Data Mining Using Grammar-Based Genetic Programming and Applications
by Man Leung Wong and Kwong Sak Leung
Kluwer Academic Publishers, February 2000
232 pp. Hardbound, ISBN: 0-7923-7746-X
NLG 270.00 / USD 115.00 / GBP 79.50
http://www.wkap.nl/book.htm/0-7923-7746-X

This is an inter-disciplinary book, in the spirit of data mining. It integrates three research areas, namely data mining, genetic programming and inductive logic programming (ILP), into a coherent framework.

In essence, data mining consists of extracting valid, comprehensible, interesting knowledge from data. There are, of course, many kinds of methods that can be used to extract knowledge from data, including mainly methods from machine learning (a branch of artificial intelligence) and statistics. The difference between these areas and data mining is somewhat fuzzy and controversial. There is a significant intersection between data mining, machine learning and statistics. In general, however, data mining researchers emphasize more the importance of discovering comprehensible, interesting knowledge than researchers in other areas. By "interesting" we usually mean knowledge that is novel for the user and potentially useful for intelligent decision making.

ILP has gained increased attention in machine learning and data mining, due to the use of a rule representation language with greater expressiveness power than the one used by most rule discovery algorithms. However, most of the current ILP algorithms are based on local, greedy search. Hence, the integration of the powerful representation language of ILP with the powerful, robust global search performed by evolutionary algorithms seems a promising research direction.

The book contains 9 chapters, including an introduction (Chapter 1) and a conclusion (Chapter 9), and 2 appendices. Chapters 2 through 4 are review chapters aiming at providing the background knowledge required for a better understanding of the other chapters. Chapters 5 to 8 are more advanced chapters introducing the main contribution of the book. Appendix A shows the rule sets discovered in several experiments performed by the authors. Appendix B shows the grammars used to discover rules in two medical data sets.

More precisely, Chapters 2, 3 and 4 present a review of data mining, evolutionary algorithms and inductive logic programming, respectively. In general these chapters are well-written, discussing some important concepts and results produced by research on the corresponding area. However, these chapters are relatively brief, considering the complexity of their themes. Therefore, a certain previous familiarity with the three research areas is definitely a plus for a better understanding of the text.

For instance, Chapter 2 briefly overviews about ten data mining algorithms/methods, including decision-tree building and rule induction algorithms, association-rule discovery algorithms, and some statistical algorithms. This chapter has 17 pages, so that each algorithm can only be described briefly.

Chapter 5 introduces the main contribution of the book, the LOGENPRO (LOgic grammar-based GENetic PROgramming) system, which integrates grammar-based GP and inductive logic programming into a flexible, powerful framework for data mining. This chapter discusses in detail how genetic operators are defined to operate on trees derived from a grammar, so that these operators produce valid offspring – i.e. offspring respecting the grammar.

The next three chapters discuss applications of LOGENPRO. In Chapter 6 LOGENPRO is used to induce decision trees and to learn logic programs. The decision trees it produced were competitive with several algorithms used in the Statlog project (http://www.amsta.leeds.ac.uk/~charles/statlog/). While, for logic programs, LOGENPRO was compared with several variants of FOIL, a well-known inductive logic programming algorithm. Several experiments were run with different amounts of noise introduced into the data being mined. Overall, the results showed that LOGENPRO achieved better classification accuracy than FOIL's variants at most noise levels.

In Chapter 7 LOGENPRO is used to induce IF-THEN prediction rules. Although LOGENPRO successfully learns rules with relatively high classification accuracy, overall the results achieved by the system are not as good as the results achieved by other rule induction algorithms.

It should be noted that in both Chapters 6 and 7 the experiments involved a few small, public domain data sets. Hence, if the reader is interested in data mining applications involving large data sets she will be disappointed.

Chapter 8 is, in our opinion, the most interesting chapter of the book. It discusses the application of LOGENPRO to two real-world medical data sets, namely a fracture data set and a scoliosis data set. The authors show how a grammar can be easily used to specify background knowledge and bias the search towards rules that are more interesting for the user. The discovery of interesting rules is also supported by the use of a rule interestingness measure.

In Chapter 8, LOGENPRO is used for data mining in the sense of extracting valid, comprehensible, interesting knowledge from data. More precisely, Chapter 8 shows that data mining goes beyond maximization of predictive accuracy. Data mining involves the search for comprehensible, interesting rules or patterns, which are really useful for the user. A rule can be very interesting and useful even though its predictive accuracy is not so high. For instance, quoting some results obtained by the authors in the scoliosis data set:

"...the system found rules with confidence factors around 40% to 60%. Nevertheless, the rules ..... show something different in comparison with the rules suggested by the clinicians [i.e. physicians]....After discussion with the domain expert [i.e. a physician who is an expert on scoliosis], it is agreed that the existing rules are not defined clearly enough, and our rules are more accurate than theirs. Our rules provide hints to the clinicians to re-formulate their concepts." (page 166)

Nowadays there are several good books on genetic programming and data mining separately, but so far this is the only book specialized in combining these two areas.

Overall the book is well written. It contains several detailed examples, and its main ideas are discussed in increasing order of complexity. However, as mentioned above, the review chapters are relatively short, considering the complexity of the research areas addressed by the book. Therefore, for a good understanding of the book the reader must have a good background on data mining, genetic programming and inductive logic programming. If the reader is not familiar with any of these areas, he should probably read another book about that (those) area(s) before starting to reading this book.

In addition, the book contains a lot of technical details, including several mathematical and logical formulas. Although the formulas are not very complex, if the reader does not like mathematics nor logic this book will not be easy to follow.

This is a research-oriented book. It contains no exercise for students. It is not the kind of book that you would usually recommend for an undergraduate student. The book is recommended for readers interested in advanced research in both data mining and evolutionary algorithms. The book tends to be particularly useful for Ph.D. students doing research in related areas, as well as to their supervisors. It is also the kind of book that a university library should have.

Acknowledgments: I thank Bill Langdon for his useful comments on an earlier draft of this book review.

Alex A. Freitas
PUC-PR, Postgraduate Program in Computer Science
Rua Imaculada Conceicao, 1155. Curitiba, PR. 80215-901. Brazil
E-mail: alex@ppgia.pucpr.br
Web page: http://www.ppgia.pucpr.br/~alex