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:

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

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].

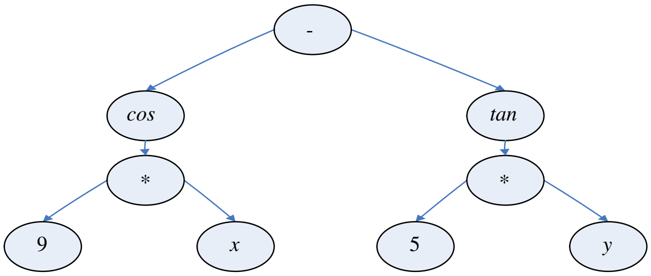

Fig. 1. Example of GP parse-tree representation [15].

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 annual number of trapped lynxes in the Mackenzie River district of Northern Canada from 1821 to 1934. The data set has been analyzed in some of the literature on hybrid time series forecasting models [11,14]. The data that are plotted in Fig. 2 reveal a periodicity of around ten years. In this data, Zhang [11] and Aladag et al. [14] adopted ARIMA and SARIMA to estimate the linear component (L_t), respectively. Hence, to compare different hybrid ARIMA models, this study adopts ARIMA model to estimate the linear component of this time series. The Canadian lynx data are grouped into a training set (100 data observations) and a testing set (last 14 observations).

The second data, the energy consumption data of China are utilized from 1957 to 2007, giving a total of 51 observations (see Fig. 3). The study of energy issue is important for policy makers and related organizations [26,27]. The energy consumption data is regarded as nonlinear and is utilized to demonstrate the effectiveness of nonlinear models. In this data analysis, China's energy consumption data are grouped into a training set (43 data points) and a testing set (last 6 data points). Finally, the US quarterly GDP financial data are utilized from 1947 to 2003, giving 228 data points in the time series (see Fig. 4). In econometric field, the financial data is often used to forecast future trend through time series models. The numbers of training set and testing set about this data are first 210 data points and last 14 data points, respectively. Hence, the three data sets are utilized to evaluate effectiveness of the proposed hybrid forecasting model.

3.2. Results

In this study, all ARIMA modeling is obtained using the SPSS statistical package. The results reveal that the best ARIMA model involved the application of the AIC [18,19]. Besides these hybrid ARIMA models (i.e., ARIMA–ANN model, ARIMA–SVM model, and proposed hybrid model), the ANN model and GP model are also added to forecast the three data sets. Furthermore, in compare with these forecasting models, this study uses three evaluation indices. The first index is root mean square error (RMSE), which compares forecast value with real value. Notably, RMSE is defined as:

$$\mathrm{RMSE} = \sqrt{\sum_{t=1}^{N} \frac{(f_t - o_t)^2}{N}}, \tag{6}$$

where f_t is the forecast value for the tth year, o_t is the real value for the tth year, and N is the number of observations. The second index is mean absolute percentage error (MAPE). This index measures the accuracy of time series data fitted using a statistical method. Notably, MAPE is defined as:

$$\mathrm{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{f_t - o_t}{o_t}\right| \times 100\%. \tag{7}$$

Similar with the MAPE, the third index is mean absolute error (MAE). This index is defined as:

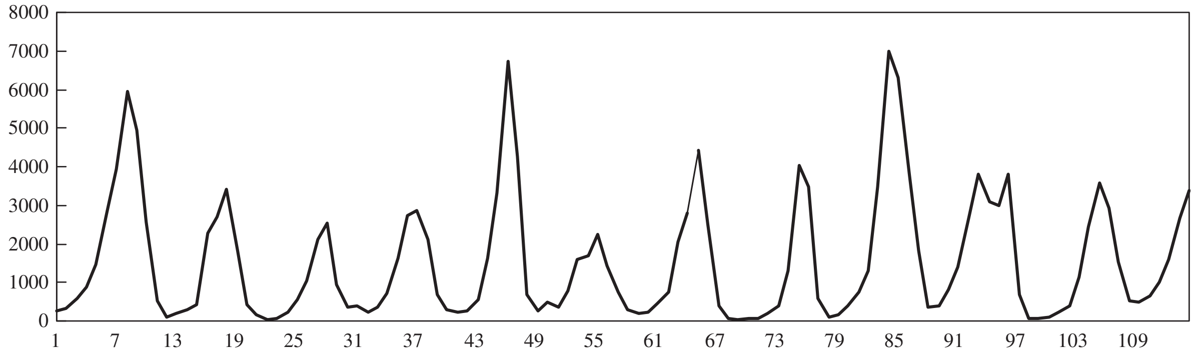

Fig. 2. Canadian lynx data (1821–1934).

