ELSEVIER

Expert Systems with Applications 29 (2005) 927–940

Expert Systems with Applications

www.elsevier.com/locate/eswa

The use of data mining and neural networks for forecasting stock market returns

David Enke*, Suraphan Thawornwong

Laboratory for Investment and Financial Engineering, Smart Engineering Systems Lab, Intelligent Systems Center, University of Missouri, Rolla, MO 65409-0370, USA

Abstract

It has been widely accepted by many studies that non-linearity exists in the financial markets and that neural networks can be effectively used to uncover this relationship. Unfortunately, many of these studies fail to consider alternative forecasting techniques, the relevance of input variables, or the performance of the models when using different trading strategies. This paper introduces an information gain technique used in machine learning for data mining to evaluate the predictive relationships of numerous financial and economic variables. Neural network models for level estimation and classification are then examined for their ability to provide an effective forecast of future values. A cross-validation technique is also employed to improve the generalization ability of several models. The results show that the trading strategies guided by the classification models generate higher risk-adjusted profits than the buy-and-hold strategy, as well as those guided by the level-estimation based forecasts of the neural network and linear regression models.

Keywords: Stock return forecasting; Data mining; Information gain; Neural networks; Trading strategies

1. Introduction

Over the past two decades many important changes have taken place in the environment of financial markets. The development of powerful communication and trading facilities has enlarged the scope of selection for investors (Elton and Gruber, 1991). Traditional capital market theory has also changed and methods of financial analysis have improved (Poddig and Rehkugler, 1996). Forecasting stock return or a stock index is an important financial subject that has attracted researchers' attention for many years. It involves an assumption that fundamental information publicly available in the past has some predictive relationships to the future stock returns or indices. The samples of such information include economic variables such as interest rates and exchange rates, industry specific information such as growth rates of industrial production and consumer price, and company specific information such as income statements and dividend yields. This is opposed to the general perception of market efficiency (Fama, 1970). In fact, the efficient market hypothesis states that all available information affecting the current stock values is constituted by the market before the general public can make trades based on it (Jensen, 1978). Therefore, it is impossible to forecast future returns since they already reflect all information currently known about the stocks. This is still an empirical issue because there is considerable evidence that markets are not fully efficient, and it is possible to predict the future stock returns or indices with results that are better than random (Lo and MacKinlay, 1988).

Recently, Balvers, Cosimano, and McDonald (1990), Breen, Glosten, and Jagannathan (1990), Campbell (1987), Fama and Schwert (1977), Fama and French (1988, 1989), Ferson (1989), Keim and Stambaugh (1986), and Schwert (1990) among others, provide evidence that stock market returns are predictable by means of publicly available information such as time-series data on financial and economic variables, especially those with an important business cycle component. These studies identify that such variables as various interest rates, monetary growth rates, changes in industrial production, and inflation rates are statistically important for predicting a portion of the stock returns. However, most of the above studies attempting to

* Corresponding author. Tel.: +573-341-4565; fax: +573-341-6567
E-mail address: enke@umr.edu (D. Enke).

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.06.024

capture the relationship between the available information and the stock returns rely on simple linear regression assumptions. There is no evidence to support the assumption that the relationship between the stock returns and the financial and economic variables is perfectly linear. This is due to the fact that there exists the significant residual variance of the actual stock return from the prediction of the regression equation. Therefore, it is possible that nonlinear models are able to explain this residual variance and produce more reliable predictions of the stock price movements (Mills, 1990; Priestley, 1988).

