Review on Air Quality Prediction Using ARIMA and Neural Network

Avinash S B[1], Chaluvaraj M[2], N V Devanand[3], Raju H[4], Mrs. M G Kousar[5]

[1,2,3,4]*B.E. Student, Department of CSE, Sir M Visvesvaraya Institute of Technology, Bengaluru, India*

[5]*Assistant Professor, Department of CSE, Sir M Visvesvaraya Institute of Technology, Bengaluru, India*

---***---

Abstract - *In recent years, people have been paying more and more attention to air quality because it directly affects people's health and daily life. Effective air quality prediction has become one of the hot research issues. However, this paper is suffering many challenges, such as the instability of data sources and the variation of pollutant concentration along time series. Aiming at this problem, we propose an improved air quality prediction method based on the ARIMA model to predict the PM2.5 concentration at the 35 air quality monitoring stations in Beijing over the next 24 h. In this paper, we resolve the issue of processing the high-dimensional large-scale data by employing the ARIMA model and innovatively take the forecasting data as one of the data sources for predicting the air quality. With exploring the forecasting data feature, we could improve the prediction accuracy with making full use of the available spatial data. Given the lack of data, we employ the sliding window mechanism to deeply mine the high-dimensional temporal features for increasing the training dimensions to millions. We compare the predicted data with the actual data collected at the 35 air quality monitoring stations in Beijing. The experimental results show that the proposed method is superior to other schemes and prove the advantage of integrating the forecasting data and building up the high-dimensional statistical analysis.*

***Key Words*: Instability, Prediction, Employing, Forecasting, Data Sources, Predicted Data, Actual Data.**

1.INTRODUCTION

In recent years, people are beginning to pay more and more attention to the impact of the environment on health, and the information related to air quality has become the focus of people's daily life. The existing air quality monitoring instruments, stations and satellite meteorological data can provide real-time air quality monitoring information. However, this is far from sufficient, and it is entirely necessary to predict the trend of air pollutants in the future. Currently, the forecast data on weather conditions is of high reliability and accuracy. Based on this, we propose to fuse the predictive data, i.e., the forecast data on weather conditions, with the available air quality historical data and meteorological data, supported by machine learning means, to explore mining data correlation and build a well-performed model of predicting the future air quality conditions. This contribution enables an efficient solution to construct a predictive data feature exploration-based air quality prediction approach with the improved performance. The primary goal of air quality prediction is to predict the concentration of pollutants for a while in the future based on historical air quality data sets, meteorological data sets, etc., such as the work proposed .By learning the previous research results, we found that the existing methods are employing the historical data-based prediction, using some neural networks, such as LSTM proposed in machine learning based solution proposed .Extreme Learning Machine (ELM) in the simple regression methods. However, on the one hand, such the methods failed to make full use of the existing air quality big data for deeply mining the temporal features and statistical data features. On the other hand, the simple regression methods are less efficient in processing high-dimensional big data and cause the performance of the model accuracy relatively limited. It can be seen that the existing methods have some certain limitations and it is necessary to carry out further research. Nowadays, as meteorological data measurements become more accurate and predictive data begins to be highly reliable, it exhibits considerable mining value. If the predictive data can be effectively combined with the historical data, the prediction effect will be significantly improved. Based on the above considerations, this paper selects to employ the model that is suitable for processing high-dimensional data and supporting the parallel learning, namely LightGBM combined with the historical datasets to predict the air quality. With this method, we use the historical air quality data within the latest 144 hours and the future 24-hour weather forecast data to carry out the time-related feature mining and to construct the relevant statistical features, and then we are enabled to predict the PM2.5 concentration at the 35 air quality monitoring stations in Beijing. After preprocessing the data, we use the sliding window mechanism to increase the feature dimension to 2262 and expand the data volume to millions of items, through which a higher accurate prediction model could be established.

