Data Mining for Profit

Objective  Data mining techniques will be employed in financial knowledge discovery problems to explore possible calendar effects in stock market, and then statistics methods will be used to evaluate the significance of the patterns that are found by data mining.

Keywords  Data Mining  Stock Market Calendar Effects

Data Mining
- Something both old and new

Data mining is defined as the process of seeking interesting or valuable information within large data sets. The information explosion is generating tremendous mount of data nowadays, like credit card transactions, genetics research, telephone calls, and web pages. Advanced storage technology enables us to store ever-greater amounts of data, and better access to those data. So our focus has moved to exploiting the content, text, hyperlinked web pages, images, sounds, movies and other multimedia data. It is an interdisciplinary subject, representing the confluence of ideas from statistics, exploratory data analysis, machine learning, pattern recognition, database technology, and other disciplines.

![Figure 1](image)

Figure 1 In 2007 the amount of information and storage capacity available

Mining Financial Data
- Hard but promising

Stock price, like many other financial data, is in the form of time series data. Interesting patterns mined from such data could be used for possible trade opportunities. Mining financial data presents special challenges. Trivial data mining algorithms will treat it as tables and forgets the time structure in the series.

---

1 A Forecast of Worldwide Information Growth Through 2010 by International Data Corporation
Finding characteristic patterns especially those seasonal effects in it will remove most non-stationarity and pave the way for further study.

Difficulties of mining financial data could have more fundamental sources. The efficient market theory states that it is practically impossible to predict financial markets for a long-term\textsuperscript{2}. However, in reality short-term trends do exist and computer programs can be written to find them. The challenge is how to find them time and cost efficiently while they are valid.

Not everybody is sold on this idea though and this problem is far away from being solved. Given enough time and data series, it is possible to draw all kinds of patterns out of most datasets. Many are worried about that the published literature is likely to be biased in favor of reporting such results. White and his colleagues examined the performance of trading rules published in 1986 on data from the period 1987-1996, and found that their performance was not significantly different from buy-and-hold strategy. In another paper, White a colleagues examine “calendar effects” that are presumed to predict systematic changes in stock price on specific days of the month of the year. After correcting for the size search space (nearly 9,500 different possible calendar effects) they conclude that no calendar rule appears to be capable of outperforming the benchmark market index over the period 1897-1996\textsuperscript{3}. That is very strong evidence that calendar effects do not exist at all. Eamonn Keogh and Jessica Lin even proved that theoretically it is meaningless to do that\textsuperscript{4}. So there are still a lot of debates going on, and I hope a project on this topic can help me understand it better.

**Calendar Effects**

- Find it or lose it

Calendar effects (seasonal effects in time series) are cyclical anomalies in returns, where the cycle is based on the calendar. Many anomalies and statistically significant predictable patterns in the stock returns have been uncovered in the literature and the most important calendar anomalies are the January effect and the weekend effect. The January effect\textsuperscript{5} (also known as the turn-of-the-year effect or the January anomaly) claims that the returns on common stocks in January are much higher than in other months, and this phenomenon is due to smaller-capitalization stocks in the early days of the month. The weekend effect\textsuperscript{6} (also known as the Monday effect, the day-of-the-week effect or the Monday seasonal) refers to the tendency of stocks to exhibit relatively large returns on Fridays compared to those on Mondays.

\textsuperscript{4} Eamonn Keogh, Jessica LinDOI, 2005, Clustering of time-series subsequences is meaningless: implications for previous and future research, Knowledge and Information Systems, Volume 8, Issue 2, 2005-08-01, Pages 154-177
Some other possible calendar effects are Weekday-of-the-month, Week-of-the-month, Semi-month, Turn-of-the-month, End-of-Year and Holiday-effects, etc.

However, most patterns are not robust and dependable in different sample periods, and many of these patterns, even if they did exist, could disappear in the future, as many of them have already done. Indeed, any truly repetitive and exploitable pattern that can be discovered in the stock market and can be arbitraged away will self-destruct. For example, the January effect became undependable after it received considerable publicity.

**Finding New Patterns**
- No sign of breakthrough

Overall financial markets are efficient in a sense that prices of traded assets already reflect all known information. Here information or news is defined as anything that may affect prices that is unknowable in the present and thus appears randomly in the future. Suppose there is a general tendency for stock prices to under-react to some events, leading to abnormal returns to investors who exploit the lack of full immediate adjustment (many of the predictable patterns that have been discovered may simply be the result of data mining). Quantitative investors will then develop trading strategies to exploit the pattern. But many calendar effects have diminished, disappeared altogether or even reversed since they were discovered. The more potentially profitable a discoverable pattern is, the less likely it is to survive.

How can we find new significant correlations between financial variables or among financial and nonfinancial datasets and avoid spurious results? Well-known and commonly used data mining methods in finance are some attribute-based learning methods such as neural networks, the nearest neighbors’ method, and decision trees. Some non attribute-based learning methods like relational data mining are reported to have more advantages.

**Summary**

Many data mining techniques can be used to discover new calendar effects (or other patterns) in stock market. The challenge comes from both the size of the datasets and stochastic nature of the financial time series. Two important questions need to be addressed: how to identify certain patterns from data efficiently and prove they are significant.

---

7 *The Efficient Market Hypothesis and Its Critics* by Burton G. Malkiel
8 *Data Mining in Finance: Advances in Relational and Hybrid Methods* by Boris Kovalerchuk and Evgenii Vityaev