Expert Systems with Applications 36 (2009) 7947–7951

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression

Sheng-Hsun Hsu [a,*], JJ Po-An Hsieh [b,1], Ting-Chih Chih [c], Kuei-Chu Hsu [a]

[a] Department of Business Administration, Chung Hua University, No. 707, Sec. 2, WuFu Road, Hsinchu, Taiwan
[b] Department of Management and Marketing, The Hong Kong Polytechnic University, Hong Kong
[c] Department of Information Management, Chung Hua University, Taiwan

ARTICLE INFO

Keywords:
Stock price prediction
Support vector machine
Self-organizing map

ABSTRACT

Stock price prediction has attracted much attention from both practitioners and researchers. However, most studies in this area ignored the non-stationary nature of stock price series. That is, stock price series do not exhibit identical statistical properties at each point of time. As a result, the relationships between stock price series and their predictors are quite dynamic. It is challenging for any single artificial technique to effectively address this problematic characteristics in stock price series. One potential solution is to hybridize different artificial techniques. Towards this end, this study employs a two-stage architecture for better stock price prediction. Specifically, the self-organizing map (SOM) is first used to decompose the whole input space into regions where data points with similar statistical distributions are grouped together, so as to contain and capture the non-stationary property of financial series. After decomposing heterogeneous data points into several homogenous regions, support vector regression (SVR) is applied to forecast financial indices. The proposed technique is empirically tested using stock price series from seven major financial markets. The results show that the performance of stock price prediction can be significantly enhanced by using the two-stage architecture in comparison with a single SVR model.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

Stock price prediction is an important financial subject that has attracted researchers' attention for many years. In the past, conventional statistical methods were employed to forecast stock price. However, stock price series are generally quite noisy and complex. To address this, numerous artificial techniques, such as artificial neural networks (ANN) or genetic algorithms are proposed to improve the prediction results (see Table 1). Recently, researchers are using support vector regressions (SVRs) in this area (see Table 1). SVR was developed by Vapnik and his colleagues (Vapnik, 1995). Most comparison results show that prediction performance of SVR is better than that of ANN (Huang, Nokamori, & Wang, 2005; Kim, 2003; Tay & Cao, 2001a). Reasons that are often cited to explain this superiority include the face that SVRs implement the structural risk minimization principle, while ANNs use the empirical risk minimization principle. The former seeks to minimize the misclassification error or deviation from correct solution of the training data; whereas the latter seeks to minimize the upper bound of generalization error. Solution of SVR may be global optimum while neural network techniques may offer only local optimal solutions. Besides, in choosing parameters, SVRs are less complex than ANNs.

Although researchers have shown that SVRs can be a very useful for stock price forecasting, most studies ignore that stock price series are non-stationary. That is, stock price series do not exhibit identical statistical properties at each point of time and face dynamic changes in the relationship between independent and dependent variables. Such structural changes, which are often caused by political events, economic conditions, traders' expectations and other environmental factors, are an important characteristic of equities' price series. This variability makes it difficult for any single artificial technique to capture the non-stationary property of the data. Most artificial algorithms require a constant relationship between independent and dependent variables, i.e., the data presented to artificial algorithms is generated according to a constant function. One potential solution is to hybridize several artificial techniques. For example, Tay and Cao (2001b) suggest a two-stage architecture by integrating a self-organizing map (SOM) and SVR to better capture the dynamic input–output relationships inherent in the financial data. This architecture was originally proposed by Jacobs, Jordan, Nowlan, and Hinton (1991), who were inspired by the divide-and-conquer principle that is often

* Corresponding author. Tel.: +886 3518 6061.
E-mail addresses: spolo@chu.edu.tw (S.-H. Hsu), JJ.Hsieh@inet.polyu.edu.hk (JJ Po-An Hsieh).
[1] Tel.: +852 2766 7359; fax: +852 2765 0611.

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.10.065

Table 1
Previous research result.

Research	Comparison algorithms	Experimental data	Results
Tay and Cao (2001a)	Back-propagation neural network (BPN) and SVR	Futures and Bonds	SVR forecasts better than the BPN
Kim (2003)	BPN and SVR	Korea composite stock price index	SVR forecasts better than the BPN in terms of movement direction
Huang et al. (2005)	SVR, Linear Discriminant analysis, Quadratic discriminant analysis, and Elman BPN	Japan NIKKEI 225 Index	SVR forecasts better than other techniques in terms of movement direction
Pai and Lin (2005)	Hybrid model of ARIMA and SVM	Ten company stocks	The hybrid model forecasts better than SVR and ARIMA
Wittkemper and Steiner (1996)	Neural networks whose topologies were optimized by genetic algorithm	Sixty-seven German stocks	The new topology can yield good forecast results

used to attack complex problems, i.e., dividing a complex problem into several smaller and simpler problems so that the original problem can be easily solved. In the two-stage architecture, the SOM serves as the "divide" function to decompose the whole financial data into regions where data points with similar statistical distribution are grouped together. After decomposing heterogeneous data into different homogenous regions, SVRs can better forecast the financial indices. Although this architecture is interesting and promising, Tay and Cao (2001b) tested the effectiveness of the architecture only on futures and bonds. Whether the architecture can be employed for stock price prediction remains to be answered.

