Expert Systems with Applications 39 (2012) 11607–11615

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Forecasting trends of high-frequency KOSPI200 index data using learning classifiers

Youngdoo Son [a], Dong-jin Noh [b], Jaewook Lee [a,*]

[a] Department of Industrial Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-744, South Korea
[b] Department of Industrial and Management Engineering, POSTECH, San 31, Hyoja-dong, Nam-gu, Pohang 790-784, South Korea

ARTICLE INFO

Keywords:
Financial forecasting
Statistical learning
Binary classification
Market lead-lag relationship
High-frequency trading

ABSTRACT

Recently many statistical learning techniques have been applied to the prediction of financial variables. The aim of this paper is to conduct a comprehensive study of the applications of statistical learning techniques to predict the trend of the return of high-frequency Korea composite stock price index (KOSPI) 200 index data using the information from the one-minute time series of spot index, futures index, and foreign exchange rate. Through experiments, it is observed that the spot index change is better predictable with high-frequency time series data and the futures index information significantly improves the prediction accuracy of the return trends of the spot index for high-frequency index data, while the information of exchange rate does not. Also, dimension reduction process before training helps to increase the accuracy and dramatically for some classifiers. In addition, the trained classifiers with which a virtual trading strategy is applied to, noticeable better profits can be achieved than just a buy-and-hold-like strategy.

1. Introduction

The financial variable prediction has been a long and yet active research theme targeted by many researchers since successful prediction helps to make profits as well as avoid risks. There have been many approaches especially for the stock market. Many people, called chartists, have been trying to predict the stock price or some other financial variables by technical analysis. On the other hand, fundamentalists have used some fundamental values of a firm like sales or earnings and their ratios. In 1960s, the structural models like the capital asset allocation model appeared from academia to predict the return of stocks (Sharpe, 1964). According to the well-known efficient market hypothesis it is argued that the stock price is fully random walk without new unpredictable information, making it almost impossible to predict it. There are, however, several counter-evidences that the stock price process does not follow the random walk leaving aside some controversial issues. Two typical such counter-evidences are the momentum effect and the mean reversion which show that the autocorrelations of the return of a stock are positive in short horizons and negative for long horizons.

Inspired by these empirical findings, during the last decades many statistical learning techniques have been applied to predict various financial variables including stock price in many financial markets (Chang, Wang, & Zhou, 2011; Chen, Shih, & Wu, 2006; Gestel et al., 2001; Kara, Boyacioglu, & Baykan, 2011; Kim, 2003; Refenes, Zapranis, & Francis, 1995; Shazly & Shazly, 1999; Steiner & Wittkemper, 1995; Tay & Cao, 2001; Tsibouris & Zeidenberg, 1995; Wittkemper & Steiner, 1996; Zhang & Wu, 1996). Korean stock index was also investigated by several past studies. Chen et al. (2006) compared the performance of support vector machines and back propagation neural networks in forecasting the six major Asian stock markets and got about 0.55 with support vector machines and 0.56 with artificial neural networks in directional accuracy for Korea composite stock price index (KOSPI). Kim (2003) adopted support vector machines to KOSPI index in forecasting Korean stock markets and reported 0.57 in directional accuracy with support vector machines.

There were also many empirical results that the derivative markets led the cash (or spot) markets. Fleming, Ostdiek, and Whaley (1996) found the returns in cost efficient index derivative markets led those in stock markets. For S&P 500 index, Boyle, Byoun, and Park (2002) found the options market led the cash index. For Korean markets, Kang, Lee, and Lee (2006) analyzed Korean KOSPI 200 index and found that both the futures and options markets led the cash index up to ten minutes. With these results, we can notice that the derivative markets lead the cash index. This suggests the use of independent variables from not only the spot index time series data but also the time series of futures, options, and other financial variables to increase the prediction accuracy of spot index or return.

However, the learning approaches above were mostly based on low- or medium-frequency, weekly or daily data. Two main questions to be addressed when we apply statistical learning techniques to high-frequency data are whether the stock price

* Corresponding author.
E-mail address: jaewook@snu.ac.kr (J. Lee).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.04.015

change is better predictable with high-frequency time series data and whether the futures index or foreign exchange rate data can help the prediction of the spot index in high-frequency time window. To address these, we conduct a comprehensive study of the applications of statistical learning techniques to predict the trend of the return of high-frequency KOSPI 200 index data using the information from the one-minute time series of spot index, futures index, and foreign exchange rate. For the statistical learning techniques, we used four state-of-the-art classifiers: linear regression, logistic regression, artificial neural networks, and support vector machines. The independent variables were extracted from the time series of KOSPI 200 index, KOSPI 200 futures index and the exchange rate of Korean Won and US Dollars. These results help to explain *the tail wagging the dog* phenomenon.

The organization of this paper is as follows. In the next section, we introduce previous works on employed statistical learning techniques. In Section 3, we describe the data and experimental procedure with the way of making independent variables from time series that we used. Then, in Section 4, we explain the experimental results by comparing the classifiers with and without dimension reduction and test the simulated returns of them applied to a virtual trading strategy. Finally, in Section 5, we report conclusions and directions for future works.

