2018 9th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP)

Time Series Forecasting using Sequence-to-Sequence Deep Learning Framework

Shengdong Du, Tianrui Li
School of Information Science and Technology,
Southwest Jiaotong University
Chengdu 611756, China
e-mail: {sddu, trli }@swjtu.edu.cn

Shi-Jinn Horng
Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology
e-mail: horngsj@yahoo.com.tw

***Abstract*—**Time series forecasting has been regarded as a key research problem in various fields. such as financial forecasting, traffic flow forecasting, medical monitoring, intrusion detection, anomaly detection, and air quality forecasting etc. In this paper, we propose a sequence-to-sequence deep learning framework for multivariate time series forecasting, which addresses the dynamic, spatial-temporal and nonlinear characteristics of multivariate time series data by LSTM based encoder-decoder architecture. Through the air quality multivariate time series forecasting experiments, we show that the proposed model has better forecasting performance than classic shallow learning and baseline deep learning models. And the predicted PM2.5 value can be well matched with the ground truth value under single timestep and multi-timestep forward forecasting conditions. The experiment results show that our model is capable of dealing with multivariate time series forecasting with satisfied accuracy.

Keywords- Time series forecasting, LSTM, Encoder-decoder, PM2.5, Sequence-to-sequence deep learning.

I. INTRODUCTION

Time series forecasting covers a wide range of real-life problems, which has important research value in various fields such as financial index forecasting, anomaly detection, traffic prediction, medical monitoring, intrusion detection, air pollution prediction and so on. And research on time series forecasting is very important and has always been regarded as a key issue in a lot of fields [1]. In general, a time series is a collection of observations made in chronological order. Time series data has unique characteristics, e.g. a large amount of data, high dimensionality, and update constantly, so it is very difficult to analyze and model them effectively.

For decades, many researchers have been devoted to time series forecasting and have achieved important results in theory and applications [2]. However, most of these studies do rely on mathematical equations or simulation techniques to describe the evolution of time series data. These traditional methods are represented by the classic statistical method and shallow machine learning algorithms. For example, Autoregressive Integral Moving Average (ARIMA) is often used as a univariate time series model, which is widely applied to time series forecasting problems [3]. Pai et al. proposed a seasonal support vector regression (SSVR) model to forecast time series data which outperforms both SVR and SARIMA models in terms of forecasting accuracy [4]. Zhang et al. investigated the issue of how to effectively model time series with both seasonal and trend patterns by artificial neural network modeling [5].

In the big data era, with the rapid growth of sensor data acquisition technology, time series forecasting models need to match increasingly complicated and big datasets. Moreover, shallow learning models have bottlenecks in handling big data (especially complex multivariable, high dimensional and noisy big data sets.), new time series forecasting task needs data-driven model support [6]. Deep learning is currently the most popular data-driven model [7], which can automatically extract and learn the deep representation of various time series data. Since 2012, deep learning has made great progress in research and applications of image processing, audio processing, and natural language understanding etc. [8] [9] [10]. Although time series forecasting task usually adopts traditional shallow machine learning methods, the deep learning model for time series analysis is getting more and more attention [6] [11].

In this paper, we propose an end-to-end model to solve the time series forecasting problems (such as air pollution forecasting, traffic flow forecasting etc.) via sequence-to-sequence deep learning framework, which used an encoder-decoder learning structure [10] [13]. The proposed model can learn the long temporal dependencies of multivariates time series data. The air quality time series data experiments indicated that the proposed model has good forecasting performance and generalization ability.

The rest of the paper is organized as follows: Section II presents the related work. Section III shows an overview of the proposed deep learning framework, including the overall design of our method and model in detail, e.g. how to design model structure which used for time series processing. Section IV describes the comparative experiments, and the effectiveness of the proposed framework is analyzed and evaluated. We draw conclusions and directions for future research in the last section.

