Available online at www.sciencedirect.com
ScienceDirect
Energy Reports 7 (2021) 1047–1054
www.elsevier.com/locate/egyr

2021 International Conference on Energy Engineering and Power Systems (EEPS2021), August 20–22, 2021, Hangzhou, China

Photovoltaic power prediction of LSTM model based on Pearson feature selection

Hailang Chen*, Xianfa Chang

School of EIE Heyuan Polytechnic Heyuan, Guangdong 517000, China

Received 4 September 2021; accepted 29 September 2021

Abstract

Accurate photovoltaic power prediction is the basis for realizing high-efficiency utilization of new energy in large-scale regional power grids. In order to deal with the influence and restriction of many factors such as ambient temperature, relative temperature and solar irradiance in the prediction of photovoltaic power generation, a photovoltaic power prediction method based on Pearson coefficient is proposed in this paper. In the prediction model, Pearson coefficients were used for correlation tests to remove irrelevant features. The remaining features were modeled using a long short-term memory network for regression prediction, and the final conclusions were drawn. The results of the algorithm show that the modified long short-term memory network has improved the mean absolute error and mean squared error of the predicted values. The prediction method, which can achieve short-term prediction of PV power and can reduce the impact of noise on PV power prediction. This research provides important support for the engineering application of energy internet related technologies to guarantee the stable operation of the power grid as well as to arrange reasonable dispatch.

Peer-review under responsibility of the scientific committee of the International Conference on Energy Engineering and Power Systems, EEPS, 2021.

Keywords: New energy; Machine learning; Correlation coefficient; Long short-term memory networks

1. Introduction

Energy is an important material basis for human activities and social development [1]. In the development of modern society, the excessive exploitation and use of fossil energy has caused energy shortage and environmental pollution, which has seriously restricted the global economic development and also brought the global problem of energy scarcity. In the past, the use of energy was mainly concentrated on traditional fossil energy such as oil and coal. Traditional fossil energy is non-renewable, and excessive development and utilization has led to a series of problems such as the deterioration of the ecological environment and the shortage of energy. Therefore, the development and utilization of renewable energy and the realization of sustainable energy development have become major measures of energy development strategies in all countries in the world.

* Corresponding author.
E-mail address: 110133@hypt.edu.cn (H. Chen).

https://doi.org/10.1016/j.egyr.2021.09.167

Peer-review under responsibility of the scientific committee of the International Conference on Energy Engineering and Power Systems, EEPS, 2021.

As one of the renewable energy sources with simple exploitation conditions and high commercial value, solar energy is receiving more and more attention from more and more countries. In 2019, the cumulative installed capacity of solar power in North America was 70048 MW, an increase of 11,209 MW from 2018. The cumulative installed capacity of solar power in Central and South America was 8750 MW, an increase of 1631 MW from 2018; the cumulative installed capacity of solar power in Europe was 146,666 MW, an increase of 2003 MW from 2018. The cumulative installed capacity of solar power in the Middle East is 5583 MW, an increase of 2271 MW from 2018. the cumulative installed capacity of solar power in Africa is 7236 MW, an increase of 1268 MW from 2018. the cumulative installed capacity of solar power in the Asia-Pacific region is 346,441 MW, an increase of 60401 MW from 2018. The new installed capacity of solar power in China is 53.06 GW, which is about 25 times of the installed capacity in 2011 and accounts for half of the new installed capacity worldwide, gradually becoming one of the indispensable new energy sources in China's energy system [8]. According to the forecast, the cumulative global PV installation is expected to reach 1721 GW by 2030 and further increase to 4670 GW by 2050 [9].

Solar energy in the new energy is widely distributed, easy to mine and pollution-free. It stands out among many new energy sources and is valued by more researchers. However, solar energy has the characteristics of uncertainty, fluctuation and randomness [10]. Therefore, it is particularly important to accurately predict photovoltaic power. Currently, more and more methods have been widely used in the field of PV power prediction. J Shi et al. elaborated and experimented on the feasibility of deep learning algorithms for PV power prediction based on the characteristics of power generation systems and PV arrays [11]. A Moosavi et al. performed uncertainty analysis through the relationship that exists between physical factors in numerical weather forecasting on the accuracy of the prediction [12]. Wang et al. built FOS- ELM based on the online sequential limit learning machine approach for short-term PV power prediction [13]. Q Zhang et al. proposed a support vector regression prediction model based on the cuckoo search algorithm for the problem of using the radiation intensity from the previous day's short-term prediction and considering that the support vector machine (SVM) of the learning parameters has a great impact on the generalization ability of the model [14]. JL Sánchez-García et al. proposed a hybrid model for PV power prediction by using statistical methods and combining the artificial neural network advantages [15]. M Rana et al. used correlated variables as the input data of neural networks and support vector regression algorithms for prediction [16]. R Aler et al. used support vector machines and gradient boosting regression methods to improve the accuracy of photovoltaic power prediction [17]. MK Behera et al. used an extreme learning machine (ELM) technology to predict photovoltaic power and compared its performance with the performance of existing models such as the BP prediction model [18].

