

Preface

Data mining is the science, art and technology of exploring large and complex bodies of data in order to discover useful patterns. Theoreticians and practitioners are continually seeking improved techniques to make the process more efficient, cost-effective and accurate. One of the most promising and popular approaches is the use of decision trees. Decision trees are simple yet successful techniques for predicting and explaining the relationship between some measurements about an item and its target value. In addition to their use in data mining, decision trees, which originally derived from logic, management and statistics, are today highly effective tools in other areas such as text mining, information extraction, machine learning, and pattern recognition.

Decision trees offer many benefits:

- Versatility for a wide variety of data mining tasks, such as classification, regression, clustering and feature selection
- Self-explanatory and easy to follow (when compacted)
- Flexibility in handling a variety of input data: nominal, numeric and textual
- Adaptability in processing datasets that may have errors or missing values
- High predictive performance for a relatively small computational effort
- Available in many data mining packages over a variety of platforms
- Useful for large datasets (in an ensemble framework)

This is the first comprehensive book about decision trees. Devoted entirely to the field, it covers almost all aspects of this very important technique.

The book has twelve chapters, which are divided into three main parts:

- Part I (Chapters 1-3) presents the data mining and decision tree foundations (including basic rationale, theoretical formulation, and detailed evaluation).
- Part II (Chapters 4-8) introduces the basic and advanced algorithms for automatically growing decision trees (including splitting and pruning, decision forests, and incremental learning).
- Part III (Chapters 9-12) presents important extensions for improving decision tree performance and for accommodating it to certain circumstances. This part also discusses advanced topics such as feature selection, fuzzy decision trees, hybrid framework and methods, and sequence classification (also for text mining).

We have tried to make as complete a presentation of decision trees in data mining as possible. However new applications are always being introduced. For example, we are now researching the important issue of data mining privacy, where we use a hybrid method of genetic process with decision trees to generate the optimal privacy-protecting method. Using the fundamental techniques presented in this book, we are also extensively involved in researching language-independent text mining (including ontology generation and automatic taxonomy).

Although we discuss in this book the broad range of decision trees and their importance, we are certainly aware of related methods, some with overlapping capabilities. For this reason, we recently published a complementary book "Soft Computing for Knowledge Discovery and Data Mining", which addresses other approaches and methods in data mining, such as artificial neural networks, fuzzy logic, evolutionary algorithms, agent technology, swarm intelligence and diffusion methods.

An important principle that guided us while writing this book was the extensive use of illustrative examples. Accordingly, in addition to decision tree theory and algorithms, we provide the reader with many applications from the real-world as well as examples that we have formulated for explaining the theory and algorithms. The applications cover a variety of fields, such as marketing, manufacturing, and bio-medicine. The data referred to in this book, as well as most of the Java implementations of the pseudo-algorithms and programs that we present and discuss, may be obtained via the Web.

We believe that this book will serve as a vital source of decision tree techniques for researchers in information systems, engineering, computer

science, statistics and management. In addition, this book is highly useful to researchers in the social sciences, psychology, medicine, genetics, business intelligence, and other fields characterized by complex data-processing problems of underlying models.

Since the material in this book formed the basis of undergraduate and graduates courses at Tel-Aviv University and Ben-Gurion University, it can also serve as a reference source for graduate/advanced undergraduate level courses in knowledge discovery, data mining and machine learning. Practitioners among the readers may be particularly interested in the descriptions of real-world data mining projects performed with decision trees methods.

We would like to acknowledge the contribution to our research and to the book to many students, but in particular to Dr. Barak Chizi, Dr. Shahar Cohen, Roni Romano and Reuven Arbel. Many thanks are owed to Arthur Kemelman. He has been a most helpful assistant in proofreading and improving the manuscript.

The authors would like to thank Mr. Ian Selstrup, Senior Editor, and staff members of World Scientific Publishing for their kind cooperation in connection with writing this book. Thanks also to Prof. H. Bunke and Prof P.S.P. Wang for including our book in their fascinating series in machine perception and artificial intelligence.

Last, but not least, we owe our special gratitude to our partners, families, and friends for their patience, time, support, and encouragement.

Beer-Sheva, Israel
Tel-Aviv, Israel

Lior Rokach
Oded Maimon

October 2007