II. RELATED WORK

Time series forecasting has a good study history in the literature, most of the existing works usually solve the problems of time series forecasting using classical statistical method or shallow machine learning models [1] [2], e.g. Support Vector Regression [4], ARIMA[3] and Artificial

978-1-5386-9403-9/18/$31.00 ©2018 IEEE
DOI 10.1109/PAAP.2018.00037

Neural Network [14]. Park et al. proposed a novel method that forecasts change direction (up or down) of next day's closing price of financial time series using the continuous HMM [15]. Hassan et al. proposed a hybrid method of Hidden Markov Model(HMM), Fuzzy Logic and multi-objective Evolutionary Algorithm (EA) for building a fuzzy model to predict non-linear time series data [16]. Ahmed et al. presented a large scale comparison study for the major machine learning models (include Multilayer Perceptron, Bayesian Neural Networks, K-nearest Neighbor Regression, Support Vector Regression and Gaussian Processes etc.) for time series forecasting task [17].

In recent years, time series forecasting based on big data analysis has become a research hotspot. Because real-world time series data usually have high dimensionality, dynamic and nonlinear characteristics, more and more researchers are trying to use data-driven models, especially using deep learning models [11] [12]. For example, Yao et al. proposed a deep learning framework which integrates convolutional and recurrent neural networks to exploit local interactions and extract temporal relationships to model time series dynamics[18]. Chambon et al. proposed a deep learning approach for sleep stage classification which can exploit all multivariate and multimodal Polysomnography (PSG) time series data [19]. Freeman et al. proposed an air quality time series forecasting method using deep learning model which consisting of a recurrent neural network (RNN) with long short-term memory (LSTM)[20].

More recently, Sequence-to-sequence deep learning has been widely applied to sequence data process and natural language understanding problems [10]. Sequence-to-sequence architecture is a general end-to-end approach for sequence data learning which makes minimal assumptions on the sequence structure. And it usually uses an encoder-decoder structure to encode the input sequence to a vector of a fixed dimensionality, and then decode the target sequence from the vector [13]. Venugopalan1 et al. proposed a novel end-to-end sequence-to-sequence model to generate captions for videos, which have demonstrated state-of-the-art performance in image caption generation problem [21]. Kuznetsov et al. presented the first theoretical analysis of sequence-to-sequence deep learning time series forecasting framework, and made a comparison of sequence-to-sequence modeling to classical time series models [22]. Until now sequence-to-sequence deep learning has not yet been well researched and applied for time series forecasting problems, e.g. air quality, traffic flow, electrocardiogram and power load time series forecasting etc.

In this paper, by a comparison of traditional shallow machine learning models and classic deep learning models, we propose a new end-to-end time series forecasting model via sequence-to-sequence deep learning framework, which is motivated to address deep representation and long temporal dependency problems of multivariate time series data and performing deep feature learning automatically.

III. THE PROPOSED METHOD

A. Problem and Definitions

A time-series is a sequence of values measured over timestep, in discrete or continuous time units. Time series forecasting has been a key issue in early warning and control of many industrial applications. Its goal is to anticipate changes in the future value at observation points over time. The observation time period is usually different, which is decided by the ground-based monitoring sensor, e.g. an air pollution time series of PM2.5 is shown in Fig. 1, and the observation time period is one hour.

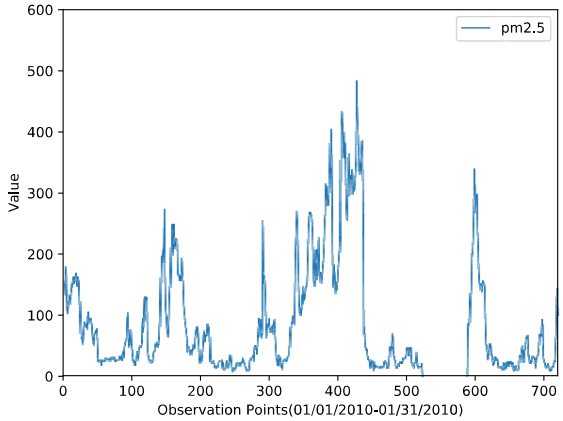

Figure1. One month PM2.5 value (01/01/2010-01/31/2010) of Beijing air pollution data set from UCI [23].

As shown in Fig. 2, air quality data is typical multivariate time series, which usually contains the real-valued PM2.5 pollutant, also has some other variables such as temperature and wind speed etc. and PM2.5 value is highly related by meteorological observation data. For example, high wind speed will reduce the concentration of PM2.5, high humidity usually aggravates air pollution, and high atmospheric pressure usually results in good air quality, etc. Therefore, the above multivariate time series characteristics are very important for air quality forecasting task.

