Journal of Computational Science 17 (2016) 1–13

Contents lists available at ScienceDirect

Journal of Computational Science

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

# Proximal support vector machine based hybrid prediction models for trend forecasting in financial markets

Deepak Kumar, Suraj S. Meghwani, Manoj Thakur*

*Indian Institute of Technology-Mandi, Mandi 175001, Himachal Pradesh, India*

ARTICLE INFO

*Article history:*
Received 10 October 2015
Received in revised form 4 July 2016
Accepted 19 July 2016
Available online 30 July 2016

*Keywords:*
Proximal support vector machines
Random forest
Feature selection
Stock index trend prediction
RReliefF technical indicators

ABSTRACT

In the recent years, various financial forecasting systems have been developed using machine learning techniques. Deciding the relevant input variables for these systems is a crucial factor and their performances depend a lot on the choice of input variables. In this work, a set of fifty-five technical indicators has been considered based on their application in technical analysis as input feature to predict the future (one-day-ahead) direction of stock indices. This study proposes four hybrid prediction models that are combinations of four different feature selection techniques (Linear Correlation (LC), Rank Correlation (RC), Regression Relief (RR) and Random Forest (RF)), with proximal support vector machine (PSVM) classifier. The performance of these models has been evaluated for twelve different stock indices, on the basis of several performance metrics used in literature. A new performance measuring criteria, called joint prediction error (JPE) is also proposed for comparing the results. The empirical results obtained over a set of stock market indices from different international markets show that all hybrid models perform better than the individual PSVM prediction model. The comparison between the proposed models demonstrates superiority of RF-PSVM over all other prediction models. Empirical findings also suggest the superiority of a certain set of indicators over other indicators in achieving better results.



## 1. Introduction

Financial markets are driven by various factors, viz. government policies, economic conditions, political issues, trader's expectations, etc. Many factors which are either unidentified or incomprehensible makes the prediction of stock prices a challenging task. Several empirical studies [1–3] have suggested that index prices follow nonlinear dynamic behavior instead of complete random behavior. This further indicates that trends in the stock markets can only be predicted up to a certain limits [1].

In the last decade, several machine learning algorithms have been developed and used for stock market trend predictions. Among these algorithms, artificial neural networks (ANNs) and support vector machines (SVMs) [4] are the extensively used algorithms. ANNs are non-parametric, empirical risk minimization based modelling approaches with the capability of approximating any non-linear function up to arbitrary precision, regardless of any prior assumptions about data and input space [5,6]. ANNs have been employed for financial trend prediction in [7–9]. Though, ANNs have been applied to a wide-range of real life problems as learning paradigms, but they are found to have several limitations like non-convex objective function and difficulty in deciding criterion for the number of hidden layers. In addition ANNs also suffer from the over-fitting problem, which arise due to large number of parameters in the model and is a major drawback of empirical risk minimization principle. In financial time series forecasting, Kim and Co [10,11] reported inconsistency in the performance of ANN, arising due to over-fitting problem.

Over-fitting is a major drawback of the empirical risk minimization principle. Due do this drawback of empirical risk minimization based approach, the research in the recent past has been diverted towards another approach called structural risk minimization (SRM) principle which was proposed by Vapnik [12]. Vapnik [13,14] formulated the SRM based classification as an optimization problem which tries to minimize the upper limit of the expected error and found it to be superior than empirical risk minimization. Support vector machine, developed by Cortes and Vapnik [4] (SVM) is a SRM based technique. Mathematically, SVMs are formulated as a quadratic optimization problem which ensures global optimal solution and are found to be better than ANNs [15] as far as the over-fitting problem is concerned. SVMs have been successfully applied for trend predication in financial markets. Several empirical studies [11,10,16] have shown that SVMs provide a promising alternative to ANNs.

* Corresponding author.
*E-mail address:* manojpma@gmail.com (M. Thakur).

http://dx.doi.org/10.1016/j.jocs.2016.07.006
1877-7503/

Deciding input variables or features for trend prediction in the stock market is a challenging task. In literature [17–20], factors like historical price patterns, macroeconomic variables and technical indicators have been applied as inputs to the trend prediction models. Though, none of the feature set is expected to exactly replicate market dynamics which depends upon many other undefined and incomprehensible factors. Nevertheless, deriving features from available quantitative information in market is found to be a promising strategy in trend prediction.

Recently, use of feature selection techniques in wrapper and filter methods is used to improve the computational complexity and interpretability of the models [21,22]. Various feature selection techniques have been employed in combination with SVM for financial forecasting in [23–25,19]. All of the above studies have concluded that two-stage architecture in financial forecasting achieves higher performance as compared to the individual SVMs or other learning paradigms. However, improving the effectiveness and efficiency of financial forecasting still remains the major challenge for the researchers in this area.

This study proposes and compares several two-stage architectures that can predict the next-day direction of the stock markets. These models combine various feature selection methods with Proximal Support Vector Machine (PSVM) [26] as a classifier for twelve stock indices. Four feature selection techniques, viz. Linear correlation (LC), rank correlation (RC), regression relief (RR) and random forest (RF) are used in conjunction with PSVM. Proposed methods are named as RF-PSVM, RR-PSVM, LC-PSVM and RC-PSVM.

