Application of Machine Learning Method in Jinan Temperature Prediction

Mingzhao Liu and Zhongmu Li*
Institute of Astronomy, Dali University, Dali 671003, China

Keywords: Machine Learning, LSTM, Temperature Prediction.

Abstract: Machine learning algorithms based on big data have been widely used in many fields, such as healthcare, education, manufacturing, and finance. It can extract features from existing data, learn the changing laws of existing data, and then judge and predict new data. It is practical to apply machine learning to temperature prediction. This paper uses a machine learning algorithm and Jinan temperature data to build a prediction model, and compares the predicted with the actual value. The prediction accuracy of the model is 90.1%. The results show that the model can predict the daily maximum temperature in Jinan, and the machine learning method has good applicability in the field of temperature prediction.

1 INTRODUCTION

With the increasing severity of global warming, people gradually realize the importance of temperature in daily production and life. Extreme high or low temperatures will affect the safety of people's lives and property. In particular, persistent heat has melted Arctic glaciers and raised sea levels, causing irreversible damage to coastal cities. In recent years, the frequency of extreme temperature phenomenon in Jinan is getting higher and higher. As the provincial capital, Jinan has a large population density, and the damage caused by extreme high or low temperature is particularly obvious. To better cope with temperature changes, it is necessary to predict the future temperature. Since ancient times, people have been looking for ways to predict temperature, hoping to predict temperature to a certain extent through efforts. Temperature prediction belongs to a climate factor prediction in climate prediction. At present, there are two main methods for climate prediction, namely statistical and dynamic methods. Statistical method refers to the use of statistical methods to analyze the linear relationship between various prediction factors. Common statistical methods include multiple linear regression (Cannon and McKendry, 2002; Mekanik et al., 2013), singular spectrum analysis (Chau and Wu, 2010), singular value decomposition (Yun et al., 2002; Fattorini and Brandini, 2002) and grey prediction model (Kung et al., 2003), etc. Dynamic method refers to a numerical model that adds initial conditions to partial differential equations based on physical laws, and solves to obtain future climate change. Although, statistical methods and dynamic methods are the two main methods of climate prediction, the shortcomings of these two methods are very obvious. The statistical method lacks the research on the physical mechanism, and the prediction ability of extreme values is insufficient. The dynamic method is sensitive to the initial value and cannot make full use of historical data.

In recent years, artificial intelligence has developed rapidly. Machine learning (Sammut and Webb, 2010) and deep learning (Lu et al., 2014) technology are the core content of artificial intelligence (Jordan and Mitchell, 2015), and they are widely used in various fields. The development of artificial intelligence promotes the generation of prediction methods based on data-driven models (Lu et al., 2014). Machine learning can automatically learn the patterns of changes in existing data and apply these patterns to new data. The training of machine learning model can not be separated from a large number of data. The accuracy of the model prediction depends on the amount of data. The four basic characteristics of big data are the amount of data, the speed of data update, the diversity of data, and the accuracy of data. The meteorological data has already these four characteristics (Markus et al., 2019). Therefore, machine learning is very suitable for climate prediction. Common machine learning algorithms

Liu, M. and Li, Z.
Application of Machine Learning Method in Jinan Temperature Prediction.
DOI: 10.5220/0011920100003612
In *Proceedings of the 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 252-255
ISBN: 978-989-758-622-4; ISSN: 2975-9463

include random forest algorithm (Simone et al., 2011), artificial neural network (Agatonovic-Kustrin et al., 2000), support vector machine (Tripathi et al., 2006), logistic regression (Menard, 2004) and long short-term memory network, etc. In this paper, we use a representative method in machine learning, long short-term memory network. The network and Jinan climate data are used to build the model. The model is then used to predict the daily maximum temperature in Jinan. We calculate the error between the predicted and the real value to evaluate the model performance.

2 DATA AND MODELS

2.1 Data Sources

The data used in this article comes from the daily observation data of China's surface meteorological stations of the "National Meteorological Science Data Center". We selected the daily maximum temperature data of Jinan station from year 1951 to 2019 for research. The temperature data characteristics are shown in Figure 1. Temperature data is a typical time series data with obvious periodic attribute. We input it to the prediction model, and the model automatically extracts data features and makes predictions.

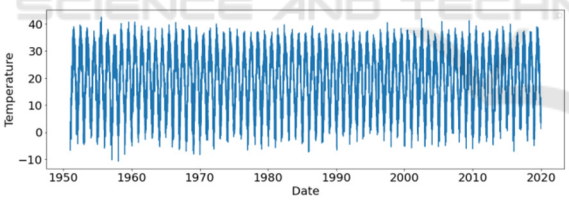

Figure 1: Temperature change graph

2.2 Data Processing

We use the info function in Python to view the relevant information of the sample data. Then we call the MinMaxScaler function in the sklearn.preprocessing package to normalize the sample data and transform the sample data to be between (0, 1). The data are selected from year 1951 to 2012 as the training set, with a total of 22,639 pieces of data; the data from year 2013 to 2019 are selected as the test set, with a total of 2,556 pieces of data. Missing data is not much in the selected data, which has little impact on model training. Therefore, we will directly delete the missing data. After the data are divided, the training data are directly passed to the model, and the model is predicted after training. The predicted value is compared with the test data, and the Explained Variance Score (EVS), Mean Absolute Error (MAE) and Mean Squared Error (MSE) are calculated.