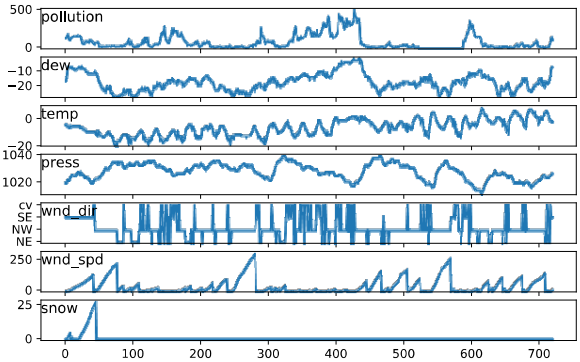

Figure2. One month(01/01/2010-01/31/2010) air quality data (include PM2.5, pressure, wind speed, snow, rain etc.) of Beijing air pollution data set from UCI [23].

Time series forecasting task is illustrated as follows. Taking air pollution data as an example, given time T, the forecasting task is to anticipate the PM2.5 value $P_{i,T+1}$ at time $T+1$ or $P_{i,T+n}$ at time $T+n$ which model the history air quality related time series dataset $AQI = \{AQI_{i,t} | i \in O, t = 1,2,\dots,T$ in the past$\}$, where AQI represents the history multivariate time series of air quality index, O means the overall observation points, and AQI not only includes PM2.5 time series itself but also includes other air quality related time series data such as press, temperature, wind speed, etc.

B. Overview of the Sequence-to-Sequence Framework

A time series forecasting method based on sequence-to-sequence deep learning architecture is proposed in this paper, which is a combination of encoder and decoder component that take into account the spatial-temporal dependence of multivariate time series data. Encoder-decoder is not a specific model, but a kind of general framework. We used LSTM as the component of encoder and decoder parts of the sequence-to-sequence deep learning framework, which has shown excellent performance for natural language processing tasks [10].

Fig. 3 shows the graphical illustration of the sequence-to-sequence deep learning framework. The overall framework consists of two main components: one is the LSTM encoder, it takes the history multivariate time series samples as input for learning and produces the fixed length vector which captures the temporal representation of the past time series data. The other component is the LSTM decoder, which based on the temporal structure representation (fixed length vector) and generates the future time series as forecasting output. As the above process, the proposed model can predict the most probable hypotheses of the future time series value using the end-to-end sequence deep learning framework.

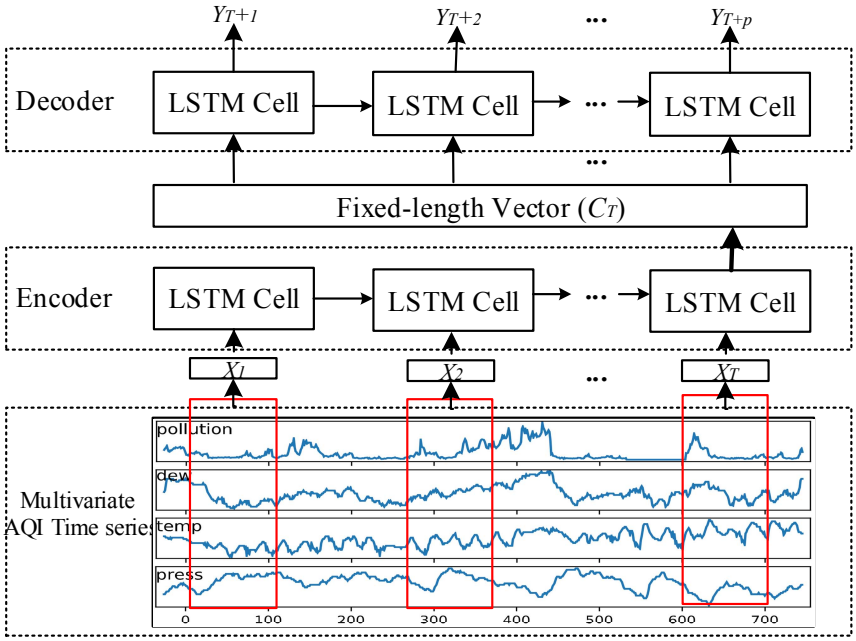

Figure 3. The diagram of the time series sequence-to-sequence deep learning framework(TSDLF). The schematic illustration shows the proposed model with encoder-decoder architecture and temporal representation learning for multivariate air quality time series forecasting.