PSVM formulation for classification problem can be interpreted as a special case of regularized least squares, which leads to a strongly convex objective function. As compared to other variants of SVMs, e.g., LS-SVM [27] and SVM-Light [28], PSVM is much faster. PSVM classifies data points by allotting the close of two parallel planes that are pushed aside as far as possible. Performance of proposed hybrid models is tested over twelve stock indices on the basis of several performance metrics like precision, recall, testing accuracy, $F_1$ score along with a new proposed metric called joint prediction error (JPE). JPE encapsulates the combined effect of $F_1$ score of stock rise and fall. The empirical study shows that PSVM has better performance in predicting stock rise and stock fall. Based on JPE and other performance metrics the performance of all hybrid models are found to be better than ALL-PSVM (PSVM with all features) and ALL-BPNN (BPNN with all features). Moreover, RF-PSVM outperforms all other hybrid models in nine out of twelve stock indices. This study is limited for predicting index trend, although this work can also be used for individual stocks. Here, our proposed software tool is applied to generate buy and sell signals based on the combination of optimal number of technical indicators.

The rest of this paper is organized as follows. Section 2 gives a brief introduction of feature selection techniques and PSVM used in the current study. Proposed hybrid models are discussed in Section 3. The details about the data used in the current work is reported in Section 4. Performance measures & implementation of prediction models is presented in Section 5. Section 6 reports empirical findings in this study. Conclusion from the current study is drawn in Section 7.

## 2. Methodology

### 2.1. Feature selection

A selection of important features enhances interpretability of the model. It can improve the performance of learning algorithms and also helps in reducing the computational complexity of the model. In this section, various feature selection techniques used in this study for trend prediction of stock indices are briefly discussed.

#### 2.1.1. Correlation criteria

One of the simplest means to draw out relevant features is by using correlation coefficient [29]. The importance score or measure of importance for feature $f$ can be assessed by value of correlation coefficient with an output variable, say $y$. Thus, importance score ($S_f^{LC}$) for feature $f$ is calculated by Pearson's coefficient:

$$S_f^{LC} = \left( \frac{cov(f, y)}{\sqrt{var(f)var(y)}} \right)^2 \tag{1}$$

where $cov(.,.)$ and $var(.)$ denote the covariance and variance respectively. A slightly different correlation criterion for feature selection is Spearman's rank correlation, which does not depend upon the distribution of participating variables as opposed to normality assumption in Pearson's correlation coefficient. Spearman correlation is defined as Pearson correlation of ranks of the variables. Importance scores $S_f^{RC}$ for feature $f$ using Spearman coefficient criterion can be calculated as

$$S_f^{RC} = \left( 1 - \frac{6 \times \sum d_i^2}{n \times (n^2 - 1)} \right) \tag{2}$$

Here $d_i$ is a difference in ranks of feature $f$ and output variable $y$.

#### 2.1.2. Regression Reflief feature selection

From the nearest neighbor algorithm, Relief algorithm was proposed by Kira et al. [30]. Relief algorithm and its extensions are mostly used for the feature pruning task in many regression and classification algorithms [31,32]. In this algorithm, importance score of all features are first initialized with zero. Then, for a randomly selected instance ($X_i$) from training instances, Relief algorithm searches for its $k$ proximal instances of same and opposite class which are denoted by $N_i^+$ and $N_i^-$ respectively. Let $N_i^+[f]$, $N_i^-[f]$, and $X_i[f]$ represent the numerical value of $f$th feature in respective instances, then importance score of feature $f$th, denoted by ($S_f^{RR}$), is updated by the following rule

$$S_f^{RR} = S_f^{RR} + \sum_k \frac{d(X_i[f], N_i^-[f]) - d(X_i[f], N_i^+[f])}{k} \tag{3}$$

where

$$d(X_i[f], N_i^-[f]) = \frac{|X_i[f] - N_i^-[f]|}{\max(f) - \min(f)}, \quad d(X_i[f], N_i^+[f]) = \frac{|X_i[f] - N_i^+[f]|}{\max(f) - \min(f)}$$

$max(f)$ and $min(f)$ are the maximum and minimum value of feature $f$ in training instances. This procedure is repeated for $m$ randomly selected instances from training instances. One of the extension of relief algorithm is Regression Relief Feature (RReliefF) selection (proposed in [33]). Since, target variable in regression problem is continuous, concept of proximal classes instance is not possible. Here, proximity is modelled as the relative distance between predicted values of two instances.

#### 2.1.3. Random forest

Random forest [34] is one of the most popular classification and regression algorithm. It has many advantageous characteristics like better generalization capability, robustness, feature pruning ability and simplicity to do non-linear classification. The principle behind the random forest algorithm is the construction of many unpruned decision trees with each tree using bootstrapped training data. In random forest rather than determining the best split among all the features, only a subset of randomly selected features is considered.