It is not difficult to find from the literature that machine learning methods are increasingly used in the field of photovoltaic power prediction. Among them, machine learning algorithms include: BP neural network, support vector machines with good performance for regression problems, and gray model theory for small sample processing. Each algorithm has its advantages and disadvantages (as shown in Table 1). Therefore, it is inevitable to explore better photovoltaic power prediction models.

Table 1. Comparison of several common methods for photovoltaic power prediction.

Main methods	Advantages	Disadvantages
GM	Small samples and samples are irregularly distributed, and the amount of calculation is small, reflecting the overall trend of change [2].	Even if the prediction conditions are met, it may lead to a large error.
BP	It is used with empirical risk minimization principle with non-linear mapping capability, the ability of self-learning and generalizing [3].	Using gradient descent method with slow speed, it may enter the local minimum and training failure, needing a large number of samples for training [4].
SVM	With structural risk minimization principle, it is used by kernel functions instead of mapping in higher dimensional spaces based on nonlinear mapping theory [5].	In spite of large limitation, it can solve binary classification problems, but has difficulties in multi-classification problems.
RBF	It is characterized by non-linear mapping capability, parallel processing capability, and strong interpretation capability for mapping between input and output [6].	There is no way to explain the reasoning process for the insufficient data, and the network is unable to perform feedback information.
GVM	With the principle of structural risk minimization underlying the theory of nonlinear mappings, the mapping of higher dimensional spaces is replaced by the use of kernel functions	Gradient descent is slow, if the sample size is large, the training time is longer [7].

2. Correlative work

2.1. Overview of pearson coefficient

The Pearson coefficient was first proposed by statistician Carl Pearson and was originally used as a statistical indicator to analyze and study the degree of linear correlation between variables. There are several different correlation coefficients for different research subjects, and the most commonly used correlation coefficient is the Pearson correlation coefficient used in the paper. The correlation coefficients used in this paper are calculated by the product-difference method, i.e., first calculating the means of two variables separately, and then calculating the deviations between the two variables and their respective means separately, and multiplying the two deviations to obtain the correlation coefficients, which reflect the degree of correlation between the variables, and the Pearson coefficient is the most common one among many kinds of correlation coefficients. According to the characteristics of different correlated variables, different statistical indicators are selected for analysis. For example, if the statistical indicator of linear correlation is selected as the correlation coefficient and its square is used as the determination coefficient, the Pearson coefficient of this two-dimensional vector can be obtained by the formula.

$$r_{xy} = \frac{\sum_{i-1}^{m}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i-1}^{m}(x_i - \overline{x})^2}\sqrt{\sum_{i-1}^{m}(y_i - \overline{y})^2}} \tag{1}$$

In this formula, $\overline{x}$ is the mean of the elements in vector 1, $\overline{y}$ is the mean of the elements in vector 2. r_{xy} is the degree of correlation between different variables, m is the number of data in the sequence.

2.2. LSTM principles and applications

Long short-term memory (LSTM) was first proposed by Hochreiter et al. [19], and in 2000 Schmidhuber et al. improved the LSTM network by proposing the method of the forget gate, which is applicable to the prediction of continuity [20,21]. The LSTM was then improved and generalized in Grave's book [22]. LSTM has been quite successful and widely used in many issues [23]. the predecessor of LSTM neural networks is recurrent neural networks. Recurrent neural networks (RNNs) are neural networks that learn sequential patterns through internal loops. there are many recurrent loops in an RNN network, which can pass information on continuously. The learning and adjustment of weights adopts the chain rule back propagation. When the value is propagated back to the activation function, such as Sigmoid and Tanh functions, the slope will become extremely small (or extremely large), and the problem of gradient disappearance (or gradient explosion) occurs. The LSTM model was developed to avoid these problems. Hochreiter et al. proposed memory cells and gates. Such a structure can store information for a long time while forgetting unnecessary information.