C. LSTM based Encoder-Decoder Model

Although traditional shallow learning models such as HMM and SVR can process time series, the efficiency is not so good, especially under big data condition. Long Short-term Memory network (LSTM) is another good option, which is a popular model for sequence data process [24]. The memory cell of each LSTM block contains four main components. The collaboration of these components enables cell learning and memory long dependency features. An LSTM cell is a recurrent unit which uses the input i_t,the hidden state activation $\tilde{s}_t$ and memory cell activation s_t to compute the hidden state activation at at time t. It uses a combination of a memory cell and three types of gates: input gate i_t, forget gate f_t, and output gate o_t to decide the amount of information which should transfer to the next time step or to output. The typical LSTM block computing process is as follows:

$$i_t = \sigma\left(U^{(i)}x_t + W^{(i)}h_{t-1} + b_i\right) \quad (1)$$

$$f_t = \sigma\left(U^{(f)}x_t + W^{(f)}h_{t-1} + b_f\right) \quad (2)$$

$$o_t = \sigma\left(U^{(o)}x_t + W^{(o)}h_{t-1} + b_o\right) \quad (3)$$

$$\tilde{s}_t = tanh\left(U^{(c)}x_t + W^{(c)}\ h_{t-1} + b_c\right) \quad (4)$$

$$s_t = f_t \circ s_{t-1} + i_t \circ \tilde{s}_t \quad (5)$$

$$h_t = o_t \circ tanh(s_t) \quad (6)$$

As shown in the above formulas, $\tilde{s}_t$ is a neuron with a self-recurrent cell like RNN. s_t is the internal memory cell of LSTM block which is summed by two parts: one is the previous internal memory state s_{t-1} and forget gate f_t. the second part is calculated by element wise multiplication of self-recurrent state $\tilde{s}_t$ and input gate i_t.

The LSTM based encoder-decoder model for multivariate time series forecasting task was first proposed in this paper. The model includes an encoder and decoder component, and the LSTM encoder learning the input multivariate time series $X_1, X_2, \ldots, X_T$ of the length T and produces the temporal representation of the past time series data through the cell state vector C_T, each $X_i = [o_1, o_2, \ldots, o_m]$ represents multiple observation variables at i time-step. Then the LSTM decoder uses this representation C_T as initial state to reconstruct the past time series as the target prediction, and a linear layer on top of the LSTM decoder layer is used to predict the target.

The final modal training problem is to minimize the overall error C_i of training samples for each time window time series as follows:

.

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ C_i = \sum_{i=1}^{T}\sum_{j=1}^{m} ||{Y_i}^j - {X_i}^j||^2 \quad (7)$$

where i indicates each time-step input of multivaratie time series data (i=1, 2, ..., T), T represents the input sequence length of a time window data, and j indicates the jth observation variable of multivaratie time seires at time-step i, and $\boldsymbol{\theta}$ is the parameter space including W_l^i and b_l^i of each layer.

IV. EXPERIMENTS

In this section, we use real multivariate time series data set (mainly air quality time series data) to conduct experiments to analyze and evaluate the proposed model. Through the comparison of classical shallow learning model, baseline deep learning model, and our model, the forecasting performance and effectiveness of the proposed model are validated.

A. Datasets and Experimental Setup

The real air quality multivariate time series datasets are used for experiments. Details of the experimental dataset are described as follows:

Air quality dataset(Beijing PM2.5): The BeijingPM2.5 dataset is a typical multivariate time series which includes meteorological data and PM2.5 data and published on UCI [23]. The data is collected every hour and is sourced from the data interface released by the US Embassy in Beijing. The dataset used for our experiments is ranged from 01/01/2010 to 31/12/2014, which has 43824 records.

Then we describe the hardware and software environment of the experiment and the configuration of relevant parameters. The open source deep learning libraries such as Keras and Tensorflow are used to build the experimental deep learning models, and Scikit-learn is used to build shallow learning models. All experiments are conducted on a PC Server, whose configuration is Intel(R) Xeon(R) CPU E5-2623 3.00GHz, 4 GPUs each is 12G NVIDIA Tesla K80C, and memory is 128GB.