Since many of the current modeling techniques are based on linear assumptions, a new kind of financial analysis that considers the nonlinear analysis of integrated financial markets needs to be considered. Even though there exists a number of non-linear regression techniques, most of these techniques require that the non-linear model must be specified before the estimation of parameters can be determined. One non-linear modeling technique that may overcome these problems involves the use of neural networks (Hill, O'Connor, and Remus, 1996). In fact, neural networks offer a novel technique that does not require a pre-specification during the modeling process because they independently learn the relationship inherent in the variables. This is especially useful in security investment and other financial areas where much is assumed, and little is known about the nature of the processes determining asset prices (Burrell and Folarin, 1997). Neural networks also offer the flexibility of numerous architecture types, learning algorithms, and validation procedures. As a result, the discovery and use of non-linearity in financial market movements and analysis to produce better predictions of future stock returns or indices has been greatly emphasized by various researchers and financial analysts during the last few years (see Abhyankar, Copeland, and Wong, 1997). Current studies that reflect an interest in applying neural networks to answer future stock behaviors include Chenoweth and Obradovic (1996), Desai and Bharati (1998), Gencay (1998), Leung, Daouk, and Chen (2000), Motiwalla and Wahab (2000), Pantazopoulos et al. (1998), Qi and Maddala (1999), and Wood and Dasgupta (1996).

To this end, it has been found that stock trading driven by a certain forecast with a small forecasting error may not be as profitable as trading guided by an accurate prediction of the sign of stock return (Aggarwal and Demaskey, 1997; Leung et al., 2000; Maberly, 1986; Wu and Zhang, 1997). Nonetheless, having an accurate prediction of a certain stock or stock index return still has numerous benefits. Given the existence of a vast number of articles addressing the predictabilities of stock market return, most of the proposed models relied on various assumptions and often employ a particular series of input variables without justification as to why they were chosen. Obviously, a systematic approach to determine what inputs are important is necessary. In regard to this, the present paper will begin with the discussion of the methodology for data selection and then introduce an information gain data mining technique for performing the variable relevance analysis. Two neural network approaches that can be used for classification and level estimation will also be briefly reviewed in the third section, followed by a discussion of the neural network models, including the generalized regression, probabilistic, and multi-layer feed-forward neural networks that were developed to estimate the value (level) and classify the direction (sign) of excess stock returns on the S&P 500 stock index portfolio. In addition, the five-fold cross validation and early stopping techniques were also implemented in this study to improve the generalization ability of the feed-forward neural networks. The resulting data selection and model development, empirical results, and discussion and conclusion will then be presented, respectively. Finally, the data sources and descriptions are given in the Appendix.

2. Methodology for data selection

In general, large-scale deterministic components, such as trends and seasonal variations, should be eliminated from the inputs since the network will attempt to learn the trend and use it in the prediction (Nelson et al., 1999; Pantazopoulos et al., 1998). Therefore, the data collected in this study, excluding *DIV*, *T1*, *SP*, *DY*, and *ER*, were seasonally adjusted allowing the networks to concentrate on the important details necessary for an accurate prediction (the source and definition of all the variables are given in the Appendix). In addition, due to the lag associated with the publication of macroeconomic indicators as mentioned by Qi and Maddala (1999), certain data, particularly *PP*, *IP*, *CP*, and *M1*, were included in the base set with a two-month time lag while the rest of the variables were included with a one-month time lag. Constructing the data in this manner ensures that the forecasting models using these variables will be similar to real-world practice. Specifically, only observable, but not future data were employed as inputs to the forecasting models. As a result, these time lags were used throughout the experiment to maintain realistic situations when data are gathered.

In this study, the differences $[P_t - P_{t-1}]$ of variables were provided to the networks so that different input variables can be compared in terms of change to the monthly stock returns, since the level changes of the variable may be more meaningful to the models than the original values when forecasting a financial time series. Monthly data from March 1976 to December 1999, for a total of 286 periods and for each of 31 financial and economic variables, were collected and analyzed. These variables, including PP_{t-1}, CP_{t-1}, IP_{t-1}, $M1_{t-1}$, $T3_t$, $T6_t$, $T12_t$, $T60_t$, $T120_t$, $CD1_t$, $CD3_t$, $CD6_t$, AAA_t, BAA_t, DIV_t, $T1_t$, SP_t, DY_t, $TE1_t$, $TE2_t$, $TE3_t$, $TE4_t$, $TE5_t$, $TE6_t$, $DE1_t$, $DE2_t$, $DE3_t$, $DE4_t$, $DE5_t$, $DE6_t$, and $DE7_t$, were primarily employed to predict the level and to classify the sign of the excess stock returns (ER_{t+1}) on the S&P 500