Applied Soft Computing 13 (2013) 947–958

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

Support vector regression with chaos-based firefly algorithm for stock market price forecasting

Ahmad Kazem[a], Ebrahim Sharifi[a], Farookh Khadeer Hussain[b,*], Morteza Saberi[c], Omar Khadeer Hussain[d]

[a] Department of Industrial Engineering, University of Tafresh, Iran
[b] Decision Support and e-Service Intelligence Lab, Quantum Computation and Intelligent Systems, School of Software, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW, Australia
[c] Islamic Azad University, Tafresh Branch, Young Researchers Club, Tafresh, Iran
[d] School of Information Systems, Curtin University, Perth, WA, Australia

ARTICLE INFO

Article history:
Received 29 January 2012
Received in revised form 29 June 2012
Accepted 17 September 2012
Available online 8 October 2012

Keywords:
Support vector regression
Firefly algorithm
Chaotic mapping
Stock market price forecasting

ABSTRACT

Due to the inherent non-linearity and non-stationary characteristics of financial stock market price time series, conventional modeling techniques such as the Box–Jenkins autoregressive integrated moving average (ARIMA) are not adequate for stock market price forecasting. In this paper, a forecasting model based on chaotic mapping, firefly algorithm, and support vector regression (SVR) is proposed to predict stock market price. The forecasting model has three stages. In the first stage, a delay coordinate embedding method is used to reconstruct unseen phase space dynamics. In the second stage, a chaotic firefly algorithm is employed to optimize SVR hyperparameters. Finally in the third stage, the optimized SVR is used to forecast stock market price. The significance of the proposed algorithm is 3-fold. First, it integrates both chaos theory and the firefly algorithm to optimize SVR hyperparameters, whereas previous studies employ a genetic algorithm (GA) to optimize these parameters. Second, it uses a delay coordinate embedding method to reconstruct phase space dynamics. Third, it has high prediction accuracy due to its implementation of structural risk minimization (SRM). To show the applicability and superiority of the proposed algorithm, we selected the three most challenging stock market time series data from NASDAQ historical quotes, namely Intel, National Bank shares and Microsoft daily closed (last) stock price, and applied the proposed algorithm to these data. Compared with genetic algorithm-based SVR (SVR-GA), chaotic genetic algorithm-based SVR (SVR-CGA), firefly-based SVR (SVR-FA), artificial neural networks (ANNs) and adaptive neuro-fuzzy inference systems (ANFIS), the proposed model performs best based on two error measures, namely mean squared error (MSE) and mean absolute percent error (MAPE).

1. Introduction

Stock market price prediction is regarded as one of the most challenging tasks of financial time series prediction. The difficulty of forecasting arises from the inherent non-linearity and non-stationarity of the stock market and financial time series. In the past, Box–Jenkins models [1], such as the autoregressive (AR) model and the autoregressive integrated moving average (ARIMA) model, were proposed to tackle this problem. However, these models were developed based on the assumption that the time series being forecasted are linear and stationary. In recent years, nonlinear approaches have been proposed, such as autoregressive conditional heteroscedasticity (ARCH) [2], generalized autoregressive conditional heteroscedasticity (GARCH) [3], artificial neural networks (ANNs) [4–9], fuzzy neural networks (FNN) [10–13], and support vector regression (SVR) [14–22].

ANN has been widely used for modeling stock market time series due to its universal approximation property [23]. Previous researchers have indicated that ANN, which implements the empirical risk minimization principle in its learning process, outperforms traditional statistical models [4]. However, ANN suffers from local minimum traps and the difficulty of determining the hidden layer size and learning rate [24,25]. By contrast, support vector regression, originally introduced by Vapnik [24,26], has a global optimum and exhibits better prediction accuracy due to its implementation of the structural risk minimization principle which considers both the training error and the capacity of the regression model [25,27]. The main problem with SVR is the determination of its hyperparameters, which requires practitioner experience. Unsuitably chosen kernel functions or hyperparameter settings may lead to significantly poor performance [27–30].

* Corresponding author.
E-mail address: Farookh.Hussain@uts.edu.au (F.K. Hussain).

1568-4946/$ – see front matter. Crown Copyright © 2012 Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2012.09.024

