Preface

Data mining is a very active research area with many successful real-world applications. It consists of a set of concepts and methods used to extract interesting or useful knowledge (or patterns) from real-world datasets, providing valuable support for decision making in industry, business, government, and science.

Although there are already many types of data mining algorithms available in the literature, it is still difficult for users to choose the best possible data mining algorithm for their particular data mining problem. In addition, data mining algorithms have been manually designed; therefore they incorporate human biases and preferences.

This book proposes a new approach to the design of data mining algorithms. Instead of relying on the slow and ad hoc process of manual algorithm design, this book proposes systematically automating the design of data mining algorithms with an evolutionary computation approach. More precisely, we propose a genetic programming system (a type of evolutionary computation method that evolves computer programs) to automate the design of rule induction algorithms, a type of classification method that discovers a set of classification rules from data. We focus on genetic programming in this book because it is the paradigmatic type of machine learning method for automating the generation of programs and because it has the advantage of performing a global search in the space of candidate solutions (data mining algorithms in our case), but in principle other types of search methods for this task could be investigated in the future.

This new approach opens up some exciting avenues for the research and practice of data mining. First of all, once the process of designing a data mining algorithm – normally considered a process requiring a lot of human creativity – has been automated, researchers and practitioners can benefit from a much faster creation of new data mining algorithms. More importantly, the proposed genetic programming system can be used to create rule induction algorithms tailored to the target application domain or the dataset being mined. That is, users are no longer limited to trying to select the best existing algorithm tailored to their data; they can simply ask the computer to automatically generate a new data mining algorithm tailored to their data. It is also interesting to compare automatically designed data mining
algorithms with human-designed ones, since findings derived from this comparison can potentially advance the research related to data mining algorithm design.

This is a research-oriented book, and so it is particularly recommended for researchers and postgraduate students in the areas of data mining and evolutionary computation; but we hope it will also provide some useful ideas for data mining practitioners in general. We also hope this book will stimulate further research in the areas of data mining and evolutionary computation.

The book is organized as follows. First, the Introduction explains the motivation for automating the design of data mining algorithms and presents an overview of the system proposed for this task. Next, the book contains two chapters with an overview of data mining and evolutionary computation methods (Chapters 2 and 3 respectively), as well as a chapter with a discussion of research projects related to the topics of automated algorithm design in data mining and optimization (Chapter 4). These three chapters focus on concepts and methods of data mining, evolutionary computation, and optimization that are particularly useful for a better understanding of the new system proposed in the book. The main contribution of the book, a new genetic programming system to automate the design of rule induction algorithms, is described in detail in Chapter 5, and Chapter 6 reports the results of computational experiments evaluating the effectiveness of the proposed system. Finally, Chapter 7 discusses future directions for this emerging area of automation of the design of data mining algorithms.

The research that led to the writing of this book was carried out mainly at the University of Kent, UK, but the last part of the research (including the writing of the book) was carried out both at Kent and at the Federal University of Minas Gerais (UFMG), Brazil. The authors thank their colleagues in both institutions for creating a productive, friendly research environment and for useful discussions about data mining with evolutionary computation methods.

The first author gratefully acknowledges financial support for this research provided by two Brazilian research support agencies, namely CAPES (process number 165002-5) and FAPEMIG.

Belo Horizonte, Brazil, and Canterbury, UK, February 2009

Gisele L. Pappa
Alex A. Freitas