially: Rich Caruana, Frank Dellaert, Geoff Gordon, Andrew McCallum, Peter Stone, Astro Teller and Belinda Thom.

I am very happy to see that the number of Chinese students in the Robotics Institute has kept growing. A friendly social community is important to everybody, especially for foreign students. Fortunately, our RoboChinese, Heng Cao, Peng Chang, Mei Chen, Mei Han, Yingli Tian, Yanghai Tsin, Huadong Wu, Xingxing Yu, Dongmei Zhang, Li Zhang, Liang Zhao, as well as their families, is such a friendly community.

Almost all the CMU faculty to whom I have talked are very supportive. I want to thank following professors for their help these years: Avrim Blum, Christos Faloutsos, Tom Mitchell, Matt Mason, Roy Maxion, Sebastian Thrun, Manuela Veloso and Yangsheng Xu. Thanks also go to Suzanne Crow, Marie Elm, Carlyn Ludwig, Sandy Rocco, Ruth Wiehagen and Marce Zaragoza.

Since my undergraduate background is not in computer science, without the help of my folks: Yanbin Jia, Tang Lei, Harry Shum, Jiawen Su, Yalin Xiong and Dajun Zeng, I can not imagine how I could have survived in the Robotics Ph.D. program. I also appreciate my former advisor, Katia Sycara, as well as Dundee Navin-Chandra, the other principal researcher in our group, for their support and guidance.

Finally, my special thank goes to my dear wife, Xuemei Gu. A happy family life is the solid foundation of my pursuit of a career.