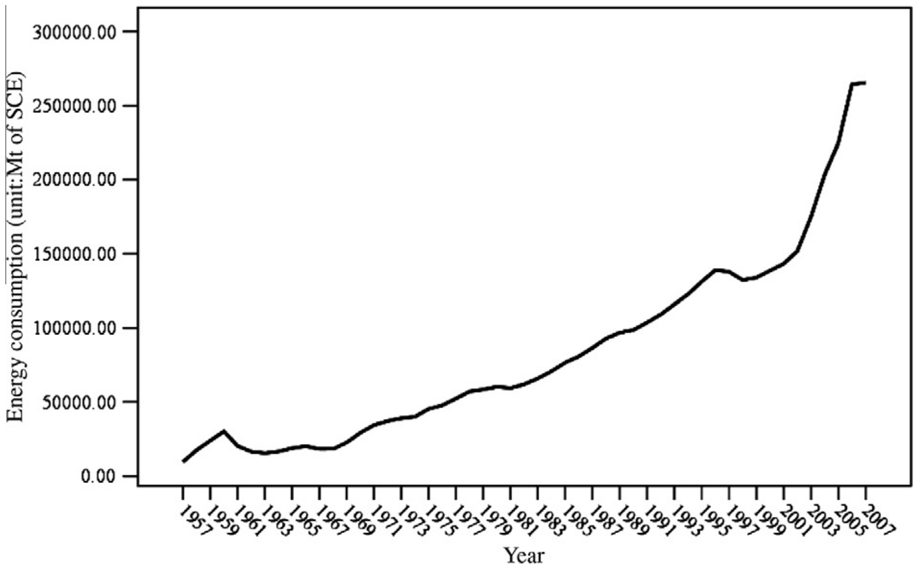

Fig. 3. China's energy consumption data series (1957–2007).

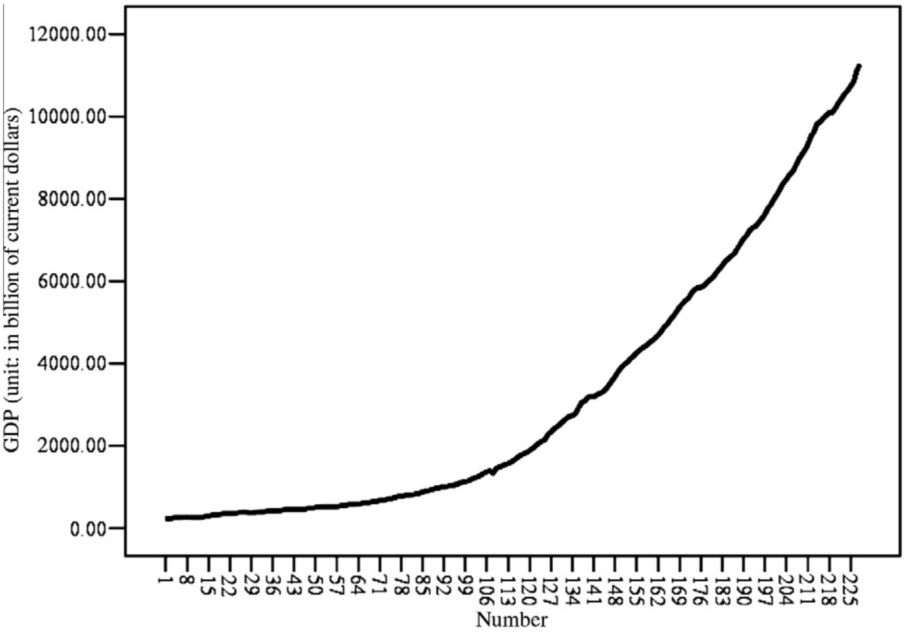

Fig. 4. US quarterly GDP data series (1947/Q1–2003/Q4).

Table 1
The appropriate parameter settings of GP.

Parameter	Value
Population size	100
Maximum number of generation	1000
Crossover rate	0.9
Mutation rate	0.01

$$\mathrm{MAE} = \sum_{t=1}^{N} \frac{|f_t - o_t|}{N}. \tag{8}$$

Notably, the literature about how to determine the residual lagged variables (i.e., $r_{t-1}, r_{t-2}, \ldots, r_{t-n}$) are not consistent. For example, Zhang [11] adopted the given network structure or trial and error to determine an appropriate setting in his analyzed cases; Aladag et al. [14] also adopted trial and error to determine the appropriate residual network model, and Chen and Wang [13] adopted their given residual lagged variables to estimate the residual value. In this study, trial and error is utilized to determine the appropriate forecasting residual model.

In the Canadian lynx data, following other time series studies [11,28], the logarithms (base 10) of the data are utilized to reduce the degree of asymmetry very greatly (a not unusual result with biological observations) in the original data. In order to stabilize the variance and render it stationary, the first-order differencing is utilized. The derived appropriate model, ARIMA(2, 1, 1), satisfies statistical assumptions, according to Box–Pierce and White Tests. In the ANN model, the settings of nodes of input layer, hidden layer, and output layer are 7 × 5 × 1 following Zhang [11]. As for the GP model, the input, output variables are $(y_{t-1}, y_{t-2}, \ldots, y_{t-7})$ and y_t, respectively. Like the ANN modeling, the network structure of nonlinear component of ARIMA–ANN model is 7 × 5 × 1. Similarly with GP, the input variables of nonlinear components of proposed model (ARIMA–GP) are $(r_{t-1}, r_{t-2}, \ldots, r_{t-7})$. Finally, the ARIMA–SVM model must consider the nonlinear component parameter settings (i.e., kernel function type, C, σ^2, and ε). Chen and Wang [13] pointed that no standard procedure exists to determine C, σ^2, and ε parameters. Some studies [12,13,29] utilized Gaussian kernel function type can yield better prediction performances. In this lynx data, the parameters settings are given as (Gaussian kernel function, $C = 10$, $\sigma^2 = 0.413$, and $\varepsilon = 0.5$) can obtain an appropriate forecasting model. In the energy consumption data, the appropriate linear ARIMA model is found to be ARIMA(0, 1, 0). This means that this time series will be stationary situation utilizing first differencing. A neural network of 2 × 1 × 1 is

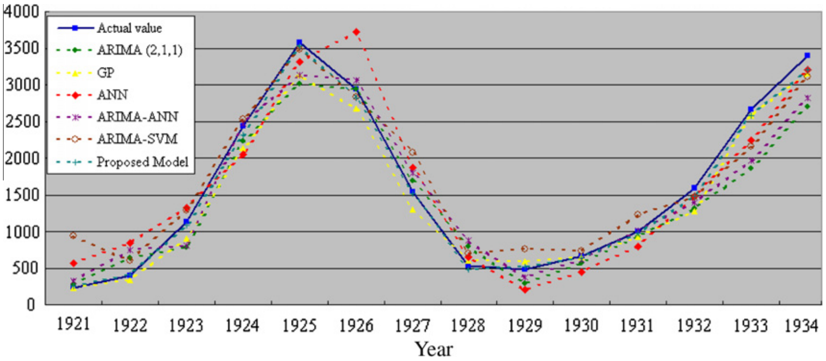

Fig. 5. Actual and fitted values for Canadian lynx data.

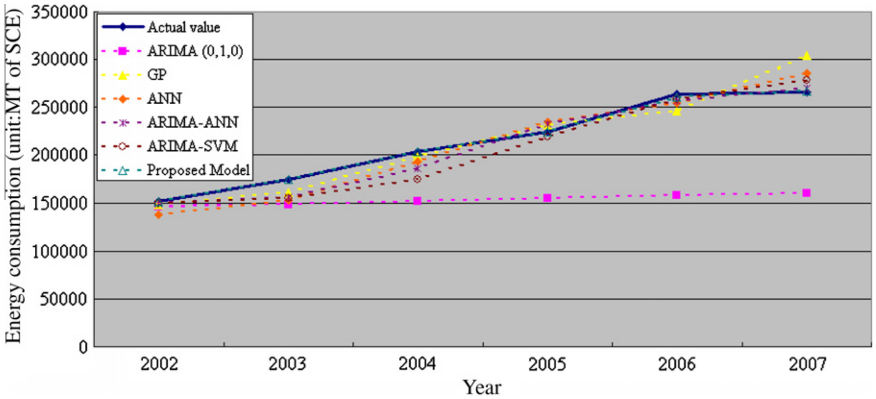

Fig. 6. Actual and fitted values for energy consumption of China.

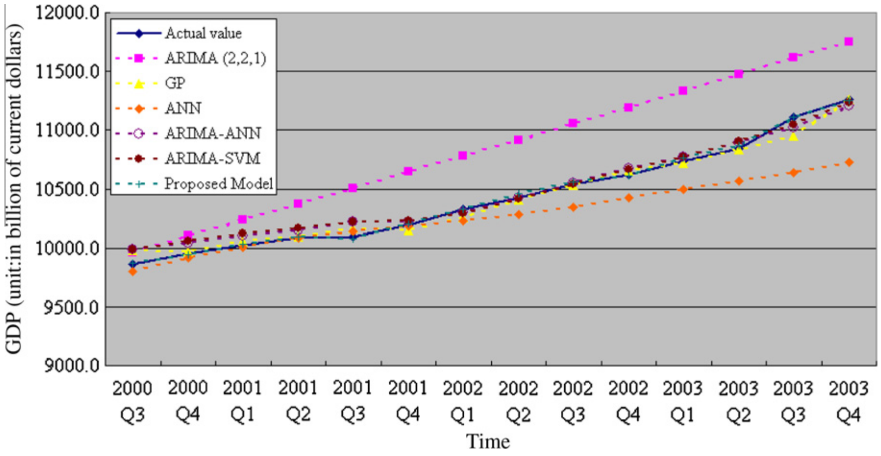

Fig. 7. Actual and fitted values for US quarterly GDP time series.

Table 2
Canadian lynx data forecasting results.

Time	Actual value	ARIMA(2, 1, 1)		GP		ANN		ARIMA–ANN		ARIMA–SVM		Proposed	
		Model value	Error[a]	Model value	Error[a]	Model value	Error[a]	Model value	Error[a]	Model value	Error[a]	Model value	Error[a]
1921	229	278.48	21.61	231.61	1.14	565.50	146.94	319.36	39.46	945.24	312.77	250.52	9.4
1922	399	633.56	58.79	332.75	16.6	843.47	111.40	747.54	87.35	607.51	52.26	410	2.76
1923	1132	785.92	30.57	907.13	19.86	1330.91	17.57	801.64	29.18	1286.00	13.60	1089.24	3.78
1924	2432	2233.54	8.16	2136.79	12.14	2047.87	15.79	2475.45	1.79	2534.00	4.19	2315.67	4.78
1925	3574	3008.92	15.81	3135.55	12.27	3308.55	7.43	3127.43	12.49	3486.70	2.44	3528.14	1.28
1926	2935	2936.89	0.06	2680.17	8.68	3721.21	26.79	3058.85	4.22	2837.00	3.34	2814.59	4.1
1927	1537	1686.05	9.7	1300.68	15.38	1863.11	21.22	1798.99	17.05	2076.50	35.10	1520.67	1.06
1928	529	793.88	50.07	600.76	13.57	646.28	22.17	878.93	66.15	693.00	31.00	480.74	9.12
1929	485	289.56	40.3	586.96	21.02	206.88	57.34	369.56	23.80	755.12	55.69	518.75	6.96
1930	662	564.75	14.69	644.94	2.58	448.20	32.30	613.98	7.25	740.42	11.85	640.25	3.29
1931	1000	929.26	7.07	932.31	6.77	791.46	20.85	1017.07	1.71	1230.12	23.01	980.78	1.92
1932	1590	1313.26	17.41	1277.87	19.63	1485.18	6.59	1398.70	12.03	1475.23	7.22	1480.65	6.88
1933	2657	1851.84	30.3	2580.56	2.88	2241.52	15.64	1969.42	25.88	2153.39	18.95	2578.54	2.95
1934	3396	2698.72	20.53	3157.96	7.01	3207.75	5.54	2814.40	17.13	3104.98	8.57	3205.89	5.6
RMSE		367.62		213.05		347.66		328.42		317.28		80.79	
MAPE (%)		23.22		11.39		36.25		24.67		41.43		4.56	
MAE		282.29		171.69		304.86		259.72		254.11		62.51	

[a] ER $= \frac{|\hat{y}_t - y_t|}{y_t} \times 100\%$.

utilized to model the data set. As for the GP model, the input, output variables are (y_{t-1}, y_{t-2}) and y_t, respectively. Similarly with the settings of nonlinear component of Canadian lynx data, the network structure of nonlinear component of ARIMA–ANN model is $2 \times 1 \times 1$; the input variables of nonlinear components of ARIMA–GP are (r_{t-1}, r_{t-2}). Finally, the ARIMA–SVM model, the parameters settings which are given as (Gaussian kernel function, $C = 3$, $\sigma^2 = 0.4$, and $\varepsilon = 0.3$) can obtain an appropriate forecasting model. In the US quarterly GDP data, the appropriate linear ARIMA model is found to be ARIMA(2, 2, 1). A neural network of $4 \times 3 \times 1$

Table 3
China's energy consumption data forecasting results (unit: MT tons of SCE).

Year	Actual value	ARIMA(0, 1, 0)		GP		ANN		ARIMA–ANN		ARIMA–SVM		Proposed	
		Model value	Error[a]	Model value	Error[a]	Model value	Error[a]	Model value	Error[a]	Model value	Error[a]	Model value	Error[a]
2002	151,797	146038.83	3.79	148076.89	2.45	138193.19	8.96	149206.29	1.71	149761.76	1.34	151678.56	0.08
2003	174,990	148878.67	14.92	160626.89	8.21	152792.78	12.68	155822.05	10.95	154800.65	11.54	174830.86	0.09
2004	203,227	151718.50	25.35	198414.89	2.37	193067.89	5.00	185641.91	8.65	174834.81	13.97	203880.14	0.32
2005	224682	154558.33	31.21	231695.89	3.12	234853.03	4.53	232850.88	3.64	219285.91	2.40	224228.47	0.20
2006	264,270	157398.17	40.44	246368.89	6.77	254192.53	3.81	256609.32	2.90	258743.02	2.09	260415.17	1.46
2007	265,583	160238.00	39.67	304089.89	14.50	285628.84	7.55	270267.10	1.76	278311.98	4.79	267100.83	0.57
RMSE		71652.59		18689.21		15208.71		11766.58		15489.57		1724.14	
MAPE (%)		25.89		6.23		7.09		4.94		6.02		0.45	
MAE		60953.08		14386.2		14375.75		9976.58		12378.14		1126.15	

[a] $\mathrm{ER} = \frac{|\hat{y}_t - y_t|}{y_t} \times 100\%$.

Table 4
US quarterly GDP (unit: in billion of current dollars).

Time	Actual value	ARIMA(2, 2, 1)		GP		ANN		ARIMA–ANN		ARIMA–SVM		Proposed	
		Model value	Error[a]	Model value	Error[a]	Model value	Error[a]	Model value	Error[a]	Model value	Error[a]	Model value	Error[a]
2000/Q3	9862.1	9962.64	1.02	9979.22	1.19	9805.16	0.58	9987.22	1.27	9984.17	1.24	9868.8	0.07
2000/Q4	9953.6	10102.51	1.5	9971.51	0.18	9918.03	0.36	10043.89	0.91	10055.72	1.03	9947.58	0.06
2001/Q1	10024.8	10238.03	2.13	10057.92	0.33	10007.47	0.17	10105.40	0.80	10117.94	0.93	10036.95	0.12
2001/Q2	10088.2	10373.38	2.83	10092.52	0.04	10086.57	0.02	10149.14	0.60	10168.54	0.80	10097.26	0.09
2001/Q3	10096.2	10508.8	4.09	10163.29	0.66	10136.96	0.40	10216.59	1.19	10223.36	1.26	10074.7	0.21
2001/Q4	10193.9	10644.69	4.42	10145.36	0.48	10183.05	0.11	10221.95	0.28	10230.68	0.36	10197.95	0.04
2002/Q1	10329.3	10781.13	4.37	10251.52	0.75	10228.91	0.97	10296.64	0.32	10289.84	0.38	10330.35	0.01
2002/Q2	10428.3	10918.19	4.7	10409.2	0.18	10282.23	1.40	10427.07	0.01	10412.94	0.15	10460.23	0.31
2002/Q3	10,542	11055.86	4.87	10535.26	0.06	10340.89	1.91	10552.40	0.10	10540.08	0.02	10555.29	0.13
2002/Q4	10623.7	11194.16	5.37	10653.76	0.28	10420.64	1.91	10670.14	0.44	10662.82	0.37	10611.8	0.11
2003/Q1	10735.8	11333.09	5.56	10719.6	0.15	10496.41	2.23	10772.97	0.35	10774.78	0.36	10758.69	0.21
2003/Q2	10846.7	11472.66	5.77	10835.74	0.1	10566.12	2.59	10891.70	0.41	10905.02	0.54	10872.51	0.24
2003/Q3	11,107	11612.85	4.55	10945.86	1.45	10635.57	4.24	11016.47	0.82	11045.27	0.56	11106.37	0.01
2003/Q4	11,262	11753.68	4.37	11258.37	0.03	10724.49	4.77	11210.20	0.46	11238.10	0.21	11250.15	0.11
RMSE		448.19		63.09		234.05		69.14		71.02		15.72	
MAPE (%)		3.97		0.42		1.55		0.57		0.59		0.12	
MAE		418.43		43.84		167.33		58.62		60.03		12.77	

[a] $\mathrm{ER} = \frac{|\hat{y}_t - y_t|}{y_t} \times 100\%$.

is utilized to model the data set. As for the GP model, the input, output variables are $(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4})$ and y_t, respectively. The network structure of nonlinear component of ARIMA–ANN model is $4 \times 3 \times 1$; the input variables of nonlinear components of ARIMA–GP are $(r_{t-1}, r_{t-2}, r_{t-3}, r_{t-4})$. Finally, the ARIMA–SVM model, the parameters settings which are given as (Gaussian kernel function, $C = 11$, $\sigma^2 = 0.25$, and $\varepsilon = 0.16$) can obtain an appropriate forecasting model.

To reduce the forecast error of nonlinear component of ARIMA–GP model, the objective function of GP can be expressed as:

$$\text{Minimize}: \sum_{t=1}^{n} |(\hat{r}_t - r_t)|, \tag{9}$$

where r_t represents the actual residual value, and the $\hat{r}_t$ represents the forecasted value of r_t. In the operation of GP, the operators {+, −, ×, ÷, log(base = e), sin, cos, and exp} are adopted. The parameters of the GP model in the three residual data sets from ARIMA model are determined from the lagged values of r_{t-1} to r_{t-n}, and are used to predict r_t. Table 1 presents the appropriate parameter settings for GP utilized to estimate these residual data sets. Following Huang et al. [24], the appropriate setting values of GP are obtained.

These above all time series models are utilized to compare the forecasting accuracy in training set of each data set. Figs. 5–7 plot these forecasting trends among three data sets. Tables 2–4 present the forecasts and errors with all forecasting models in these data sets. From the MAPE index, the proposed model has lower values than other time series models in Canadian lynx data, China's energy data, and US quarterly GDP financial data, respectively (i.e., 4.56%, 0.45%, and 0.12%). These results reveal that the proposed hybrid model has lower model errors than other models. Furthermore, the hybrid models have better forecasting accuracy than utilizing only ARIMA model. This proves that combining different nonlinear method can improve the forecasting performance of utilizing only linear time series model. Compared with other developed hybrid ARIMA models, the proposed improved ARIMA model has highly forecasting accuracy. Moreover, the proposed hybrid forecasting model can display a mathematical form in given period t. For example, the forecasting value of Canadian lynx data in 1921 year can be display as follows:

$$\hat{y}_{1921} = 0.0012 + 1.388y_{1920} - 0.733y_{1919} - 0.998\varepsilon_{1920} + \frac{(r_{1917} - r_{1914}) - r_{1916}}{r_{1914} + \cos(\cos(r_{1917})) + \cos(r_{1914})}, \tag{10}$$

where the Eq. (10) is composed of linear component and nonlinear component. The linear component (i.e., $\hat{L}_{t=1921}$) is obtained by $0.0012 + 1.388y_{1920} - 0.733y_{1919} - 0.998\varepsilon_{1920}$; the nonlinear component (i.e., $\hat{N}_{t=1921}$) is obtained by $\frac{(r_{1917}-r_{1914})-r_{1916}}{r_{1914}+\cos(\cos(r_{1917}))+\cos(r_{1914})}$.

Similarly, the forecasting value of China's energy data in 2002 year can be expressed as follows:

$$\hat{y}_{2002} = 2839.8333 + y_{2001} + \exp^{[\exp^{(\cos(r_{2000}))} + \exp^{(\sin(r_{2000}))}]}. \tag{11}$$

Finally, the forecasting value of US quarterly GDP data in 2000 Q3 can be expressed as follows:

$$\hat{y}_{2000,q3} = 0.6320 - 0.0246y_{2000,q2} + 0.1175y_{2000,q1} - 0.7450\varepsilon_{2000,q2} - 0.1578\varepsilon_{2000,q1} + (r_{2000,q1} - r_{2000,q2}). \quad (12)$$

Through the algorithm of ARIMA–GP model, the forecasting values from the ARIMA–GP model can be obtained in Tables 2–4.

Hence, from the above analysis, it is obviously that the proposed hybrid forecasting model has more accuracy in forecasting the Canadian lynx data, China's energy consumption data, and US quarterly GDP data than the other methods.

4. Conclusions

The traditional statistical forecasting methods can effectively model linear time series, but to accurately forecast nonlinear time series is difficult. Recently, ANN and SVM time series models have been developed to enhance the forecasting accuracy. However, when dealing with the real-world problems, it is not easy to judge whether linear or nonlinear structure is appropriate. In this case, the hybrid methodology can be a valid way to enhance the forecasting performance. This study is motivated by evidence that different forecasting models can complement each other in modeling data sets, and proposed a novel hybrid methodology which combines the ARIMA and GP models. From the empirical results, the proposed hybrid methodology is more outperform than other forecasting models. In future studies, the proposed method will be applied to forecast more different time series to demonstrate the universality of the proposed hybrid forecasting model.

References

[1] K. Kumar, V.K. Jain, Autoregressive integrated moving averages (ARIMA) modelling of a traffic noise time series, Applied Acoustics 58 (3) (1999) 283–294.

[2] V.S. Ediger, S. Akar, ARIMA forecasting of primary energy demand by fuel in Turkey, Energy Policy 35 (3) (2007) 1701–1708.

[3] R.E. Abdel-Aal, Modeling and forecasting electric daily peak loads using abductive networks, Electrical Power and Energy Systems 28 (2) (2006) 133–141.

[4] C. Hamzacebi, Forecasting of Turkey's net electricity energy consumption on sectoral bases, Energy Policy 35 (3) (2007) 2009–2016.

[5] Y.S. Murat, H. Ceylan, Use of artificial neural networks for transport energy demand modeling, Energy Policy 34 (17) (2006) 3165–3172.

[6] K.J. Kim, Financial time series forecasting using support vector machines, Neurocomputing 55 (1–2) (2003) 307–319.

[7] U. Thissen, R. van Brakel, A.P. de Weijer, W.J. Melssen, L.M.C. Buydens, Using support vector machines for time series prediction, Chemometrics and intelligent laboratory systems 69 (1–2) (2003) 35–49.

[8] P.C. Chang, C.H. Liu, C.Y. Fan, Data clustering and fuzzy neural network for sales forecasting: a case study in printed circuit board industry, Knowledge-based Systems 22 (5) (2009) 344–355.

[9] E. Hadavandi, H. Shavandi, A. Ghanbari, Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting, Knowledge-based Systems 23 (8) (2010) 800–808.

[10] G. Zhang, B.E. Patuwo, Y.M. Hu, Forecasting with artificial neural networks: the state of the art, International Journal of Forecasting 14 (1) (1998) 35–62.

[11] G.P. Zhang, Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing 50 (2003) 159–175.

[12] P.F. Pai, C.S. Lin, A hybrid ARIMA and support vector machines model in stock price forecasting, Omega 33 (6) (2005) 497–505.

[13] K.Y. Chen, C.H. Wang, A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan, Expert Systems with Applications 32 (1) (2007) 254–264.

[14] C.H. Aladag, E. Egrioglu, C. Kadilar, Forecasting nonlinear time series with a hybrid methodology, Applied Mathematics Letters 22 (9) (2009) 1467–1470.

[15] C.S. Ong, J.J. Huang, G.H. Tzeng, Building credit scoring models using genetic programming, Expert Systems with Applications 29 (1) (2005) 41–47.

[16] G.E.P. Box, G.M. Jenkins, Time Series Analysis: Forecasting and Control, revised ed., Holden-Day, San Francisco, 1976.

[17] G.E.P. Box, G.M. Jenkins, Intervention analysis with applications to economic and environmental problems, Journal of the American Statistical Association 70 (1975) 70–79.

[18] A.C. Harvey, Time Series Models, John Wiley & Sons, Inc., New York, 1981.

[19] H. Akaike, A new look at the statistical model identification, IEEE Transaction on Automatic Control AC-19 (1974) 716–723.

[20] J. Koza, Genetic Programming: On the Programming of Computers by Natural Selection, MIT Press, Cambridge, MA, 1992.

[21] J.W. Davidson, D.A. Savic, G.A. Walters, Symbolic and numerical regression: experiments and applications, Information Sciences 150 (1–2) (2003) 95–117.

[22] D.G. Lee, B.W. Lee, S.H. Chang, Genetic programming model for long-term forecasting of electric power demand, Electric Power Systems Research 40 (1) (1997) 17–22.

[23] N. Muttil, J.H.W. Lee, Genetic programming for analysis and real-time prediction of coastal algal blooms, Ecological Modelling 189 (3–4) (2005) 363–376.

[24] J.J. Huang, G.H. Tzeng, C.S. Ong, Two-stage genetic programming (2SGP) for the credit scoring model, Applied Mathematics and Computation 174 (2) (2006) 1039–1053.

[25] S.Y. Liong, T.R. Gautam, S.T. Khu, V. Babovic, N. Muttil, Genetic programming: a new paradigm in rainfall-runoff modelling, Journal of American Water Resources Association 38 (3) (2002) 557–584.

[26] K. Kavaklioglu, H. Ceylan, H.K. Ozturk, O.E. Canyurt, Modeling and prediction of Turkey's electricity consumption using artificial neural networks, Energy Conversion and Management 50 (11) (2009) 2719–2727.

[27] P. Zhou, B.W. Ang, K.L. Poh, A trigonometric grey prediction approach to forecasting electricity demand, Energy 31 (14) (2006) 2839–2847.

[28] M.J. Campbell, A.M. Walker, A survey of statistical work on the MacKenzie River series of annual Canadian lynx trappings for the years 1821–1934, and a new analysis, Journal of the Royal Statistical Society Series A 140 (1977) 411–431.

[29] A.J. Smola, Learning with kernels, PhD Thesis, Department of Computer Science, Technical University, Berlin, Germany, 1998.