2. Review of compared classifiers

In this section, we review the compared state-of-the-art classifiers: linear regression, logistic regression, artificial neural networks, and support vector machines. These classification algorithms are briefly explained below.

2.1. Linear regression

Linear regression, one of the simplest data fitting methods, aims at finding parameters β such that the line $y_{pred} = X\beta$ best fits the dataset where X is a dataset matrix (Du, Han, & Chen, 2004). It can be treated as a least square sense minimization problem as follows:

$$\hat{\beta} = \arg\min_{\beta} \|y_{true} - X\beta\|^2 \tag{1}$$

where $\hat{\beta}$ is a parameter vector that minimizes the difference between y_{true}, the true value of the target, and y_{pred}, the predicted value. The solution of Eq. (1) is then given by

$$\hat{\beta} = (X^T X)^{-1} X^T y_{true}. \tag{2}$$

For classification, there should be a target treating criterion unlike the case of regression since the target value is the class taking discrete values. In this paper, the value of y_{ture} is set to be a target vector of training dataset, consisting of two values, $\{1, -1\}$. After finding $\hat{\beta}$ from Eq. (2), we obtain y_{pred} of the test dataset, which is a vector consisting of continuous values. To get the final predicted target of the test dataset, we apply the following procedure as follows:

$$\text{target}_i = \begin{cases} 1 & \text{if } y_{pred,i} > 0, \\ -1 & \text{otherwise.} \end{cases} \tag{3}$$

2.2. Logistic regression

Logistic regression, similarly to the linear regression, aims at finding the best parameters β that solves the following minimization problem.

$$\hat{\beta} = \arg\min_{\beta} \sum_i \|y_{true,i} - f(x_i^T \beta)\|^2. \tag{4}$$

where x_i is the ith instance of the dataset and $f(\cdot)$ called the logistic function is a real function whose domain is the set of real numbers and range is $(0,1)$. A widely used logistic function is given by

$$f(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}, \tag{5}$$

Since the logistic function takes the values in the range of $(0,1)$ and is increasing, it is usually considered as a proxy for probability. For binary classification, it is considered as the probability of one class, say $\mathcal{C}_1$ (denoted by +1) and the probability of the other class, $\mathcal{C}_2$ as $1 - p(\mathcal{C}_1)$ (denoted by -1).

Since Eq. (4) is a nonlinear optimization problem, so there is no closed form solution for Eq. (4) and an iterative method, like *Newton–Rhapson method*, is widely used to find $\hat{\beta}$ (Bishop, 2006). The remaining procedures are almost the same with the linear regression case. With given parameter $\hat{\beta}$, we can find the probability that each test instance belongs to $\mathcal{C}_1$ by calculating $f\left(x_i^T \hat{\beta}\right)$, from which we classify the instances as follows:

$$\text{target}_i = \begin{cases} 1 & \text{if } f\left(x_i^T \hat{\beta}\right) > 0.5, \\ -1 & \text{otherwise.} \end{cases} \tag{6}$$

2.3. Artificial neural networks

Artificial neural networks (ANN) are another widely used classifier which is highly nonlinear. By mimicking a human brain, it consists of perceptrons, each of which has several input values and usually one output value. The output value is determined by a function of input values, called an activation function. The activation function can be a variety of functions such as step function, linear function, sigmoid function and so forth (Grossberg, 1982). In this paper, we used a hyperbolic tangent sigmoid function and a linear function as an activation function. The brief shape of a perceptron is shown on Fig. 1.

In an artificial neural network system, these perceptrons are linked with weighted edges as illustrated in Fig. 2. There are three kinds of layers which contain perceptrons. The input layer is the first layer whose perceptrons get the input values from the instances in a given dataset. The output layer has perceptrons, in many times only one perceptron, which make the overall outputs. Between input layer and output layer, there can be a number of hidden layers. Perceptrons on hidden layer get input values from the previous layer and give output values to the next layer. For supervised learning problems like classification, the weights of edges can be computed with training data using a back-propagation method (Hecht-Nielsen, 1989).

2.4. Support vector machines

Support vector machine (SVM) is a recently emerged nonlinear binary classifier (Boser, Guyon, & Vapnik, 1992; Vapnik, 1995) and successfully extended to regression and clustering problems (Jung,

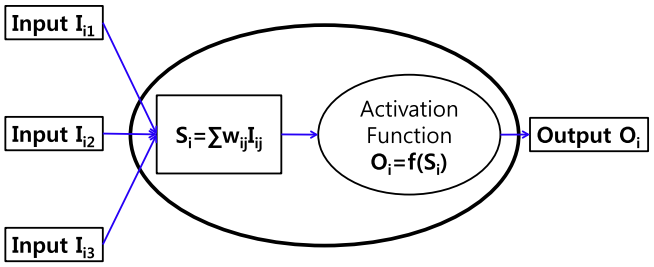

Fig. 1. The shape of a perceptron.