Our model is compared with one classic shallow learning model (SVR) and three baseline deep learning models (RNN, LSTM, GRU). They are summarized as follows.

SVR is a kernel method of machine learning which also can be used for time series forecasting, and three SVR models with a different kernel (RBF, poly and linear) are used in the experiments. RNN is a popular deep learning method for handling sequence tasks. LSTM and GRU (Gated Recurrent Units) are the most popular variants of RNN.

We select mean square error (MSE) as the loss function of model training, and the activation function of the output layer is a linear function, which is used for regression prediction. Moreover, we apply the min-max function to normalize the multivariate time series data to [0,1]. Additionally, we select the first four years data for training and select the last year data for testing from Beijing PM2.5 dataset. Hyper-parameters of deep learning model are listed in the following; batch size sets to 96, dropout rate sets to 0.8, learning rate sets to 0.01, each hidden layer units number sets to 128, the default hidden layers number of baseline deep learning model sets to 1, and select Adam as optimizer [25]. Lastly, we use RMSE as the model error evaluation indicators, which is used to evaluate the model performance.

B. Results Analysis

The experiment results of air quality multivariate time series forecasting task are reported in Table 1, which gives RMSE comparative analysis of SVR (rbf, linear and poly kernel), RNN, LSTM, GRU and our proposed model TSDLF. As shown in the table, our model is superior to other models in terms of PM2.5 single-step forward and multi-step forward prediction performance (lookup size set 6 which is also called input window size, which represents historical time series observations input size of the model). Compared to the baseline shallow and deep learning models, our model reduces error to 0.031 when forward prediction size is 1, also has the lowest error when forward prediction size is 3 or 6, which improves the forecasting accuracy.

Moreover, the prediction performance of baseline deep learning methods is also better than the classic shallow learning method. The primary reason is that our model can learn deep temporal representation and long-term dependencies features of air quality multivariate time series data. Furthermore, it also makes use of the interdependent spatial-temporal features of multivariate time series data. This means that deep learning models are more effective for multivariate time series forecasting than traditional shallow

learning methods. In short, Experimental results show that our sequence-to-sequence deep learning framework for multivariate time series forecasting has better prediction performance than those of baseline models.

Table 1: RMSE of our model TSDLF and comparisons with other baseline models for the PM2.5 forward forecasting task.

Type	***Models***	***RMSE***		
		(6,1)	***(6,3)***	***(6,6)***
Shallow model	SVR-POLY	0.068	0.116	0.098
	SVR-RBF	0.045	0.055	0.067
	SVR-LINEAR	0.034	0.057	0.072
Deep model	LSTM	0.051	0.048	0.058
	GRU	0.042	0.063	0.178
	RNN	0.099	0.098	0.097
	TSDLF(Ours)	**0.031**	**0.040**	**0.049**

Note: A pair of numbers in double parentheses like (6,1) in the table represents (lookup size, forward-timestep prediction size). training epoch of deep learning models set 100, the ground truth and predicted value for RMSE computing which is normalized to [0,1].

In addition, it is found by experiments that the choice of forward prediction size has a great influence on the forecasting performance. As Table 1 shows, the performance of multi-step forward prediction is significantly lower than that of single step forward prediction. As the forward prediction size increases, the forecasting performance of these models gradually decrease. But we can observe that compared to baseline deep models, our model TSDLF also has the lowest prediction error versus different forward prediction size.