In the process of conducting air quality predictions, we are facing many challenges. First, the air quality is always affected by a variety of factors, such as traffic factors,big events, etc. Such the impact factors are difficult to acquire or model in advance. Second, the air quality exhibits high uncertainty in the time dimension .PM2.5 pollutant concentration values of the given air quality monitoring station at the same moment along the two days is hugely different. Moreover, even during the same day, the PM2.5 concentration value varies a lot, and the difference between the highest and lowest concentration values could reach 100 (mg/m3). Third, the geographical distribution of the air quality monitoring stations presents a significant difference among the pollutant concentration values. Within the same day, the pollutant concentration curves of three different monitoring points in Beijing showed significant differences due to their different locations.

2. LITERATURE SURVEY

1.Nowadays more and more urban residents are aware of the importance of the air quality to their health, especially who are living in the large cities that are seriously threatened by air pollution. Meanwhile, being limited by the spare sense nodes, the air quality information is very coarse in resolution, which brings urgent demands for high-resolution air quality data acquisition. In this paper, we refer the real-time and fine-gained air quality data in city-scale by employing the crowd source automobiles as well as their built-in sensors, which significantly improves the sensing system's feasibility and practicability. The main idea of this paper is motivated by that the air component concentration within a vehicle is very similar to that of its nearby environment when the vehicle's windows are open, given the fact that the air will exchange between the inside and outside of the vehicle though the opening window. Therefore, this paper first develops an intelligent algorithm to detect vehicular air exchange state, then extracts the concentration of pollutant in the condition that the concentration trend is convergent after opening the windows, finally, the sensed convergent value is denoted as the equivalent air quality level of the surrounding environment. Based on our Internet of Things cloud platform, real-time air quality data streams from all over the city are collected and analyzed in our data center, and then a fine-gained city level air quality map can be exhibited elaborately. In order to demonstrate the effectiveness of the proposed method, experiments crowd sourcing 500 floating vehicles are conducted in Beijing city for three months to ubiquitously sample the air quality data. Evaluations of the algorithm's performance in comparison with the ground truth indicate the proposed system is practical for collecting air quality data in urban environment.

2.With the rapid development of urbanization and industrialization, many developing countries are suffering from heavy air pollution. Governments and citizens have expressed increasing concern regarding air pollution because it affects human health and sustainable development worldwide. Current air quality prediction methods mainly use shallow models; however, these methods produce unsatisfactory results, which inspired us to investigate methods of predicting air quality based on deep architecture models. In this paper, a novel spatiotemporal deep learning (STDL)-based air quality prediction method that inherently considers spatial and temporal correlations is proposed. A stacked autoencoder (SAE) model is used to extract inherent air quality features, and it is trained in a greedy layer-wise manner. Compared with traditional time series prediction models, our model can predict the air quality of all stations simultaneously and shows the temporal stability in all seasons. Moreover, a comparison with the spatiotemporal artificial neural network (STANN), auto regression moving average (ARMA), and support vector regression (SVR) models demonstrates that the proposed method of performing air quality predictions has a superior performance.

3.Exposure to high concentrations of fine particulate matter ($PM_{2.5}$) can cause serious health problems because $PM_{2.5}$ contains microscopic solid or liquid droplets that are sufficiently small to be ingested deep into human lungs. Thus, daily prediction of $PM_{2.5}$ levels is notably important for regulatory plans that inform the public and restrict social activities in advance when harmful episodes are foreseen. A hybrid EEMD-GRNN (ensemble empirical mode decomposition-general regression neural network) model based on data preprocessing and analysis is firstly proposed in this paper for one-day-ahead prediction of $PM_{2.5}$ concentrations. The EEMD part is utilized to decompose original $PM_{2.5}$ data into several intrinsic mode functions (IMFs), while the GRNN part is used for the prediction of each IMF. The hybrid EEMD-GRNN model is trained using input variables obtained from principal component regression (PCR) model to remove redundancy. These input variables accurately and succinctly reflect the relationships between $PM_{2.5}$ and both air quality and meteorological data. The model is trained with data from January 1 to November 1, 2013 and is validated with data from November 2 to November 21, 2013 in Xi'an Province, China. The experimental results show that the developed hybrid EEMD-GRNN model outperforms a single GRNN model without EEMD, a multiple linear regression (MLR) model, a PCR model, and a traditional autoregressive integrated moving average (ARIMA) model. The hybrid model with fast and accurate results can be used to develop rapid air quality warning systems.

