



A hybrid stock selection model using genetic algorithms and support vector regression

Chien-Feng Huang*

Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan, ROC

ARTICLE INFO

Article history:

Received 28 January 2011
 Received in revised form 1 June 2011
 Accepted 23 October 2011
 Available online 31 October 2011

Keywords:

Stock selection
 Support vector regression
 Genetic algorithms
 Parameter optimization
 Feature selection
 Model validation

ABSTRACT

In the areas of investment research and applications, feasible quantitative models include methodologies stemming from soft computing for prediction of financial time series, multi-objective optimization of investment return and risk reduction, as well as selection of investment instruments for portfolio management based on asset ranking using a variety of input variables and historical data, etc. Among all these, stock selection has long been identified as a challenging and important task. This line of research is highly contingent upon reliable stock ranking for successful portfolio construction. Recent advances in machine learning and data mining are leading to significant opportunities to solve these problems more effectively. In this study, we aim at developing a methodology for effective stock selection using support vector regression (SVR) as well as genetic algorithms (GAs). We first employ the SVR method to generate surrogates for actual stock returns that in turn serve to provide reliable rankings of stocks. Top-ranked stocks can thus be selected to form a portfolio. On top of this model, the GA is employed for the optimization of model parameters, and feature selection to acquire optimal subsets of input variables to the SVR model. We will show that the investment returns provided by our proposed methodology significantly outperform the benchmark. Based upon these promising results, we expect this hybrid GA–SVR methodology to advance the research in soft computing for finance and provide an effective solution to stock selection in practice.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Stock selection has been a challenging and important research area in finance and investment decision-making. This line of research is highly contingent upon reliable prediction of future performance of stocks and successful portfolio construction. Recent advances in computational intelligence and data mining are leading to significant opportunities to solve these problems more effectively. Feasible quantitative models include methodologies stemming from soft computing [1] for prediction of financial time series, multi-objective optimization of expected investment return and risk reduction, and portfolio management – selection of investment instruments based on asset ranking using a variety of input variables and historical data, etc. [2,3]. All these research efforts were in an attempt to facilitate the task of decision-making for investment.

In the research area of stock selection and portfolio optimization, several machine learning methodologies have been developed, including fuzzy systems, artificial neural networks (ANNs), evolutionary algorithms (EAs) as well as support vector

machines (SVMs). Earlier work includes several fuzzy approaches; for instance, Chu et al. [4] used fuzzy multiple attribute decision analysis to select stocks for portfolio construction. Analogously, Zargham and Sayeh [5] employed a fuzzy rule-based system to evaluate a set of stocks for the same purpose. Although these fuzzy approaches denote early efforts in employing computational intelligence for financial applications, they usually lack sufficient learning ability.

Quah and Srinivasan [6] studied an ANN stock selection system to choose stocks that are top-ranked performers. They showed their proposed model outperformed the benchmark model in terms of compounded actual returns overtime. Chapados and Bengio [7] also trained neural networks for estimation and prediction of asset behavior in order to facilitate decision-making in asset allocation. Although these models worked in some applications, they often suffer from the overfitting problem and may tend to fall into a local optimum.

For portfolio optimization, Kim and Han [8] proposed a genetic algorithm (GA) approach to feature discretization and the determination of connection weights for ANNs to predict the stock price index. They suggested that their approach was able to reduce the numbers of attributes and the prediction performance was enhanced. In addition, Caplan and Becker [9] employed genetic programming (GP) to develop a stock ranking model for the high

* Tel.: +886 7 5919798; fax: +886 7 5919514.
 E-mail address: cfhuang15@nuk.edu.tw

technology manufacturing industry in the U.S. More recently, Becker et al. [10] explored various single-objective fitness functions for GP to construct stock selection models for particular investment specifics with respect to risk. In a nutshell, these GP-based models rank stocks from high to low according to a pre-defined objective function.

Because stock market data is highly noisy and complex in dimensionality, it often occurs that most of the aforementioned approaches exhibit inconsistent and unpredictable performance. These challenges arise mainly from the fact that the characteristics and processes of the underlying system that generate time series are generally nonlinear and non-stationary, and for these systems the models solving the relevant applications are usually unknown a priori. An advanced class of novel machine learning algorithms – support vector machines – that improve upon the deficiency of well-known linear techniques for solving these complex applications, was thus developed by Vapnik [11]. As opposed to the traditional empirical risk minimization principle employed by ANNs that minimizes the error on training data, SVMs employ the principle of structural risk minimization that aims to minimize the upper bound of generalization error, and over-fitting is less likely to occur. In general, the optimal solution to SVMs may also be global whereas other neural-network models tend to fall into a local optimal solution. As a result, SVM research thus far has showed that this methodology can outperform other non-linear methods, including neural-network based non-linear prediction, case based reasoning, Linear Discriminant Analysis, Quadratic Discriminant Analysis and Elman Back-propagation Neural Networks [12–15]. In this study, we therefore adopt this methodology for the investment problem investigated here.