LSTM networks use memory cells instead of neurons. Fig. 1 is a schematic diagram of LSTM memory cell. An LSTM cell consists of one memory cell (c_t) and three gate structures, including input gate (i_t), forget gate (f_t), output gate (o_t). At the moment t, x_t, t represents the input data, h_t represents the hidden layer state. The symbol $\times$ represents the vector outer product and the symbol $+$ represents the superposition operation. The operation formula of LSTM is shown in Eq. (2)-equation (7). Where, U and W represent matrix weights, b represents offset, σ is sigmoid function, and the symbol * represents vector outer product.

$$f_t = \sigma\left(U_f x_t + W_f H_{t-1} + b_f\right) \tag{2}$$

$$i_t = \sigma\left(U_i x_t + W_i H_{t-1} + b_i\right) \tag{3}$$

$$u_t = \tanh\left(U_u x_t + W_u H_{t-1} + b_u\right) \tag{4}$$

$$c_t = f_t \times C_{t-1} + i_t \times u_t \tag{5}$$

$$o_t = \sigma\left(U_o x_t + W_o H_{t-1} + b_u\right) \tag{6}$$

$$h_t = o_t \times \tanh\left(c_t\right) \tag{7}$$

The forget gate calculates the weighted sum of t, ht-1, bf, and obtains f_t ($f_t \in (0, 1)$) through the sigmoid function, as shown in Eq. (2). f_t represents the weight of the information that needs to be forgotten in the last memory cell

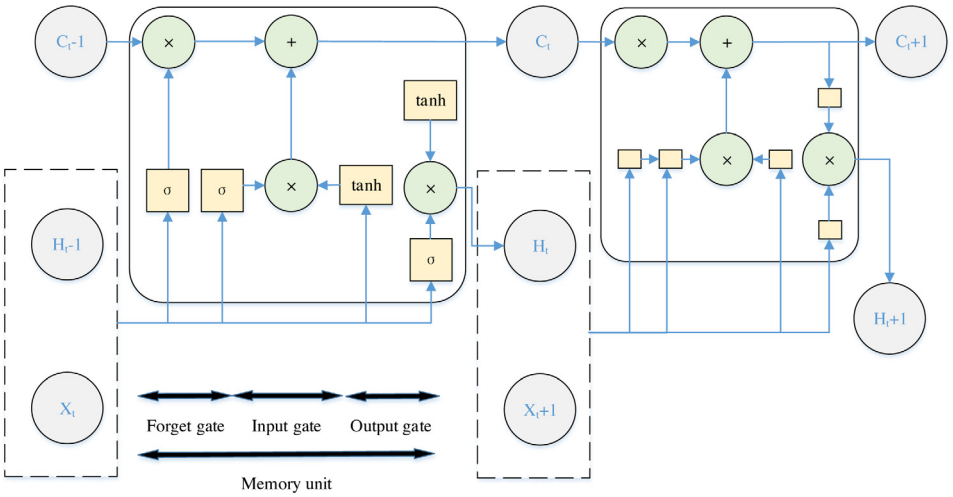

Fig. 1. LSTM memory unit.

(Ct-1). In other words, the forget gate is used to control the amount of information retained in the previous memory cell, as shown in Eq. (5). The input gate determines how much new information to receive to the memory cell (Ct)) (see Eq. (3)). Ct is the storage weight of memory cells (see Eq. (4)). The original information and the new information are controlled by the forget gate and the input gate respectively, and the current memory cell (Ct) is obtained (see formula (5)). Finally, filter the memory cells (Ct) through the output gate of formula (6). The updated memory cell obtains the current hidden layer state h_t through equation (7). Finally, back-propagation is performed to obtain the LSTM model composed of these storage blocks. LSTM makes LSTM training robust through multi gate cooperation and avoids gradient dispersion [24].

3. LSTM model based on pearson feature selection

3.1. Overall design of the model

In order to deal with the problem that the traditional neural network needs a large number of sample training and local optimal solution, this paper introduces the LSTM model with long-term memory function, good generalization ability and suitable for the problems highly related to time series [25]. At the same time, in order to improve the prediction effect of LSTM, this paper proposes an LSTM model based on Pearson feature selection (Fig. 2). Firstly, the model preprocesses the original data by means interpolation and normalization to form regular, orderly and complete data. Then, the Pearson feature is used to select the data to filter, and the filtered data is used as the input of LSTM model. Through continuous training and learning, the final prediction model is obtained.

