Knowledge-Based Systems 24 (2011) 66–72

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming

Yi-Shian Lee *, Lee-Ing Tong

Department of Industrial Engineering and Management, National Chiao Tung University, 1001, Ta-Hsuch Rd., Hsinchu 300, Taiwan

ARTICLE INFO

Article history:
Received 15 March 2010
Received in revised form 26 June 2010
Accepted 14 July 2010
Available online 17 July 2010

Keywords:
ARIMA
Hybrid model
Genetic programming
Forecasting
Artificial neural network

ABSTRACT

The autoregressive integrated moving average (ARIMA), which is a conventional statistical method, is employed in many fields to construct models for forecasting time series. Although ARIMA can be adopted to obtain a highly accurate linear forecasting model, it cannot accurately forecast nonlinear time series. Artificial neural network (ANN) can be utilized to construct more accurate forecasting model than ARIMA for nonlinear time series, but explaining the meaning of the hidden layers of ANN is difficult and, moreover, it does not yield a mathematical equation. This study proposes a hybrid forecasting model for nonlinear time series by combining ARIMA with genetic programming (GP) to improve upon both the ANN and the ARIMA forecasting models. Finally, some real data sets are adopted to demonstrate the effectiveness of the proposed forecasting model.

1. Introduction

Many approaches for forecasting time series have been developed. Of conventional statistical methods, the autoregressive integrated moving average (ARIMA) is extensively utilized in constructing a forecasting model. For instance, Kumar and Jain [1] employed ARIMA to develop a model for forecasting traffic-noise time series. Ediger and Akar [2] applied ARIMA model to forecast demand for fuel in Turkey. However, ARIMA cannot be utilized to produce an accurate model for forecasting nonlinear time series. In recent years, the artificial neural network (ANN) and the support vector machines (SVM) have been successfully utilized to develop a nonlinear model for forecasting time series [3–9]. These approaches usually yield better results than the ARIMA model in nonlinear time series. Zhang et al. [10] reviewed forecasting models using ANN for time series.

Since determining whether a linear or nonlinear model should be fitted to a real-world data set is difficult, several investigations have developed some hybrid forecasting models that combine different methods to reduce the forecast error. Zhang [11] developed a hybrid forecasting model that combines ARIMA with ANN to forecast the Canadian lynx time series more accurately than either of the models used separately. Pai and Lin [12] employed a hybrid ARIMA and SVM to construct a model for forecasting stock price. Chen and Wang [13] presented a hybrid seasonal time series ARIMA (SARIMA) and SVM to forecast the production values of the machinery industry in Taiwan. Like Zhang [11], Aladag et al. [14] developed a hybrid model that combined ARIMA and Elman's recurrent neural networks (ERNN) to forecast Canadian lynx time series.

The above hybrid models [11–14] can be employed to combine the linear and nonlinear forecasting system with high overall forecasting accuracy. The hybrid models can be expressed as follows:

$$y_t = L_t + N_t, \tag{1}$$

where y_t represents the original positive time series at time t; L_t represents the linear component, and N_t is the nonlinear component of the model, respectively. The residuals can be obtained using the ARIMA model:

$$r_t = y_t - \hat{L}_t, \tag{2}$$

where r_t is estimated using such nonlinear methods as ANN, SVM, or ERNN. $\hat{L}_t$ is the forecasted value of L_t and is estimated using the ARIMA model. Accordingly, the residual can be rewritten as follows:

$$r_t = f(r_{t-1}, r_{t-2}, \ldots, r_{t-n}) + \varepsilon_t, \tag{3}$$

where $f(r_{t-1}, r_{t-2}, \ldots, r_{t-n})$ represents the nonlinear function that is constructed using ANN, SVM, or ERNN and ε_t is the random error term. The hybrid model for forecasting time series is:

$$\hat{y}_t = \hat{L}_t + \hat{N}_t. \tag{4}$$

Although these hybrid models exhibited favorable overall forecasting performance, the hidden layers in ANN are difficult to

* Corresponding author. Tel.: +886 3 5712121x57356; fax: +886 3 5722392.
E-mail addresses: bill.net.tw@yahoo.com.tw (Y.-S. Lee), litong@cc.nctu.edu.tw (L.-I. Tong).

0950-7051/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2010.07.006

explain and the relationship between the input variables and output variable(s) in ANN or SVM cannot be expressed by a mathematical equation. Furthermore, the ANN model needs large data sets to train a robust network model [15]. Accordingly, this study proposes a novel hybrid model for forecasting time series that combines the ARIMA model with genetic programming (GP). The proposed hybrid model takes the advantages of the ARIMA and GP models in linear or nonlinear modeling and $f(r_{t-1}, r_{t-2}, \ldots, r_{t-n})$ in Eq. (3) can be obtained using GP. Furthermore, unlike ANN, which requires for large data sets to train an appropriate network model, GP can perform well even with small data sets [15]. Thus, the proposed hybrid model can easily be constructed in practice for either large or small data sets. This study is organized as follows. Section 2 describes the procedure of combining the ARIMA and GP model to construct the proposed hybrid model. Section 3 employs some real-world data sets to demonstrate the effectiveness of the proposed method and the proposed method is also compared with other time series forecasting models. Section 4 draws conclusions.