This study aims to test the effectiveness of the architecture for stock price prediction by comparing the predictive performance of the two-stage architecture with a single SVM technique. Seven stock market indices were used for this study. This paper consists of five sections. Section 2 introduces the basic concept of SVR, SOM and the two-stage architecture. Section 3 describes research design and experiments. Section 4 presents the conclusions.

2. Methodology

2.1. Support vector machine

Support vector machine (SVM) is originated as an implementation of Vapnik's (1995) structural risk minimization (SRM) principle, which reduces empirical risk, based on bounds of generalization error. The fundamental concept in SVM is to transform the data into a higher dimensional space and to find the optimal hyperplane in the space that can maximize the margin between classes. The simplest SVM only deals with a two-class problem, in which the data is separated by a hyperplane defined by a number of *support vectors*. Support vectors are a subset of the training data used to define the boundary between two classes. As a result, support vectors contain all of the information needed to define the classifier. This property makes SVM highly insensitive to the dimensionality of the feature space.

2.2. Support vector regression

Support vector regression is closely related to SVM classifiers in terms of theory and implementation. Vapnik (1995) introduced the ε–insensitive zone in the error loss function. From a theoretical point of view, this zone represents the degree of precision at which the bounds on generalization ability apply. Training vectors that lie within this zone are deemed correct, whereas those outside this zone are deemed incorrect and contribute to the error loss function. These incorrect vectors become the support vectors (see Fig. 1). Vectors lying on and outside the dotted lines are support vectors, whereas those within the ε–insensitive zone are not important in terms of the regression function. The regression surface then can be determined only by support vectors.

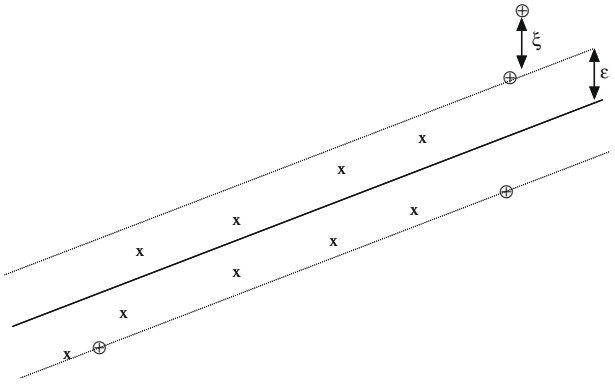

Fig. 1. Approximation function (solid line) of the SVR using an ε–insensitive zone (the area between dotted lines).

Fundamentally, SVR is linear regression in the feature space. Although it is simple and not very useful in real-world situations, it forms a building block for understanding complex SVRs. Detailed discussions of SVMs and SVRs have been given by Burges (1998), Cristianini and Shawe-Taylor (2000), and Smola and Scholkopf (1998).

Given set of training data $\{(x_1,y_1),\ldots,(x_l, y_l)\} \subset X \times R$, where X denotes the space of input patterns. The goal of SVR is to find a function $f(x)$ that deviates not more than ε from the targets y_i for all the training data, and at the same time, is as flat as possible. Let linear function $f(x)$ takes the form:

$$f(x) = w^T x + b \text{ with } w \in X, \quad b \in R \tag{1}$$

Flatness in (1) means smaller $\|w\|$. The problem can then be formulated as

$$\min \quad \tfrac{1}{2}\|w\|^2 \qquad \text{s.t.} \quad \begin{cases} y_i - w^T x_i - b \leqslant \varepsilon \\ w^T x_i + b - y_i \leqslant \varepsilon \end{cases} \tag{2}$$

However, not all problems are linearly separable. To cope with this issue, non-negative slack variables, ξ_i, ξ_i^*, are introduced to deal with the otherwise infeasible constraints of optimization problem (2). The new formation is then stated as

$$\min \quad \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \qquad \text{s.t.} \quad \begin{cases} y_i - w^T x_i - b \leqslant \varepsilon + \xi_i \\ w^T x_i + b - y_i \leqslant \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geqslant 0 \end{cases} \tag{3}$$

The constant C determines the trade-off of error margin between the flatness of $f(x)$ and the amount of deviation in excess of ε that is tolerated. To enable the SVR to predict a nonlinear sit-

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