3.2. Model validation

In this paper, the actual data of photovoltaic metering station in H city is used to verify the model. The data collection interval was once per hour. 3142 sets of data were used as the data training set from January to December 2019 (among which, 1320 sets of data were collected for 120 days of sunny days and 825 sets of data were collected for 75 days of cloudy days 3142 sets of data (of which 141 sets of data for 16 days for sunny days and 88 sets of data for 10 days for cloudy days) are available for the data test set in January 2020. The data composition categories are temperature, wind speed, temperature of PV panels, humidity, total radiation, barometric pressure and power indicators. The prediction target is the PV power output for the next hour. To facilitate the analysis of the accuracy of the prediction results, the mean relative error MAPE and the root mean square error RMSE are used as the judgment indexes of the error magnitude, and their calculation procedures are shown in Eqs. (8) and (9).

$$\mathrm{MAE} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \left| Y_I - Y_i' \right| \tag{8}$$

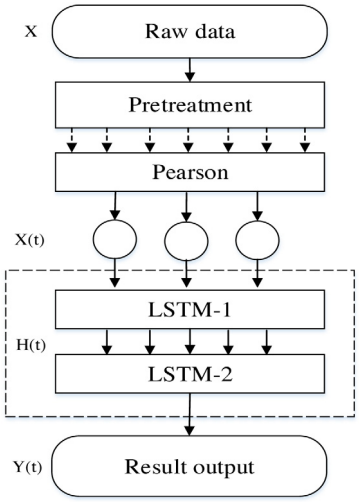

Fig. 2. LSTM model based on Pearson feature selection.

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left(Y_I - Y_i'\right)^2} \tag{9}$$

The average absolute error represents the size of the average difference between the observed value and the predicted value. The smaller the value, the smaller the difference. The degree of error dispersion between the predicted value and the true value is represented by the root mean square error. The smaller the root mean square error, the smaller the degree of difference. Both of the above evaluation indicators can effectively and accurately evaluate the prediction results of the model.

3.3. Pearson feature selection

Photovoltaic power generation refers to the conversion of irradiated light from the sun through photovoltaic panels to produce electricity. Photovoltaic output power is affected by many factors, in addition to the physical factors of the photovoltaic panels themselves, including external factors such as ambient temperature, solar radiation intensity and

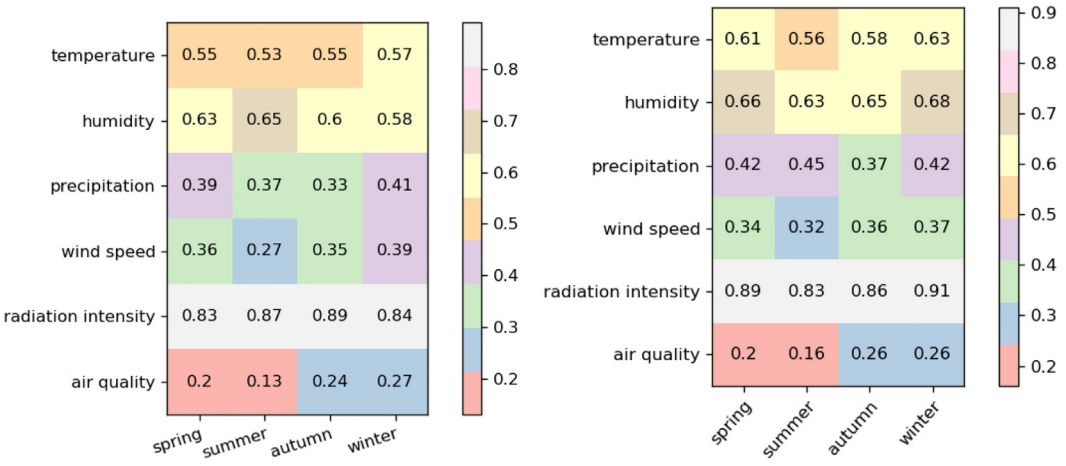

Fig. 3. (a) Correlation analysis of sunny days; (b) Correlation analysis of cloudy days.

weather conditions. According to the analysis of the data sample collection area, collection period and environmental conditions, the weather type is mainly sunny and cloudy days. Therefore, in the paper, we selected temperature, humidity, precipitation, wind speed, sunshine radiation intensity and air quality as the explanatory variables, and then, all explanatory variables were classified using Pearson coefficients to find out the high, moderate, weak and irrelevant variables among the explanatory variables, and their coefficient results are shown in Fig. 3. (Pearson coefficients of characteristic variables greater than 0.8 are highly correlated. (Between 0.5 and 0.8 is moderate correlation, between 0.3 and 0.5 is weak correlation, and less than 0.3 is almost no correlation.)