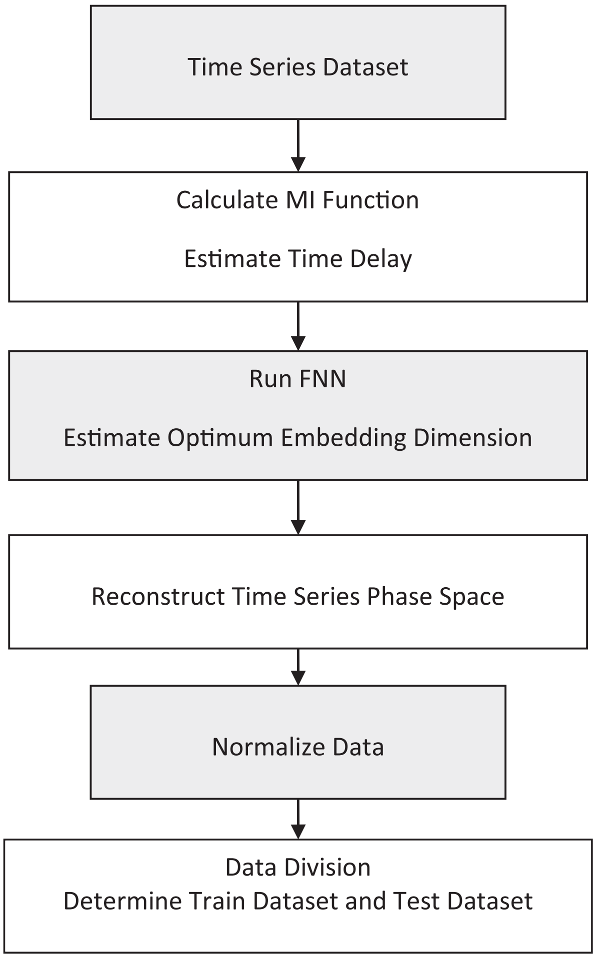

Fig. 1. Data preprocessing procedure.

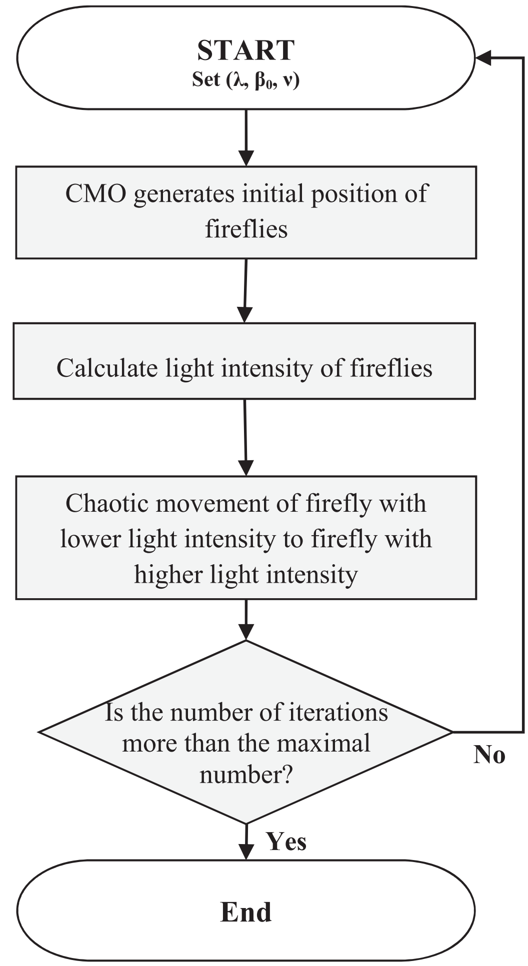

Fig. 2. Chaotic firefly algorithm.

Recently, optimization algorithms such as genetic algorithm (GA) and chaotic genetic algorithm (CGA) have been used to find the best hyperparameters for SVR [31–34].

In this paper, we propose a chaotic firefly algorithm for optimizing the SVR hyperparameters. Results show that our method performs better than SVR-FA, SVR-GA SVR-CGA, ANFIS, ANN and other previous algorithms.

The remainder of this paper is organized as follows. Section 2 introduces new prediction model, including delay coordinate embedding, logistic map, support vector regression and firefly algorithm. Section 3 defines the implementation steps of the proposed model. Section 4 describes the data used in this study and discusses the experimental findings. Conclusions and remarks are given in Section 5.

2. Support vector regression with chaotic firefly algorithm

In this section, we introduce delay coordinate embedding for phase space reconstruction, logistic map, support vector regression and firefly algorithm.

2.1. Delay-coordinate embedding

The analysis of time series generated by non-linear dynamic systems can be done in accordance with Taken's embedding theory [35]. Let univariate time series $\{x_i\}_{i=1}^{N}$, where N is the length of the time series, generated from a d-dimension chaotic attractor, a phase space R^d of the attractor can be reconstructed by using a delay coordinate defined as

$$X_i = (x_i, x_{i-\tau}, \ldots, x_{i-(m-1)\tau}) \quad (1)$$

where m is called the embedding dimension of reconstructed phase space and τ is the time delay constant. Choosing the correct embedding dimension is very important so that we can predict x_{t+1} [36]. Takens [35] considered that the sufficient condition for the embedding dimension is $m \geq 2d+1$. However, too large an embedding dimension needs more observations and complex computation. Moreover, if we choose too large an embedding dimension, noise and other unwanted inputs will be highly embedded with the real source input information, which may corrupt the underlying system dynamic information. Therefore, in accordance with [37], if the dimension of the original attractor is d then an embedding dimension of $m=2d+1$ will be adequate for reconstructing the attractor.

An efficient method of finding the minimal sufficient embedding dimension is the false nearest neighbors (FNN) procedure, proposed by Kennel et al. [38]. Two near points in reconstructed phase space are called false neighbors if they are significantly far apart in the original phase space. Such a phenomenon occurs if we select an embedding dimension lower than the minimal sufficient value and the reconstructed attractor therefore does not preserve the topological properties of the real phase space. In this case, points are projected into the false neighborhood of other points. The idea behind the FNN procedure is as follows. Suppose X_i has a nearest