Furthermore, even though SVMs have been employed as a popular research methodology in the area of financial applications, most of them focused on the forecast of future direction of either a stock market index or individual stocks [14–18]. Rather than the prediction of financial time series alone, in this study we investigate the task of stock selection using SVMs. This problem is challenging and important in investment, but it is not clear yet how SVMs can be used to advance this research area. Although there exists an earlier attempt using SVMs for this problem by Fan and Palaniswami [19], they solely employed SVMs to classify stocks into winning or losing groups, and this coarse-grained classification procedure usually failed to capture more subtle characteristics of individual stocks. In this study, we will utilize SVMs for regression (support vector regression – SVR) of stock returns, which then serve as surrogates for the actual returns of stocks to imply their quality and relative rankings. Via this improvement, we shall demonstrate SVR as an effective means for stock selection.

However, despite the promising performance of the SVM and SVR in classification and regression, respectively, its success in solving these two problems is highly contingent upon the input variables (features) to the model. Yang and Honavar [20] indicated that several classification issues are determined by the choice of features that describe given patterns presented to a classifier, such as the classification accuracy of the learned classifier, the computational overhead required for learning a classification function, the number of training examples needed for learning, and the cost associated with the features.

The goal of feature selection aims to identify useful, non-redundant subsets of features for a given data mining or machine learning task. By extracting the most essential yet least number of features, one can reduce the computational cost significantly, and construct models that are generalized enough to bring about consistent performance over unseen datasets. Furthermore, since the variables relevant to the SVM/SVR consist of not only the features but also the kernel parameters, it is expected that a successful

model along this line of research shall take into consideration these two issues simultaneously.

In the literature, simultaneous optimization on kernel parameters and feature subsets for SVM-based models has been conducted. Fröhlich et al. [21] first presented a study on this problem for SVM by using the GA, in which feature selection was the main research subject. Huang and Wang [22] then presented a different version for this sort of simultaneous optimization and showed that the classification accuracy of their proposed SVM can be improved for several UCI datasets [23]. Due to these promising results, in this stock-selection study, we thus propose to employ a SVR-based model with a hybrid feature selection and parameter optimization methodology by the GA. In our proposed framework, the task of feature selection depends on the learning algorithm that constructs the SVR model, and our scheme shall be categorized as a wrapper approach [24,25], as opposed to a filter approach. The wrapper approach for feature selection is employed in this study because of its improved performance over the filter approach [22–26]. In essence, the optimization method we adopted here is very similar to that proposed by Huang and Wang [22], yet we will demonstrate our main contribution lies in a proper setup that successfully applied this hybrid methodology to stock selection, which is a new SVR application area.

In a nutshell, the methodology we proposed here is to use the SVR to generate reliable surrogates of actual stock returns for stock rankings. Top-ranked stocks are then chosen for portfolio construction. For the simultaneous optimization on model parameters and feature subsets, we employ the GA for this task. We will report the portfolios constructed by our proposed scheme will substantially outperform the benchmark over the long period of time.

This paper is organized into five sections. Section 2 outlines the methods employed in our study. Section 3 describes the research data used in this study. In Section 4, we describe the experimental design and empirical results are reported and discussed. Section 5 presents the conclusions and future research directions.

2. Methodology

This section first reviews the SVM/SVR theory, followed by the description for our proposed stock selection model. Afterwards, model optimization, including parameter optimization and feature selection, will be performed by the GA. The detailed explanations about the SVM and GA theories may be found in the references listed in this paper.

2.1. Support vector machines

The SVM was first proposed by Vapnik [11], which aims to learn a separate function that divides training instances into distinct groups according to their class labels. By this point of view, SVMs form a class of supervised learning models with main applications to solving problems in classification and regression.

2.1.1. Classification

Through mapping input vectors x into a high-dimensional feature space, SVM models constructed in the new space may represent a linear or nonlinear decision boundary in the original space. In the new space, an optimal separation between instances of distinct classes is achieved by the hyperplane that has the maximal distance to the nearest training instances. As a result, SVMs are known as a methodology that generates the maximum margin hyperplane to provide the maximum separation between distinct classes. The maximum margin hyperplane for a given learning problem is uniquely defined by the instances that are closest to it, and these instances are known as *support vectors*. In addition, the separate function can be linear or nonlinear. In the linearly separable case,

متن کامل مقاله

دریافت فوری ←

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