2. The model development

Box and Jenkins presented the ARIMA model in 1970 [16]. The method has been widely used in financial, economic and social scientific fields [17]. In the ARIMA(p, d, q) model, p is the order of auto-regression, d is the order of differencing, and q is the order of the moving average process [16]. Generally speaking, the ARIMA model can be represented as a linear combination of the past observations and past errors as follows:

$$(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)(1-B)^d y_t = \delta+(1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q)\varepsilon_t, \quad t=2,3,\ldots, \tag{5}$$

where y_t is the actual value, B is the backward shift operator, δ is the constant item, ε_t is the random error at time t, ϕ_p and θ_q are the coefficients of the model and can be estimated utilizing the least square method. Furthermore, the model has following setups: model identification, parameter estimation, and modeling diagnosis. The appropriate ARIMA(p, d, q) model is obtained by applying the Akaike Information Criterion (AIC) rule [18,19]. Although the ARIMA model can have high forecasting performance in large or linear data set, it cannot obtain a robust forecasting ability in small or nonlinear data set. Hence, some improving ARIMA models have been proposed to solve the nonlinear or small data [11–14].

Recently, some nonlinear methods such as ANN, SVM, and ERNN are usually utilized to fit nonlinear time series. Both theoretical and empirical analyses have shown that forecasting by a hybrid ARIMA forecasting model that combines two forecasting methods is more accurate than forecasting using just a single forecasting method [11–14]. However, a hybrid forecasting model that is constructed by combining two forecasting methods cannot typically be expressed by a mathematical forecasting equation and needs large data sets to construct the appropriate model. To solve this problem, GP is utilized to fit a nonlinear forecasting time series model.

Koza [20] developed GP as a new algorithm for computer programs that exploits the concept of evolution to solve model structure identification problems and perform symbolic regression [21]. The basic concepts of GP are similar those of genetic algorithms (GAs), and include mutation, crossover and reproduction [22]. Unlike GAs, GP uses the generic parse-tree representation to replace the logic number of the genetic state (0 and 1). Hence, GP has become more popular than conventional linear forecasting methods because it can be employed to search complex nonlinear spaces. Notably, GP is also widely utilized in practical applications such as in a real-time prediction of coastal algal blooms [23], the construction of credit scoring models [15,24], emulating the rainfall-runoff process [25], and forecasting electric power demand [22].

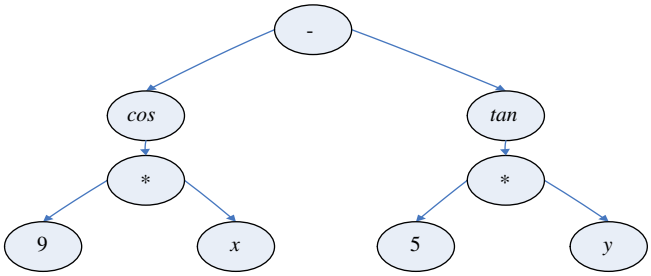

Fig. 1. Example of GP parse-tree representation [15].

Functions or statements in GP have operators ({+, −, ×, ÷, log, and exp}), a trigonometric function ({sin, cos, and tan}), and conditional statements (if, then). Hence, a GP parse tree (Fig. 1) can be applied to a simple example: cos[9x] − tan[5y]. Furthermore, GP system can yield an effective function for predicting the value of the dependent variable. When selecting input variables, GP automatically finds the variables that contribute most to the model [23] and then constructs an equation [22,23,25]. Moreover, GP does not have any restriction on the data size as compared to that of the ANN [15,24].

This study proposes a novel hybrid forecasting model, which combines ARIMA to model the linear component (L_t) of a time series and the GP to model the nonlinear component (N_t), to improve the accuracy of ARIMA forecasting. Since utilizing only linear models or nonlinear models to forecast time series data may not obtain satisfactory results. To improve the forecasting accuracy, a hybrid forecasting system that possesses both linear and nonlinear modeling abilities can be utilized. Moreover, utilizing GP to model the nonlinear component of time series can obtain a mathematical equation than ANN and SVM model no matter data sets are large or small. In practice, the forecasting values utilizing GP can be verified through the mathematical equation. For ANN and SVM models, although the application of these models is easy, the relation between the input and output variables are difficult to explain and cannot verify the forecasting value through the mathematical equation. Therefore, the proposed hybrid approach is as follows:

Step 1. The ARIMA model is utilized to model the linear component of time series. That is, $\hat{L}_t$ is obtained by using the ARIMA model.
Step 2. From Step 1, the residuals from the ARIMA model can be obtained. The residuals are modeled by the GP model in Eq. (3). That is, $\hat{N}_t$ is the forecast value of Eq. (3) by using GP.
Step 3. Using Eq. (4), forecasts of the hybrid model are obtained by adding the forecasted values of linear and nonlinear components, yield in Step 1 and Step 2, respectively.

3. Empirical results

3.1. Data sets

In this study, to demonstrate the effectiveness of the proposed hybrid forecasting model, three data sets are utilized in this study to examine the performance of the proposed hybrid model. Moreover, two literature hybrid models, developed by combining ARIMA and ANN models [11]; and by ARIMA and SVM models [12], are utilized as benchmark models. Through compared with other hybrid ARIMA models, it will be clear to see the forecasting accurately among different hybrid ARIMA models. The first data, the Canadian lynx data, are adopted as an example. The data are the