4.Learning to store information over extended time intervals via recurrent back propagation takes a very long time, mostly due to insufficient, decaying error back flow. We brie y review Hochreiter's 1991 analysis of this problem, then address it by introducing a novel, efficient, gradient-based method called Long Short-Term Memory" (LSTM). Truncating the gradient where this does not do harm, LSTM can learn to bridge minimal time lags in excess of 1000 discrete time steps by enforcing constant error flow through constant error carrousels" within special units. Multiplicative gate units learn to open and close access to the constant error flow. LSTM is local in space and time; its computational complexity per time step and weight is O(1). Our experiments with artificial data involve local, distributed, real-valued, and noisy pattern representations. In comparisons with RTRL, BPTT, Recurrent Cascade-Correlation, Elman nets, and Neural Sequence Chunking, LSTM leads to many more successful runs, and learns much faster. LSTM also solves complex, artificial long time lag tasks that have never been solved by previous recurrent network algorithms.

5.This work introduces a new technique that provides real-time feedback to a Heating, Ventilation, and Air Conditioning (HVAC) system controller with respect to the occupants' thermal preferences to avoid space overheating. We propose a non-invasive approach for automatic prediction of personal thermal comfort and mean time to warm discomfort using machine learning. The prediction framework described uses temperature information extracted from multiple local body parts to model an individual's thermal preference, with sensing measurements that capture local body part variance as well as differences between body parts. We compared the efficacy of using machine learning with classical measurements such as skin temperature along with our approach of using multi-part measurements and derived data. An analysis of the performance of machine learning shows that our method improved the accuracy of personal thermal comfort prediction by an average of 60%, and the accuracy of mean time to warm discomfort prediction by an average of 40%. The proposed thermal models were tested on subjects' data extracted from an office setup with room temperature varying from low (21.11 °C) to high (27.78 °C). When all proposed features were used, personal thermal comfort was predicted with an accuracy higher than 80% and mean time to warm discomfort with more than 85% accuracy. Further analysis of the machine learning efficacy showed that temperature differences had the highest impact on performance of individual thermal preference prediction, while the proposed approach was found not sensitive to the actual machine learning algorithm.

3. CONCLUSION

In this paper, we use the ARIMA model to process the high-dimensional data to predict the PM2.5 concentration in the 24 hours based on the historical datasets and predictive datasets. We proposed a predictive data feature exploration based air quality prediction approach. The approach enables to deeply mine and explore the high-dimensional time-related features and statistical features based on the exploratory analysis of big data. We utilize the sliding window mechanism of increasing the amount of training data to improve the training effect of the model and employed the air quality historical dataset of Beijing to evaluate the prediction model. The experimental results show that the approach outperforms the other baseline models. By incorporating the predictive data, the performance of the model could be improved under three evaluation indicators compared with the similar scheme using only the historical dataset. Meanwhile, the inclusion of statistical features in the prediction approach also has a good effect on improving the prediction performance. The approach proposed in this paper can effectively use the predictive data to deeply explore high-dimensional features, improve the model's ability to understand data, and is suitable for mining the features with strong correlation to the prediction objectives.

REFERENCES

[1] J. Huang et al., "A crowd source-based sensing system for monitoring fine grained air quality in urban environments," IEEE Internet Things J., to be published.

[2] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," Environ. Sci. Pollut. Res., vol. 23, no. 22, pp. 22408–22417, 2016.

[3] Q. