Expert Systems with Applications 42 (2015) 2162–2172

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Predicting stock market index using fusion of machine learning techniques

Jigar Patel, Sahil Shah, Priyank Thakkar *, K Kotecha

Computer Science & Engineering Department, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India

ARTICLE INFO

Article history:
Available online 25 October 2014

Keywords:
Artificial Neural Networks
Support Vector Regression
Random Forest
Stock market
Hybrid models

ABSTRACT

The paper focuses on the task of predicting future values of stock market index. Two indices namely CNX Nifty and S&P Bombay Stock Exchange (BSE) Sensex from Indian stock markets are selected for experimental evaluation. Experiments are based on 10 years of historical data of these two indices. The predictions are made for 1–10, 15 and 30 days in advance. The paper proposes two stage fusion approach involving Support Vector Regression (SVR) in the first stage. The second stage of the fusion approach uses Artificial Neural Network (ANN), Random Forest (RF) and SVR resulting into SVR–ANN, SVR–RF and SVR–SVR fusion prediction models. The prediction performance of these hybrid models is compared with the single stage scenarios where ANN, RF and SVR are used single-handedly. Ten technical indicators are selected as the inputs to each of the prediction models.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction and literature review

Prediction of stock prices is a classic problem. Efficient market hypothesis states that it is not possible to predict stock prices and that stocks behave in random walk manner. But technical analysts believe that most information about the stocks are reflected in recent prices and so if trends in the movements are observed then prices can be easily predicted. In addition, stock market's movements are affected by many macro-economical factors such as political events, firms' policies, general economic conditions, commodity price index, bank rate, bank exchange rate, investors' expectations, institutional investors' choices, movements of other stock markets, psychology of investors, etc. (Miao, Chen, & Zhao, 2007). Value of stock indices are calculated based on stocks with high market capitalization. Various technical parameters are used to gain statistical information from value of stocks prices. Stock indices are derived from prices of stocks with high market capitalization and so they give an overall picture of economy and depends on various factors.

There are several different approaches to time series modeling. Traditional statistical models including moving average, exponential smoothing, and ARIMA are linear in their predictions of the future values (Bollerslev, 1986; Hsieh, 1991; Rao & Gabr, 1984). Extensive research has resulted in numerous prediction applications using Artificial Neural Networks (ANN), fuzzy logic, Genetic Algorithms (GA) and other techniques (Hadavandi, Shavandi, & Ghanbari, 2010b; Lee & Tong, 2011; Zarandi, Hadavandi, & Turksen, 2012). Artificial Neural Networks (ANN) and Support Vector Regression (SVR) are two machine learning algorithms which have been most widely used for predicting stock price and stock market index values. Each algorithm has its own way to learn patterns. Zhang and Wu (2009) incorporated the backpropagation neural network with an Improved Bacterial Chemotaxis Optimization (IBCO). They demonstrated the ability of their proposed approach in predicting stock index for both short term (next day) and long term (15 days). Simulation Results exhibited the superior performance of proposed approach. A combination of data preprocessing methods, genetic algorithms and Levenberg–Marquardt (LM) algorithm for learning feed forward neural networks was proposed in Asadi, Hadavandi, Mehmanpazir, and Nakhostin (2012). They used data preprocessing methods such as data transformation and selection of input variables for improving the accuracy of the model. The results showed that the proposed approach was able to cope with the fluctuations of stock market values and also yielded good prediction accuracy. The Artificial Fish Swarm Algorithm (AFSA) was introduced by Shen, Guo, Wu, and Wu (2011) to train radial basis function neural network (RBFNN). Their experiments on the stock indices of the Shanghai Stock Exchange indicated that RBF optimized by AFSA was an easy-to-use algorithm with considerable accuracy. Ou and Wang (2009) used total ten data mining techniques to predict price movement of Hang Seng index of Hong Kong stock market. The approaches included Linear

* Corresponding author.
E-mail addresses: 10bce067@nirmauni.ac.in (J. Patel), 10bce089@nirmauni.ac.in (S. Shah), priyank.thakkar@nirmauni.ac.in (P. Thakkar), director.it@nirmauni.ac.in (K Kotecha).

http://dx.doi.org/10.1016/j.eswa.2014.10.031
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-nearest neighbor classification, Naive Bayes based on kernel estimation, Logit model, Tree based classification, neural network, Bayesian classification with Gaussian process, Support Vector Machine (SVM) and Least Squares Support Vector Machine (LS-SVM). Experimental results showed that the SVM and LS-SVM generated superior predictive performance among the other models. Hadavandi, Ghanbari, and Abbasian-Naghneh (2010a) proposed a hybrid artificial intelligence model for stock exchange index forecasting. The model was a combination of genetic algorithms and feed forward neural networks.

