FOREWORD

Finding Profitable Knowledge

The information revolution is generating mountains of data, from sources as diverse as astronomy observations, credit card transactions, genetics research, telephone calls, and web clickstreams. At the same time, faster and cheaper storage technology allows us to store ever-greater amounts of data online, and better DBMS software provides an easy access to those data-bases. The web revolution is also expanding the focus of data mining beyond structured databases to the analysis of text, hyperlinked web pages, images, sounds, movies and other multimedia data.

Mining financial data presents special challenges. For one, the rewards for finding successful patterns are potentially enormous, but so are the difficulties and sources of confusions. The efficient market theory states that it is practically impossible to predict financial markets long-term. However, there is good evidence that short-term trends do exist and programs can be written to find them. The data miners' challenge is to find the trends quickly while they are valid, as well as to recognize the time when the trends are no longer effective.

Additional challenges of financial mining are to take into account the abundance of domain knowledge that describes the intricately inter-related world of global financial markets and to deal effectively with time series and calendar effects. For example, Monday and Friday are known to usually have different effects on S&P 500 than other days of the week.

The authors present a comprehensive overview of major algorithmic approaches to predictive data mining, including statistical, neural networks, rule-based, decision-tree, and fuzzy-logic methods and examine the suitability of these approaches to financial data mining.

They focus especially on relational data mining, which is a learning method able to learn more expressive rules than other symbolic approaches. RDM is thus better suited for financial mining, because it is able to make better use of underlying domain knowledge. Relational data mining also has a better ability to explain the discovered rules -- ability critical for avoiding spurious patterns which inevitably arise when the number of variables examined is very large. The earlier algorithms for relational data mining, also known as ILP -- inductive logic programming, suffer from a well-known inefficiency. The authors introduce a new approach, which combines relational data mining with the analysis of statistical significance of discovered rules. This reduces the search space and speeds up the algorithms. The authors also introduce a set of interactive tools for "mining" the knowledge from the experts. This helps to further reduce the search space.

The authors' grand tour of the data mining methods contains a number of practical examples of forecasting S&P 500 and exchange rates, and allows interested readers to start building their own models. I expect that this book will be a handy reference to many financially inclined data miners, who will find the volume both interesting and profitable.

Gregory Piatetsky-Shapiro

Boston, Massachusetts