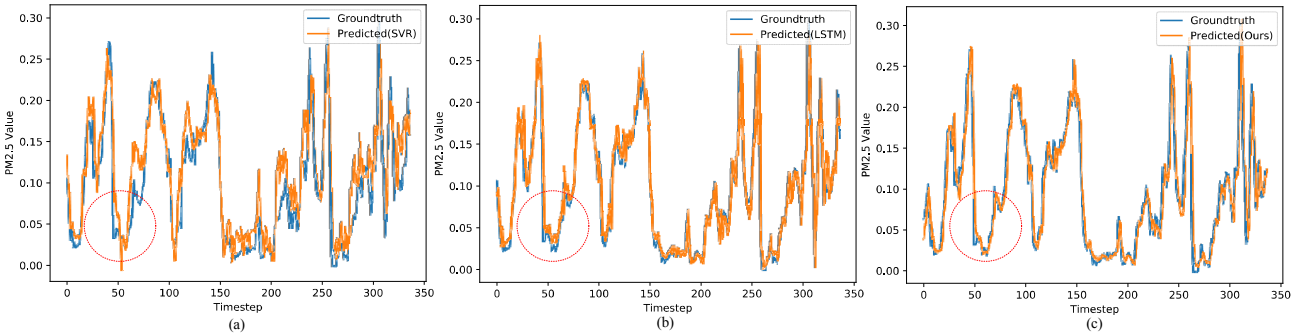

Figure 4. (a) Comparison of ground truth and one-timestep forward predicted PM2.5 value of SVR model; (b) Comparison of ground truth and one-timestep forward predicted PM2.5 value of LSTM model; (c) Comparison of ground truth and one-timestep forward predicted PM2.5 value of Our model.

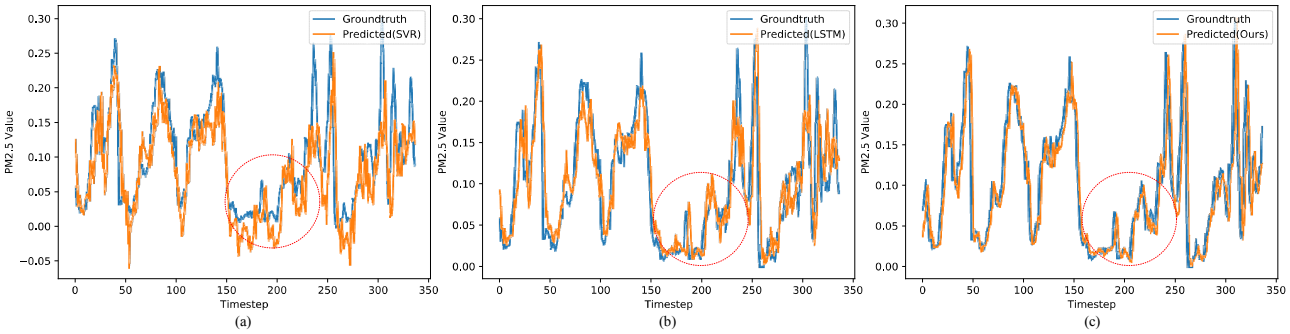

Figure 5. (a) Comparison of ground truth and three-timestep forward predicted PM2.5 value of SVR model; (b) Comparison of ground truth and three - timestep forward predicted PM2.5 value of LSTM model; (c) Comparison of ground truth and three-timestep forward predicted PM2.5 value of Our model.

In order to further analyze and compare the prediction performance of our model, we analyze the PM2.5 time series forecasting ability of our model under different timestep forward prediction over the course of two weeks (336 timestep points). Figs. 4. (a) (b) (c) give a comparison of the ground truth (expected) and one-timestep forward predicted PM2.5 values of SVR, LSTM, and our model, and x-coordinate indicates observation timestep, y-coordinate indicates a PM2.5 value which is normalized to [0,1]. As shown in the figure, the performance of our model is better than SVR and LSTM with single timestep forward forecasting task, especially in the time period of wave peak and trough of air quality time series data. Figs. 5. (a), (b) and (c) give a comparison of the ground truth (expected) and six-

timestep forward predicted PM2.5 values of SVR, LSTM, and our model, which show that the performance of our model is also better than the shallow learning model SVR and the baseline deep learning LSTM with multi-timestep forward forecasting task.

In summary, for the proposed TSDLF, the PM2.5 prediction can be well matched with the ground truth with single step forward forecasting, also has better performance than baseline models with multi-step forward forecasting, which implies the proposed sequence-to-sequence deep learning framework can effectively learn the deep features and long-term temporal dependence characteristics of multivariate air quality time series data. The proposed model which is based on sequence-to-sequence deep learning architecture can provide a useful reference for multivariate time series forecasting problem.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a sequence-to-sequence deep learning model for single timestep and multi-timestep forward forecasting of multivariate time series data, which is based on LSTM based encoder-decoder architecture. It can learn the deep representation features of local trend and long dependencies pattern of multivariate time series data. Air quality forecasting experiments showed that the proposed model has better performance than classic shallow learning and baseline deep learning models, which can explore and learn the interdependence and nonlinear correlations of multivariate air quality related time series (such as temperature, humidity, wind speed and PM2.5 itself) effectively. In future research, it is found that the forecasting performance of regular time series is better, but the forecasting performance of time series with abnormal or anomaly points is not very well, which needs to be further improved. In addition, the sequence-to-sequence deep learning model itself also needs to be researched in depth and improved with different time series data (e.g. traffic flow) and different timestep forward (e.g. longer timestep) forecasting condition.