Recently, the support vector machine (SVM) (Vapnik, 1999) has gained popularity and is regarded as a state-of-the-art technique for regression and classification applications. Kazem, Sharifi, Hussain, Saberi, and Hussain (2013) proposed a forecasting model based on chaotic mapping, firefly algorithm, and Support Vector Regression (SVR) to predict stock market price. SVR–CFA model which was newly introduced in their study, was compared with SVR–GA (Genetic Algorithm), SVR–CGA (Chaotic Genetic Algorithm), SVR–FA (Firefly Algorithm), ANN and ANFIS models and the result showed that SVR–CFA model was performing better than other models. Pai, Lin, Lin, and Chang (2010) developed a Seasonal Support Vector Regression (SSVR) model to forecast seasonal time series data. Hybrid genetic algorithms and tabu search (GA\TS) algorithms were applied in order to select three parameters of SSVR models. They also applied two other forecasting models, Sessional Autoregressive Integrated Moving Average (SARIMA) and SVR for forecasting on the same data sets. Empirical results indicated that the SSVR outperformed both SVR and SARIMA models in terms of forecasting accuracy. By integrating genetic algorithm based optimal time-scale feature extractions with support vector machines, Huang and Wu (2008) developed a novel hybrid prediction model that operated for multiple time-scale resolutions and utilized a flexible nonparametric regressor to predict future evolutions of various stock indices. In comparison with neural networks, pure SVMs and traditional GARCH models, the proposed model performed the best. The reduction in root-mean-squared error was significant. Financial time series prediction using ensemble learning algorithms in Cheng, Xu, and Wang (2012) suggested that ensemble algorithms were powerful in improving the performances of base learners. The study by Aldin, Dehnavr, and Entezari (2012) evaluated the effectiveness of using technical indicators, such as Moving Average, RSI, CCI, MACD, etc. in predicting movements of Tehran Exchange Price Index (TEPIX).

This paper focuses on the task of predicting future values of stock market indices. The predictions are made for 1–10, 15 and 30 days in advance. It has been noticed from the literature that the existing methods for the task under focus in this paper employ only one layer of prediction which takes the statistical parameters as inputs and gives the final output. In these existing methods, the statistical parameters' value of $(t)^{\text{th}}$ day is used as inputs to predict the $(t+n)^{\text{th}}$ day's closing price\value (t is a current day). It is felt that in such scenarios, as the value of n increases, predictions are based on increasingly older values of statistical parameters and thereby not accurate enough. It is clear from this discussions that there is a need to address this problem, and that motivated us about two stage prediction scheme which can bridge this gap and minimize the error stage wise. It was thought that success of the two stage proposed model could really be the significant contribution to the research as the approach can be generalized for other prediction tasks such as whether forecasting, energy consumption forecasting, GDP forecasting etc.

The two stage fusion approach proposed in this paper involves Support Vector Regression (SVR) in the first stage. The second stage of the fusion approach uses Artificial Neural Network (ANN), Random Forest (RF) and SVR resulting into SVR–ANN, SVR–RF and SVR–SVR prediction models. The prediction performance of these hybrid models is compared with the single stage scenarios where ANN, RF and SVR are used single-handedly.

The remainder of the paper is organized into following sections. Section 2 describes single stage approach while focus of the Section 3 is proposed two stage approach. Section 4 addresses experimental results and discussions on these results. Section 5 ends with the concluding remarks.

2. Single stage approach

The basic idea of a single stage approach is illustrated in Fig. 1. It can be seen that for the prediction task of n^{th} day ahead of time, inputs to prediction models are ten technical indicators describing t^{th} day while the output is $(t+n)^{\text{th}}$ day's closing price. These technical indicators which are used as inputs are summarized in Table 1. The prediction models employed are described in the following sub-sections.

2.1. Artificial Neural Networks

Three layer feed forward back propagation ANN as shown in Fig. 2 is employed in this paper. Input layer has ten neurons, one for each of the selected technical parameters. The value of the index which is to be predicted is represented by the only neuron in the output layer. Adaptive gradient descent is used as the weight update algorithm. A tangent sigmoid is used as the transfer function of the neurons of the hidden layer while the neuron in the output layer uses linear transfer function. The output of the model is a continuous value signifying the predicted value of the index.

The reason behind using adaptive gradient descent is to allow learning rate to change during the training process. It may improve the performance of the gradient descent algorithm. In adaptive gradient descent, first, the initial network output and error are calculated. The current learning rate is used to calculate new weights and biases at each epoch. Based on these new weights and biases, new outputs and errors are calculated. If the new error exceeds the old error by more than a predefined ratio (1.04, in this study), the new weights and biases are discarded. Also the learning rate is decreased (to 70% of its current value, in this study). Otherwise, new weights and biases are kept and the learning rate is increased (by 5% of the current value, in the experiments reported in this paper).

The procedure ensures that the learning rate is increased only to the extent that the network can learn without large increases in error. This allows to obtain near optimal learning rate for the local terrain. At the same time, as long as stable learning is assured, learning rate is increased. When it is too high to assure a decrease in error, it is decreased until stable learning resumes.

Number of neurons in the hidden layer and number of epochs are considered as the design parameters of the model. Comprehensive number of experiments are carried out by varying the parameter values as shown in Table 2.

2.2. Support Vector Regression

The SVR uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance ϵ is set in approximation to the SVM. Up until the threshold ε, the error is considered 0. However, the main idea is always the same: to minimize error, individualizing the hyper plane which maximizes the margin, considering that part of the error is tolerated (Parrella, 2007).

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات