Financial Time Series Data Mining

Indranil Bose
The University of Hong Kong, Hong Kong

Alvin Chung Man Leung
The University of Hong Kong, Hong Kong

Yiu Ki Lau
The University of Hong Kong, Hong Kong

INTRODUCTION

Movement of stocks in the financial market is a typical example of financial time series data. It is generally believed that past performance of a stock can indicate its future trend and so stock trend analysis is a popular activity in the financial community. In this chapter, we will explore the unique characteristics of financial time series data mining. Financial time series analysis came into being recently. Though the world's first stock exchange was established in the 18^{th} century, stock trend analysis began only in the late 20^{th} century. According to Tay et al. (2003) analysis of financial time series has been formally addressed only since 1980s.

It is believed that financial time series data can speak for itself. By analyzing the data, one can understand the volatility, seasonal effects, liquidity, and price response and hence predict the movement of a stock. For example, the continuous downward movement of the S&P index during a short period of time allows investors to anticipate that majority of stocks will go down in immediate future. On the other hand, a sharp increase in interest rate makes investors speculate that a decrease in overall bond price will occur. Such conclusions can only be drawn after a detailed analysis of the historic stock data. There are many charts and figures related to stock index movements, change of exchange rates, and variations of bond prices, which can be encountered everyday. An example of such a financial time series data is shown in Figure 1. It is generally believed that through data analysis, analysts can exploit the temporal dependencies both in the deterministic (regression) and the stochastic (error) components of a model and can come up with better prediction models for future stock prices (Congdon, 2003).

BACKGROUND

Financial time series are a sequence of financial data obtained in a fixed period of time. In the past, due to technological limitations, data was recorded on a weekly basis. Nowadays, data can be gathered for very short durations of time. Therefore, this data is also called high frequency data or tick by tick data. Financial time series data can be decomposed into several components. Kovalerchuk and Vityaev (2005) defined financial time series data as the summation of long term trends, cyclical variations, seasonal variations, and irregular movements. These special components make financial time series data different from other statistical data like population census that represents the growth trends in the population.

In order to analyze complicated financial time series data, it is necessary to adopt data mining techniques.

Figure 1. A typical movement of a stock index

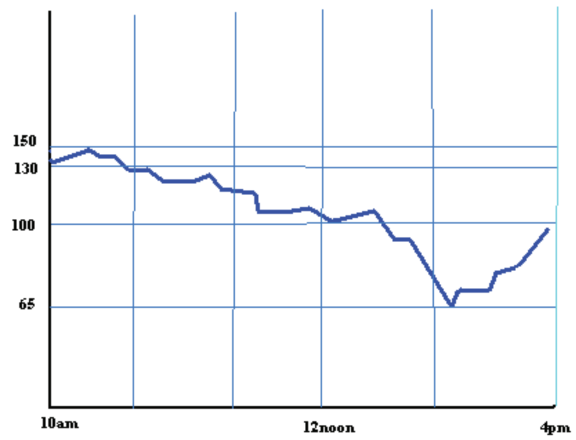

Table 1. Comparison of statistical and machine learning techniques

	Statistical techniques	Machine learning techniques
Advantages	• Relatively simple to use • Uni-variate and multi-variate analysis enable use of stationary and non-stationary models • Users can select different models to fit data and estimate parameters of models	• Easy to build model based on existing data • Computation carried out in parallel to model building which allows real time operations • Able to create own information representation during learning stages • More tolerant to noise in data
Disadvantages	• Performance and accuracy are negatively influenced by noise and non-linear components • The assumption of repeat patterns is unrealistic and may cause large errors in prediction	• Unstable for very large problems • Black box functions often do not provide any explanation of derived results

Currently, the commonly used data mining techniques are either statistics based or machine learning based. Table 1 compares the two types of techniques.

In the early days of computing, statistical models were popular tools for financial forecasting. According to Tsay (2002), the statistical models were used to solve linear time series problems, especially the stationarity and correlation problems of data. Models such as linear regression, autoregressive model, moving average, and autoregressive moving average dominated the industry for decades. However, those models were simple in nature and suffered from several shortcomings. Since financial data is rarely linear, these models were only useful to a limited extent. As a result sophisticated non-linear models like bilinear, threshold autoregression, smoothing transition autoregression, and conditional heteroscedastic autoregressions were developed to address the non-linearity in data. Those models could meet user requirements to a great extent. However, they were restricted in terms of the assumptions made by the models and were severely affected by the presence of noise in the data. It was observed that the correlation coefficient of diversification of a stock portfolio was adversely affected by the linear dependency of time series data (Kantardzic et al., 2004).

In the early 1990s, with the advent of machine learning new techniques were adopted by the field of financial forecasting. The machine learning techniques included artificial neural networks (ANN), genetic algorithms, decision trees, among others. ANNs soon became a very popular technique for predicting options prices, foreign exchange rates, stock and commodity prices, mutual funds, interest rates, treasury bonds etc. (Shen et al., 2005). The attractive features of ANNs were that it was data-driven, did not make any prior assumptions about the data, and could be used for linear and non-linear analysis. Also its tolerance to noise was high.

FINANCIAL TIME SERIES DATA MINING

In order to make useful predictions about future trends of financial time series, it is necessary to follow a number of steps of Knowledge Discovery in Databases (KDD) as shown in Figure 2. We can classify KDD into four important steps, namely, goal identification, preprocessing, data mining, and post-processing (Last et al., 2001). In order to conduct KDD, analysts have to establish a clear goal of KDD and prepare necessary data sets. Secondly, they have to preprocess the data to eliminate noise. Thirdly, it is necessary to transform the format of the data so as to make it amenable for data analysis. In the step of data mining, analysts have to supply a series of training data to build up a model, which can be subsequently tested using testing data. If the error of prediction for the training data does not exceed the tolerance level, the model will be deployed for use. Otherwise, the KDD process will be redone.

Step 1: Goal identification and data creation. To begin with KDD, users need to establish a clear goal

Related Content