Acknowledgments. This research was partially supported by the National Key Research and Development Program of China (No. 2016YFC0802209), the National Natural Science Foundation of China (No. 61773324), the "Center for Cyber-physical System Innovation" from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan and MOST under 106-2221-E-011-149-MY2 and 106-3114-E-011-008.

REFERENCES

[1] De Gooijer J G, Hyndman R J. 25 years of time series forecasting. International journal of forecasting, 2006, 22(3): 443-473.

[2] Fu T. A review of time series data mining. Engineering Applications of Artificial Intelligence, 2011, 24(1): 164-181.

[3] Zhang G P. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 2003, 50: 159-175.

[4] Pai P F, Lin K P, Lin C S, et al. Time series forecasting by a seasonal support vector regression model. Expert Systems with Applications, 2010, 37(6): 4261-4265.

[5] Zhang G P, Qi M. Neural network forecasting for seasonal and trend time series. European journal of operational research, 2005, 160(2): 501-514.

[6] Längkvist M, Karlsson L, Loutfi A. A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognition Letters, 2014, 42: 11-24.

[7] Schmidhuber J. Deep learning in neural networks: An overview. Neural networks, 2015, 61: 85-117.

[8] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks, in Proceedings of Advances in Neural Information Processing Systems. 2012: 1097-1105.

[9] Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions, in Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 3128-3137.

[10] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares,H. Schwenk, and Y. Bengio, Learning phrase representations using rnn encoder–decoder for statistical machine translation, in Proc. Of 2014 Conf. on Empirical Methods in Natural Language Processing(EMNLP), Oct. 2014, pp. 1724–1734.

[11] Gamboa J C B. Deep learning for time-series analysis . arXiv preprint arXiv:1701.01887, 2017.

[12] Lipton Z C, Kale D C, Elkan C, et al. Learning to diagnose with LSTM recurrent neural networks . arXiv preprint arXiv:1511.03677, 2015.

[13] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [C]//Advances in neural information processing systems. 2014: 3104-3112.

[14] Laptev N, Yosinski J, Li L E, et al. Time-series extreme event forecasting with neural networks at uber [C]//International Conference on Machine Learning. 2017 (34): 1-5.

[15] Park S H, Lee J H, Song J W, et al. Forecasting change directions for financial time series using hidden markov model [C]//International Conference on Rough Sets and Knowledge Technology. Springer, Berlin, Heidelberg, 2009: 184-191.

[16] Hassan M R, Nath B, Kirley M, et al. A hybrid of multiobjective Evolutionary Algorithm and HMM-Fuzzy model for time series prediction. Neurocomputing, 2012, 81: 1-11.

[17] Ahmed N K, Atiya A F, Gayar N E, et al. An empirical comparison of machine learning models for time series forecasting. Econometric Reviews, 2010, 29(5-6): 594-621.

[18] Yao S, Hu S, Zhao Y, et al. Deepsense: A unified deep learning framework for time-series mobile sensing data processing [C]//Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017: 351-360.

[19] Chambon S, Galtier M N, Arnal P J, et al. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2018.

[20] Freeman B S, Taylor G, Gharabaghi B, et al. Forecasting air quality time series using deep learning. Journal of the Air & Waste Management Association, 2018: 1-21.

[21] Venugopalan S, Rohrbach M, Donahue J, et al. Sequence to sequence-video to text [C] //Proceedings of the IEEE international conference on computer vision. 2015: 4534-4542.

[22] Kuznetsov V, Mariet Z. Foundations of Sequence-to-Sequence Modeling for Time Series. arXiv preprint arXiv:1805.03714, 2018.

[23] Beijing PM2.5 Data Set [Online]Available:https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data

[24] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780.

[25] Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.