It can be seen from Fig. 3 that under different weather, the influence of solar irradiance on photovoltaic power is still the most important factor, and Pearson's value is highly correlated. In different seasons, the impact on photovoltaic power will vary to varying degrees, but the trend of change can be consistent. According to the correlation, precipitation and wind speed are almost irrelevant, indicating that the two explanatory variables are not directly or indirectly related to the PV power of the explained variable. Therefore, in this paper, solar radiation intensity, humidity and temperature are used as the final explanatory variables of the model.

4. Results of the experiment

4.1. Analysis of predicted effects

In this paper, the long and short-term memory network is used for prediction, and the prediction result is shown in Fig. 4. It can be seen from the figure that in sunny and cloudy weather conditions, the predicted results can better fit the true value. The RMSE and MAE of the error indicators under sunny conditions are 0.189 (10 MW) and 0.121 (10 MW), respectively. Under cloudy conditions, the predicted value is very close to the observed value trend, the error index MAE is 0.154 (10 MW), and the RMSE is 0.181 (10 MW).

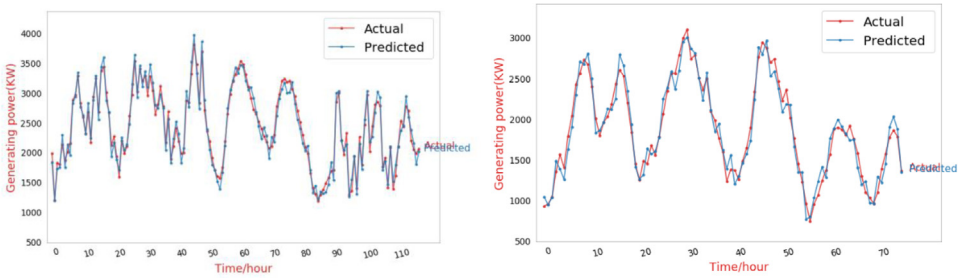

Fig. 4. (a) Power prediction in sunny days; (b) Power prediction in cloudy days.

4.2. Comparison of experimental data

In order to highlight the advantages of the algorithm in this paper, the Long short-term memory (LSTM) algorithm, neural network algorithm (Back Propagation, BP), radial basis function neural network (Radial Basis Function, RBF) and The time series algorithm (Time Series, TS) predicts the photovoltaic power, and the results are shown in Table 2. As can be seen from the data presented in Table 2, the algorithms in the paper all obtained the highest accuracy when predicting PV power generation due to the ability of the long and short term memory network to enhance the generalization and robustness of the model. Among the other algorithms, the BP neural network algorithm is susceptible to local minima in the calculation process, which makes the overall generalization not high; the radial basis function neural network is not sufficient for data filtering and information extraction, so its prediction accuracy is still lower than the method used in the paper; the time series algorithm is only based on the historical data of PV power for power prediction, so its prediction accuracy is usually lower than the other The time series algorithm only predicts power based on historical data of PV power, so its prediction accuracy is usually lower than other algorithms. In summary, the LSTM prediction model used in this paper has a high prediction accuracy and can simulate the trend of PV power better.

Table 2. Load and PV output forecasting error evaluation considering multiple models.

Weather type	Prediction	MAE (10 MW)	RMSE (10 MW)
Sunny	LSTM	0.189	0.121
	BP	0.233	0.443
	RBF	0.378	0.684
	TS	0.417	0.556
Cloudy	LSTM	0.154	0.136
	BP	0.236	1.108
	RBF	0.256	1.163
	TS	1.022	1.089

5. Conclusion

In this paper, we use the actual situation of PV power plant in H city to analyze the PV power curve, use the model based on long and short term memory network to fit the PV power curve, give full play to the advantages of the model to analyze the correlation of PV power, and propose a PV power prediction method based on Pearson feature selection for long and short term memory network. The method uses Pearson coefficients to analyze the influence of external conditions on the variation of PV power, and the model is validated by case tests. The results show that the intensity of insolation radiation, temperature and humidity influence factors play a decisive role in the variation of PV power. LSTM is compared with BP, RB and TS algorithms, and the accuracy of PV power prediction based on LSTM algorithm is better.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Fund for Guangdong Provincial Department of Education (No. 2020KQNCX216) and National Institute of Computer Basic Education Research Association (No. 2020-AFCEC-202) and Key areas fund project of Guangdong Provincial Education Department in 2020 (No. 2020zdzx3088). We would like to thank all sponsors and reviewers for their detailed comments on our paper.

