FAST KERNEL-DENSITY-BASED CLASSIFICATION AND CLUSTERING USING P-TREES

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Anne Margarete Denton

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

March 2003

Fargo, North Dakota

ABSTRACT

Denton, Anne Margarete, M.S., Department of Computer Science, College of Science and Mathematics, North Dakota State University, March 2003. Fast Kernel-density-based Classification and Clustering Using P-trees. Major Professor: Dr. William Perrizo.

Much hardware development effort is dedicated to fast bit-wise operations on large amounts of data, as is necessary for image and video processing. Database and data mining applications do not normally use this processing power. Peano Count Tree-based algorithms are an exception in replacing database scans by AND operations on a compressed bit-vector representation of the data. In this thesis, we show how AND operations on Peano Count Trees, or P-trees, can be implemented; describe an application programmer interface to use them; and develop a variety of data mining algorithms that are based on them. Two types of classification algorithms as well as one clustering algorithm that use ideas from traditional algorithms, adapt them to the P-tree setting, and introduce new improvements are described. All algorithms are fundamentally based on kernel-density estimates that can be seen as a unifying concept for much of the work done in classification and clustering. The two classification algorithms in this thesis differ in their approach to handling data with many attributes. Paper 1 demonstrates means of selecting the most important attributes, thereby reducing the space in which densities have to be evaluated. Paper 2 solves the problem of high dimensionality by using an assumption on independence of attributes. Highly correlated attributes are identified and joined in a novel way. Paper 3 describes a clustering algorithm that combines ideas from three traditional clustering techniques into one P-tree-based method. For all algorithms, we show where they outperform traditional methods.

ACKNOWLEDGMENTS

I would like to thank my adviser, Dr. William Perrizo, for his encouragement and support, and for creating an excellent research environment. Thanks to the other committee members, Dr. D. Bruce Erickson, Dr. Paul Juell, and Dr. Dogan Comez, for their interest in my thesis and for valuable comments they made. Thanks to all DataSURG members for their willingness to explore research ideas together. Last but not least, thanks to Alan, Carl, and Martha whose understanding and support made this thesis possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES