The comparison of forecasting analysis based on the ARIMA-LSTM hybrid models

Zhedong Wu
The Apartment of Applied Mathematics, The Hong Kong Polytechnic University
Shenzhen, China
Line100.zh@hotmail.com

***Abstract*—For ARIMA-LSTM hybrid model, it can be mixed in different ways. Because ARIMA model can be separated into three parts: trend, seasonality, and residuals. One way is that using LSTM model to forecast residuals of ARIMA model, while another way is that using LSTM model to forecast seasonality and residuals of ARIMA model, RMSE, MAE, MAPE are used to measure the fit of two different ARIMA-LSTM hybrid models with the data of 2010-2020 Shenzhen maximum temperature every 3 hours. In result, it is concluded that using LSTM model to forecast residuals of ARIMA model is a better way to forecast than another**

Keywords: Time Series, ARIMA, LSTM, hybrid model, forecasting, temperature

I. Introduction

The accuracy of one methodology has important meaning to forecasting, while deep learning is of great significance to forecasting. For using deep learning to make the important decision, for example, the investment of large capital, a small increase in accuracy from methodology will result a huge increase in return. On the contrary, a small behind of accuracy leads to huge failure. Time series is another important method of forecasting, while the ARIMA model is a classic and traditional time series model. Meanwhile, the LSTM model is a neural network model in deep learning that can be combine with the ARIMA model to form the ARIMA-LSTM hybrid model. However, the ARIMA-LSTM hybrid model has two slightly different method to combine these two models. This article will use the data of Shenzhen maximum temperature every 3 hours in the ten years from 2010 to 2020 to discuss the fits of two different ARIMA-LSTM hybrid models.

II. Literature Review

G. Peter Zhang (1997) proposed a hybrid method of ARIMA-LSTM model: first divide the time series into linear and non-linear parts:

$$\mathrm{e_t} = \mathrm{y_t} + \mathrm{L_t} \tag{1}$$

then define

$$e_t = f(e_{t-1}, e_{t-2}, \dots, e_{t-n}) + \varepsilon_t \tag{2}$$

where f is the nonlinear function determined by the LSTM model and εt represents the residuals at t.

In this article, this method named method 1. [1] In addition, Dongyan Fan, Hai Sun and Jun Yao et al. (2021) used the ARIMA-LSTM hybrid model to predict the production of oil. Since in actual world, oil wells are frequently switched on/off, the production of oil can be classified as the linear and nonlinear parts, and the LSTM model is used to record the time series of daily production while the ARIMA residuals is used as the input of the subsequent LSTM model. After comparing the test set and predictions, the RMSE, MAPE, MAE and Sim value these four indicators show that it has a better superiority than the ARIMA or LSTM single model. [2]

Kun Zhou, Wen Yong Wang and Teng Hu et al. (2020) compared the advantages and disadvantages of ARIMA and LSTM: learning similar to LSTM model requires more complex modeling capabilities and lacks interpretability but LSTM model has a higher accuracy of prediction than ARIMA model. [3] However, Dongyan Fan, Hai Sun and Jun Yao et al. (2021) believe that the ARIMA model has good accuracy on the stable output decline curve, but the LSTM model has a greater advantage in predicting fluctuating nonlinear data. [2]

Emmanuel Dave, Albert Leonardo and Marethia Jeanice et al. (2020) mentioned another ARIMA-LSTM hybrid model, which aims to split the target that needs to be predicted into different parts. In the forecast of Indonesia export, they proposed that *prediction[i] = TREND prediction[i] + SEASONAL prediction[i] + RESIDUAL prediction[i]*, separated the export into three parts: trend, seasonality, and residuals. According to the good performance of ARIMA model in prediction of linear relationships and the obvious advantages of LSTM model in nonlinear relationships, they determine each part to use ARIMA or LSTM model. Comparing to method 1, they did not only use the LSTM model to analyze the residuals of the ARIMA model, instead, added the seasonality into LSTM model. This method will be named as method 2 in this article. [4]

III. Method

Deep learning such as ANN and LSTM needs to divide the data set into training set and test set. In this article, the data will be divided into: The test set is 26296 sets of Shenzhen temperature data every 3 hours from 2010.1.1 to 2018.12.31, training set It is 5848 sets of Shenzhen temperature data every 3 hours from 2019.1.1 to 2020.12.31.

978-1-6654-1025-0/21/$31.00 ©2021 IEEE
DOI 10.1109/ICECEM54757.2021.00044

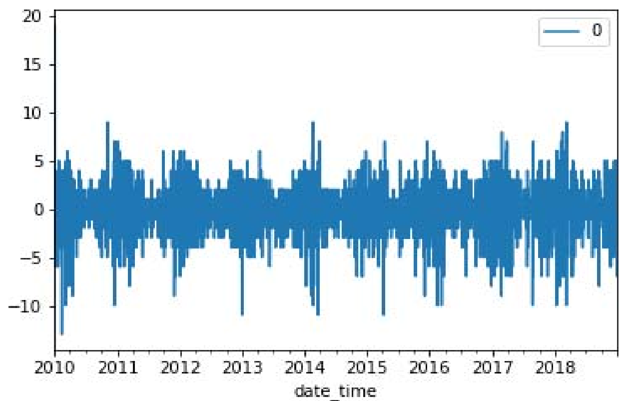

Figure 1. Tranning Set Data Plot

In addition, in Python, the way to compare the two is to set whether ARIMA is seasonal. In other words, method 1 is using the SARIMAX (p, d, q)*(P, D, Q, m) model to obtain the residual of the training set, and then use the residuals of the training set to train the LSTM model for deep learning. Furthermore, using SARIMAX (p, d, q)*(P, D, Q, m) to obtain the residuals of the test set, replacing residuals of the test set and compare the test set to obtain the final random error data. In method 2, the first step is to use the ARIMA (p, q, d) model to obtain the residual of the training set, and then perform deep learning on the residuals, replacing the random errors of the test set and compare the test set to obtain the random error of method 2 in the test set. Last, compare the random error of Method 1 and Method 2 by using the three indicators of RMSE, MAPE, and MAE to determine which method is better.

The methodology is based on: ARIMA formula is

$$x_t = \varphi_0 + \varphi_1 x_{t-1} + \cdots + \varphi_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \qquad (3)$$

Define $\hat{x}\ t$ as the prediction value, xt as the test set value, then the results of ARIMA and SARIMAX are

$$\hat{x}_t = x_t - \varepsilon_t = \varphi_0 + \varphi_1 x_{t-1} + \cdots + \varphi_p x_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

$$where\ xt\ is\ the\ test\ set \qquad (4)$$

Let εt be the input of LSTM model, Nt be the output of LSTM model, then we get

$$\hat{y}_t = \hat{x}_t + N_t = \varphi_0 + \varphi_1 x_{t-1} + \cdots + \varphi_p x_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \qquad (5)$$

And $\hat{y}_t - x_t = \theta_t$

$$where\ \theta_t\ \text{is}\ the\ \text{random error by ARIMA} - \text{LSTM model} \qquad (6)$$

Finally, compare θ_t by using RMSE, MAPE, MAE to determine that which method is a better way in forecasting the temperatures.

IV. Modeling

The ARMA (p, q) is composed of an autoregressive model AR(p) and a moving average model MA(q), where p represents the number of autoregressive terms and q represents the moving average terms. The advantage of ARMA model is that a simpler model can be obtained under same-size models. The ARIMA (p, q, d) is derived from ARMA (p, q), since ARMA (p, q) requires the variable owns stationarity, that is the mean, variance, and autocovariance are all independent of time. Therefore, ARIMA model introduces the sequence differentiation I(d) to solve this limitation.

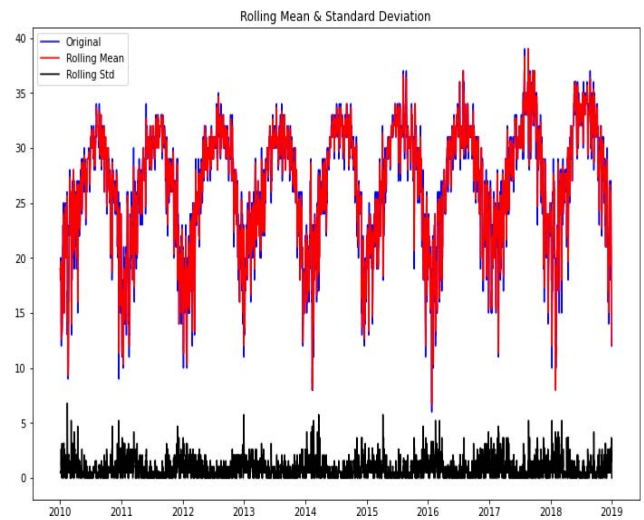

Figure 2. Rolling Mean And Standard Deviation

The seasonal ARIMA model, or SARIMAX (p, d, q) x (P, D, Q, m) is to further split the m size interval in the seasonal part. The first step in analyzing the ARIMA/SARIMAX model is confirming its stationarity. The Dickey-Fuller test is one of normally used method to check its stationarity:

Table 1. Dickey-Fuller test result

Statistical Index Name	***Result***
Test Statistic	-6.174125
p-value	0.00224573678
Number of Lags Used	13.0477527766
Number of Observations Used	143.303764392
Critical Value (1%)	-3.430599
Critical Value (5%)	-2.861650
Critical Value (10%)	-2.566829

We found that in the 8-year Shenzhen temperature data, the Test Statistic result is -6.174125, which is less than 1% level - 3.430599, and the P-value is 0.00224573678 and less than 0.01. The results show that the hypothesis that the unit root exists in the training set is well rejected, which means that we should set d = 0.

Subsequently, the parameters of p and q of the non-seasonal ARIMA model are estimated by checking the autocorrelation and partial correlation graph of the stationary series (ACF & PACF graph), it is found that the autocorrelation graph has a slow downward trend to form a tail, and the partial correlation graph is after lag 2 Confidence close to 0 forms an obvious truncation, so ARIMA (2,0,0) is a suitable time series model for this data.

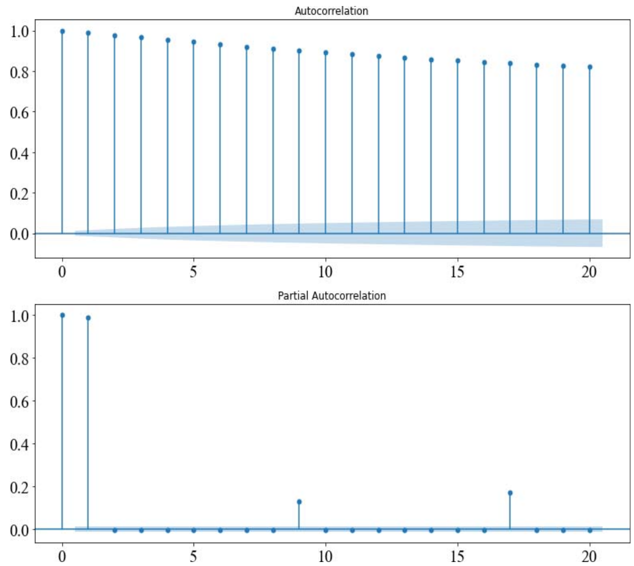

Figure 3. Acf & Pacf Graph

Table 2. ARIMA (2,0,0) Analyze

Statistical Index Name	***Result***
Log Likelihood	30287.011
AIC	60586.022
BIC	60635.081
HQIC	60601.863

The seasonality of the weather can be assumed by one day, one month, or one year, that is, 8*1, 8*30, 8*360. Furthermore, the superiority of a single SARIMAX model should be higher than that of a single ARIMA model, that is, the AIC of the SARIMAX model should be smaller than that of the ARIMA model. Under this premise, when comparing method 1 and method 2, the value of m has no significant effect on the result, so taking the smallest m=8 is a good parameter. The enumeration method is used to enumerate all the combinations from 0 to 2, and the AIC criterion is used for comparison. After comparison, it is found that the AIC of SARIMAX (2, 0, 0) x (1, 0, 2, 8) is 60580.528, which is the smallest AIC among the 27 enumeration items and is smaller than the 64011.160 of ARIMA (2,0,0).

Table 3. SARIMAX (2, 0, 0) x (1, 0, 2, 8) Analyze

Statistical Index Name	***Result***
Log Likelihood	32002.580
AIC	64011.160
BIC	64035.691
HQIC	64019.080

Finally, the 26296 residuals of SARIMAX (2, 0, 0) x (1, 0, 2, 8) and ARIMA (2, 0, 0) in the training set were put into the LSTM model for deep learning: first using the MinMaxScaler function in the sklearn. preprocessin package in Python for the data, even if the data is standardized between 0 and 1. Subsequently, 4 layers of LSTM layer are used, the prediction points are combined with the previous 60 days data, the number of neurons is 50, the Mini-Batch Gradient Descent (MBGD, batch size = 32) is used, the Dropout is set to 0.2, the Optimizer selects Adam, and the number of Output layers is 1. The MSE size is used as the training target, and the training is repeated 100 epochs.

For the SARIMAX model, we should determine whether there exists seasonality first. By splitting the trend, seasonality, and the residuals of the training set, the maximum temperature data of Shenzhen every 3 hours has an obvious seasonality.

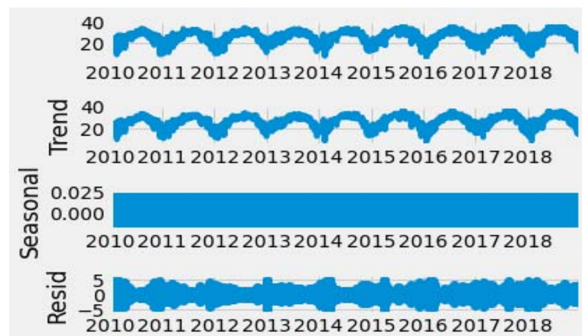

Figure 4. Trend, Seasonality, And Residuals Of Data

V. Comparison

For the method 1, first use the SARIMAX (2, 0, 0) (1, 0, 2, 8) model to obtain the residual error of the model in the test set. After 100 times of deep learning of the LSTM model on the residual, the MSE is 0.0001550. The training results on the test set are shown in the figure:

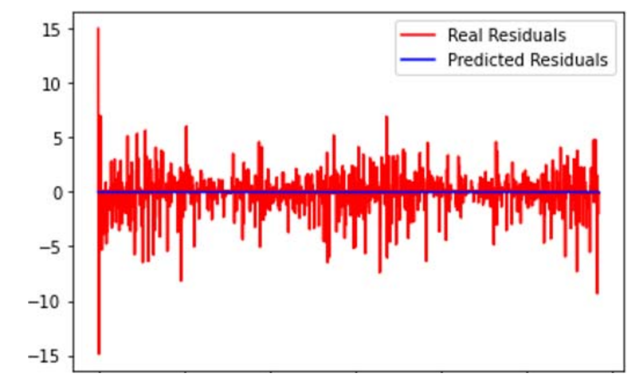

Figure 5. Real Residuals And Predicted Residuals For Method 1

However, the predicted result is always a constant -0.018672. This means that LSTM model is overfit, and it is resulted by the data are too concentrated. To solve this problem, the input data should be feature standardization, that is, the mean value of each input data becomes 0, and the variance becomes 1. This increases the standard deviation between data and the differences between data are more obvious. This standardization scales data of which magnitude is too small/large to a suitable range and improve the accuracy of model training. The scale function in *sklearn.* Preprocessing is used in Python to standardize the features of the training set. After the feature is standardized, the top ten items of the predicted value of the residual are:

Table 4. Head 10 random errors of method 1

Time	***Random Error***
2019/1/1 0:00	0.8254531264
2019/1/1 3:00	0.82207160286
2019/1/1 6:00	0.85428642248
2019/1/1 9:00	0.91790294728
2019/1/1 12:00	1.24871145755
2019/1/1 15:00	0.45674088409
2019/1/1 18:00	1.02267855051
2019/1/1 21:00	0.77702764434
2019/1/2 0:00	0.74180559196
2019/1/2 3:00	0.79148301212

For the method 2, using the non-seasonal ARIMA (2, 0, 0) model, the top ten items of the residual predicted value for:

Table 5. Head 10 random errors of method 2

Time	***Random Error***
2019/1/1 0:00	0.09926274677
2019/1/1 3:00	-0.12044415845
2019/1/1 6:00	-0.36058258019
2019/1/1 9:00	-0.42964107354
2019/1/1 12:00	-0.23118297562
2019/1/1 15:00	0.1582871278
2019/1/1 18:00	0.15409180304
2019/1/1 21:00	0.2216020938
2019/1/2 0:00	0.216578914
2019/1/2 3:00	0.21445610957

Next, use RMSE, MAPE, and MAE these three indicators to determine the fits of two methods:

Table 6. Comparison of Method 1&2 RMSE, MAPE, MAE

Statistical Index Name	***Method 1***	***Method 2***
RMSE	0.529453654	0.78876716
MAPE	129.444623	5.120623484
MAE	0.280321172	0.622153632

Both RMSE and MAE have obvious characteristics: RMSE is mathematically a measurement of the average of errors, and it gives higher weights to greater errors in a data set; while MAE prefers to show more information about skewness of the errors, that is, more analysis of the median in a data set. And compared to RMSE, MAE gives the same weight to each error. This means that the method 1 has a better average and skewness of errors than method 2, and the method 1 is significantly better than method 2 in terms of RMSE and MAE.

However, method 1 has much greater MAPE than method 2.

By discovering the formula of MAPE:

$$\mathrm{MAPE} = \frac{1}{n}\Sigma\frac{|e_t|}{d_t} \tag{7}$$

when the actual value is relatively small in a data set, the error of prediction in this value will have a greater impact on MAPE. Therefore, for a small actual value, method 1 has larger errors than method 2.

VI. Conclution

Although method 2 has better predictions for relatively small data in the data set, in general, method 1 is better than method 2 for predicting the highest temperature, that is, for the ARIMA- LSTM hybrid model, first use the seasonal SARIMAX model forecasts is better than first use the non-seasonal ARIMA model. In the ARIMA/SARIAMX model, the forecast target is divided into three terms: Trend, Seasonality, and residuals. In the data of the highest temperature every 3 hours, which has obvious seasonality, the ARIMA-LSTM hybrid model that puts both seasonal and residuals terms into LSTM model, is not as good as that only put the residuals into the LSTM model for predictions.

References

[1] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159–175. https://doi.org/10.1016/s0925-2312(01)00702-0

[2] Fan, D., Sun, H., Yao, J., Zhang, K., Yan, X., & Sun, Z. (2021). Well production forecasting based on ARIMA-LSTM model considering manual operations. Energy, 220, 119708. https://doi.org/10.1016/j.energy.2020.119708

[3] Zhou, K., Wang, W. Y., Hu, T., & Wu, C. H. (2020). Comparison of Time Series Forecasting Based on Statistical ARIMA Model and LSTM with Attention Mechanism. Journal of Physics: Conference Series, 1631, 012141.https://doi.org/10.1088/1742-6596/1631/1/012141

[4] Dave, E., Leonardo, A., Jeanice, M., & Hanafiah, N. (2021). Forecasting Indonesia Exports using a Hybrid Model ARIMA-LSTM. Procedia Computer Science, 179, 480–487. https://doi.org/10.1016/j.procs.2021.01.031