References

[1] Yusheng Xue, Chen Yu, Junhua Zhan, et al. Review on short-term and ultra-short-term wind power prediction. Autom Electr Power Syst 2015;39(06):141–51.
[2] Xu Z, Liu F. Review of GM model optimization research progress. Comput Sci 2016;43(S2):6–10.
[3] Capela F, Nouchi V, Van Deursen R, et al. Multitask learning on graph neural networks applied to molecular property predictions. 2019, arXiv preprint arXiv:1910.13124.
[4] Baozhou Huang, Junhua Yang, Siling Lu, et al. Based on improved particle swarm optimization neural network complex algorithm captured by the wave power prediction. J Solar Energy 2021;(2):302–8.
[5] Vapnik VN. The nature of statistical learning theory. New York: Springer-Verlag; 1995.
[6] Gu L, Siong Tok DK, Yu Ding-li. Development of adaptivep-step RBF network model with recursive orthogonal least squares training. Neural Comput Appl 2018;29(5):1445–54.
[7] Zhao H. General vector machine. 2016.
[8] Guiying Shen. China's installed photovoltaic power generation reached 180 million kW in the first quarter of 2019. Fine Specialty Chem 2019;(6):5–6.
[9] Shaoping Guo. Solar photovoltaic power generation development status and prospects, vol. 16. Shandong Industrial Technology; 2018, p. 163.
[10] Kim GG, Choi JH, Park SY, et al. Prediction model for PV performance with correlation analysis of environmental variables. IEEE J Photovolt 2019;9(3):832–41.
[11] Shi J, Zhang J. Ultra short-term photovoltaic refined forecasting model based on deep learning. Electric Power Constr 2017.
[12] Moosavi A, Rao V, Sandu A. A learning based approach for uncertainty analysis in numerical weather prediction models. 2018.
[13] Wang J, Ran R, Zhou Y. A short-term photovoltaic power prediction model based on an FOS-ELM algorithm. Appl Sci 2017;7(4):423.

[14] Zhang Q, Xiaomei WU, Tian M, et al. Prediction of radiation intensity for photovoltaic power plant based on cuckoo search optimization. Ningxia Electric Power; 2017.
[15] Sánchez-García JL, Espinosa-Juárez E, Flores JJ. Short term photovoltaic power production using a hybrid of nearest neighbor and artificial neural networks. In: Transmission & distribution conference and exposition-Latin America. IEEE; 2017, p. 1–6.
[16] Rana M, Koprinska I, Agelidis VG. Univariate and multivariate methods for very short-term solar photovoltaic power forecasting. Energy Convers Manage 2016;121:380–90.
[17] Aler R, Martín R, Valls JM, et al. A study of machine learning techniques for daily solar energy forecasting using numerical weather models. In: Intelligent distributed computing, vol. VIII, Springer International Publishing; 2015, p. 269–78.
[18] Behera MK, Majumder I, Nayak N. Solar photovoltaic power forecasting using optimized modified extreme learning machine technique. Eng Sci Technol 2018.
[19] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[20] Gers FA, Schmidhurber Jurgen, et al. Learning to forget: Continual prediction with LSTM. Neural Comput 2000;12(10):2451–71.
[21] Qader K, Rehman WU, Sheri AM, et al. A long short term memory (LSTM) network for hourly estimation of PM2.5 concentration in two cities of South Korea. Appl Sci 2020;10(11):3984.
[22] Graves A. Supervised sequence labelling with recurrent neural networks. Springer; 2012.
[23] Nanyun Peng, Poon H, Quirk C, et al. Cross-sentence n-ary relation extraction withgraph LSTMs. Trans Assoc Comput Linguist 2017;5(1):101–15.
[24] Jain G, Sharma M, Agarwal B. Optimiziong semantic LSTM for spam detection. Int J Inf Technol 2019;11(2):239–50.
[25] Peng W, Xin B. SPMF: A social trust and preference segmentation-based matrix factorization recommendation algorithm. EURASIP J Wirel Commun Netw 2019;272–84.