2.3 Long Short-Term Memory Principle

The long short-term memory network model is a special form of the recurrent neural network. Recurrent neural network is a kind of neural network with memory ability, and its main purpose is to process sequence data. The innovation of recurrent neural network is to store the past information in the memory unit and produce output after the interaction with the current input. However, when the amount of data increases and the data has long-term dependence, the recurrent neural network will have the problem of gradient explosion and gradient disappearance. To solve this problem, long short-term memory network was created. The long short-term memory network effectively makes up for the shortcomings of the recurrent neural network by introducing a special gating mechanism. It was first developed and applied by two scientists, Schmidhuber and Hochreiter, in 1997, and has received much attention in recent years and has been widely used in other works. The storage of the long short-term memory network is called the gating unit. The function of the gating unit is like a "gate", controlling the retention or discarding of data and allowing information to pass through selectively. The long short-term memory network has three gated units, i.e., forget gate, input gate and output gate. The three gates respectively control the output of various information. The update process is from Formula (1) to Formula (6).

$$G_f = \text{sigmoid}(w_f [H_{t-1}, X_t]) \quad (1)$$

$$G_i = \text{sigmoid}(w_i [H_{t-1}, X_t]) \quad (2)$$

$$S_t = \tanh(w_s [H_{t-1}, X_t]) \quad (3)$$

$$C_t = G_f C_{t-1} + G_i S_t]) \quad (4)$$

$$G_o = \text{sigmoid}(w_o [H_{t-1}, X_t]) \quad (5)$$

$$H_t = G_o * \tanh(C_t) \quad (6)$$

3 MODEL PREDICTION AND RESULT ANALYSIS

3.1 Experimental Platform Construction

This research uses Anaconda to build a Python3.8 environment, and uses the Tensorflow framework developed by Google to complete the construction of the entire model. The training process of machine learning model is completed on the Jupyter notebook provided by Anaconda. Anaconda is a software for installing and managing Python-related packages. It can be used to install different versions of packages and their dependencies on the same machine and switch between different environments. Tensorflow is a deep learning framework released by Google in 2015, and released version 2.0 in 2019 to make up for the defects.

3.2 Construction of Prediction Model

We use the long short-term memory network and the highest temperature data of Jinan from January 1, 1951 to December 31, 2019 to model. The network model constructed in this experiment consists of three LSTM layers, three Dropout layers and a Dense layer, where the nodes of the LSTM layer are 64, 50 and 50 respectively. The batch_size and the number of training epochs are set to 32 and 30. The optimization algorithm of the model uses Adam. We use the trained model to predict the daily maximum temperature in Jinan, and analyze the EVS, MAE and MSE of the predicted value and the real value. The results show that the model can predict the daily maximum temperature in Jinan with high accuracy.

3.3 Evaluation Standard

The purpose of prediction is to predict the future time. To analyze whether the prediction results are correct, the model needs to be evaluated after training. Model evaluation is an integral part of the model development process. It helps to discover the best model for expressing data and how well the selected model will work in the future. We need to develop evaluation criteria to evaluate the prediction effect of the model. These criteria can explain the overall prediction performance of the prediction model and compare the advantages and disadvantages of the prediction methods. In this paper, the training set data is used for model construction and training, The training set data is not involved in model evaluation. The test set data is not involved in model building, it can be used to evaluate the accuracy of model predictions. This study needs to evaluate the accuracy of the temperature prediction results of the machine learning method, the selected evaluation criteria are Explained Variance Score (EVS), Mean Absolute Error (MAE) and Mean Squared Error (MSE). The calculation formulas of these evaluation criteria are shown in Formula (8) to Formula (10).

$$EVS = 1 - \frac{var(y - y\hat{})}{var(y)} \tag{8}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y - y\hat{}| \tag{9}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y - y\hat{})^2 \tag{10}$$

3.4 Result Analysis

The model was trained by training set data to predict the daily maximum temperature in Jinan City. We compare the prediction results with the test set data. The fitting results are shown in Figure 2.

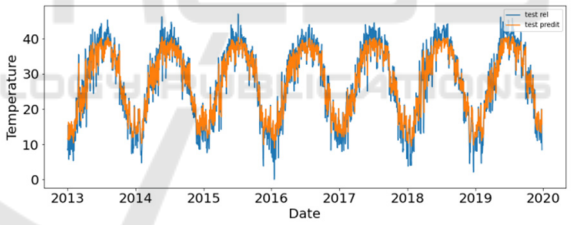

Figure 2: Comparison of predicted and actual results

It can be seen from the Figure 3 and Figure 4 that the fitting of the model is good. MSE and MAE decreased with the increase of training times.

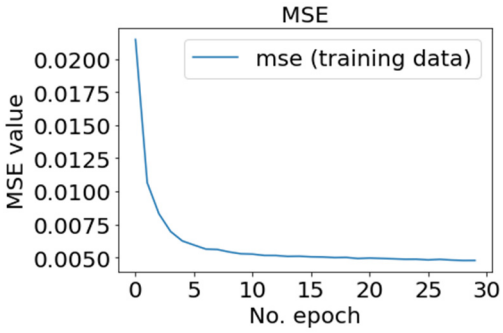

Figure 3: Mean Squared Error versus number of training sessions

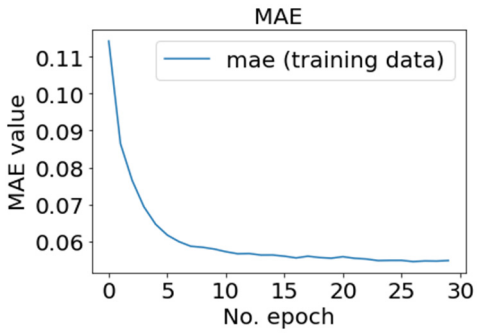

Figure 4: Mean Absolute Error versus number of training sessions

4 CONCLUSION

This paper uses a machine learning method to predict the daily maximum temperature in Jinan City. Via the comparison between the predicted and the actual value, it is found that the long short-term memory network has a good performance in fitting the prediction of the daily maximum temperature of Jinan City. The accuracy rate reaches 90%, and the values of MSE and MAE are 0.0049 and 0.0565. Therefore, long short-term memory has good prospects in temperature prediction and climate prediction, and it is a trend to apply machine learning methods to climate prediction. Temperature is characterized by randomness and uncertainty, and many climatic factors affect temperature. Therefore, the follow-up research can consider the impact of more climate factors on the temperature, and then better predict the temperature.

ACKNOWLEDGEMENTS

This work is supported by the Yunnan Academician Workstation of Wang Jingxiu (202005AF150025), National Natural Science Foundation of China (No. 11863002), and Sino-German Cooperation Project (No. GZ 1284).

REFERENCES

Agatonovic-Kustrin, S., and Beresford, R. (2000). Basic concepts of artificial neural network (ann) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical & Biomedical Analysis*, 22(5), 717-727.

Chau, K., Wu, C. (2010). A hybrid model coupled with singular spectrum analysis for daily rainfall prediction. *Journal of Hydroinformatics*, 12(4), 458-473.

Cannon, A. J., and Mckendry, I. G. (2010). A graphical sensitivity analysis for statistical climate models: application to indian monsoon rainfall prediction by artificial neural networks and multiple linear regression models. *International Journal of Climatology*, 22.

Fattorini, M., and Brandini, C. (2020). Observation Strategies Based on Singular Value Decomposition for Ocean Analysis and Forecast. *Water*, 12(12): 3445.

Jordan, M. I., and Mitchell, T. M. (2015). machine learning: trends, perspectives, and prospects. *Science*, 349(6245):255-260.

Kung, C. Y., Kung, C. J., and Tsai, S. Y. (2003). Study of computer game forecasting in Taiwan market application of grey prediction model.

Lu, C., Lei, Y., Singh, V., et al. (2014). Determination of input for artificial neural networks for flood forecasting using the copula entropy method. *Journal of Hydrologic Engineering*, 19(11), 217-226.

Lu, C., Lei, Y., Singh, V., et al. (2014). Determination of input for artificial neural networks for flood forecasting using the copula entropy method. *Journal of Hydrologic Engineering*, 19(11), 217-226.

Markus, et al. (2019). Deep learning and process understanding for data-driven earth system science. *Nature.*

Menard, S. (2004). Logistic regression. *American Statistician*, 58(4), 364.

Mekanik, F., Imteaz, M. A., Gato-Trinidad, S., and Elmahdi, A. (2013). Multiple regression and artificial neural network for long-term rainfall forecasting using large scale climate modes. *Journal of Hydrology, 503*(503), 11-21.

Sammut, C., and Webb, G. I. (2010). Machine learning. *Kluwer Academic Publishers*, 10.1007/978-0-387-30164-8(Chapter 89), 139-139.

Simone, V., Matteo, Z., et al. (2011). Application of a random forest algorithm to predict spatial distribution of the potential yield of ruditapes philippinarum in the venice lagoon, italy. *Ecological Modelling.*

Tripathi, S., Srinivas, V. V., and Nanjundiah, R.S. (2006). Downscaling of precipitation for climate change scenarios: a support vector machine approach. *Journal of Hydrology*, 330(3-4), 621-640.

Yun, W. T., Stefanova, L., and Krishnamurti, T. N. (2003). Improvement of the multimodel superensemble technique for seasonal forecasts. *Journal of Climate*, 16(22): 3834-3840.