Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations

by Ian H. Witten and Eibe Frank

Morgan Kaufmann Publishers, 2000
416 pages, Paper, $49.95
ISBN 1-55860-552-5

Review by:
James Geller, New Jersey Institute of Technology
CS Department, 323 Dr. King Blvd., Newark, NJ 07102
geller@oak.njit.edu
http://web.njit.edu/~geller/

Summary of the book

Witten and Frank's textbook was one of two books that I used for a data mining class in the Fall of 2001. The book covers all major methods of data mining that produce a knowledge representation as output. Knowledge representation is hereby understood as a representation that can be studied, understood, and interpreted by human beings, at least in principle. Thus, neural networks and genetic algorithms are excluded from the topics of this textbook. We need to say "can be understood in principle" because a large decision tree or a large rule set may be as hard to interpret as a neural network.

The book first develops the basic machine learning and data mining methods. These include decision trees, classification and association rules, support vector machines, instance-based learning, Naive Bayes classifiers, clustering, and numeric prediction based on linear regression, regression trees, and model trees. It then goes deeper into evaluation and implementation issues. Next it moves on to deeper coverage of issues such as attribute selection, discretization, data cleansing, and combinations of multiple models (bagging, boosting, and stacking). The final chapter deals with advanced topics such as visual machine learning, text mining, and Web mining.

A walk through the contents

The greatest strength of this Data Mining book lies outside of the book itself. All the algorithms described in this book are implemented and freely available through the WEKA (Waikato Environment for Knowledge Analysis) Website (www.cs.waikato.ac.nz/ml/weka). Chapter 8 of the book is a tutorial to the implemented algorithms. The integration between the book and the Web site is excellent, and the Web site is alive, thriving and growing. Thus, the number of data mining algorithms available on the Web site goes far beyond what is described in the book. Indeed, even Neural Networks have been added to the Web site since the book was first published. While many books offer an associated Web site by now, the close linkage between book and Web site and the rapid growth of the Web site are highly commendable.

Another pleasant feature of the WEKA implementation is that it is done in Java. This makes it possible to construct systems, based on Java, that capitalize on the other strengths of Java, such as access to relational databases through JDBC and easy access to Web pages from within Java programs.

Target audience

The book is written for academics and practitioners and I believe it can be well understood, even by undergraduate students.

In fact, it is probably the most accessible survey of data mining in print, without sacrificing too much of precision and rigor. The book is written in a highly redundant style, which I would like to describe as an exercise in iterative deepening. Basic concepts are repeated in several chapters, but covered to a deeper level in the later chapters. This should make it easy for students to keep reading it, without having to refer back to earlier chapters at every step of the way. On the other hand, for a person that is already familiar with the basics of data mining, this makes boring reading at some places. However, I do not recommend a streamlining of the book. Instead, I recommend that readers with some knowledge of the topic may skip paragraphs that sound familiar without any guilty feelings.

Reviewer's appreciation

The book goes to great lengths to avoid "formula shock". Formulas are developed step-by-step and well explained. Only absolutely necessary formulas are included. In many cases, where the derivation of a complex result is irrelevant to the actual data mining issues, the authors defer to statistics textbooks. While I am greatly in favor of both these approaches in writing textbooks, I feel that they have gone too far at a few places. At a number of places, the authors avoid introducing ``one more letter'' to keep the text readable. However, the price they pay for that is that many of their formulas have no equal signs. Thus, a sentence is terminated with a colon and followed with a formula, which is presumably equal to the quantity described by the sentence. This is done on many pages, e.g., 132--135, 137, 196, 207, 222, etc. Not in my wildest dreams would I have thought that I could ever criticize a book author for having too few formulas and too few variables. But this is exactly what I need to do here. While I do not recommend eliminating the previously mentioned redundancy of description, I do recommend for the next edition (which this book will undoubtedly have) to strengthen the formulas, without necessarily adding new ones.

At a few places, the book could also be improved by adding more explanations to figures. Figure 3.6 is a prime example for this issue. I found myself spending time verifying that instance counts in two subfigures truly add to the same total (of 209). They do. The reader could be spared this effort by a better caption or a better description in the body of the text. Similarly, the Apriori algorithm is introduced in a figure, but only in the "Further Reading" subsection (following much later) is the name of the algorithm mentioned. A better figure caption would help the scholarly advancement of students who might not take the "Further Reading" section that seriously.

In America we say "Actions speak louder than words". Thus, instead of summarizing the book I will describe some actions that I intend to take (or that I am already taking).

(1) I am using WEKA for my research.
(2) If I teach the same course again, I will use Witten and Frank's book again.
(3) If the book appears in a second edition, I will acquire it.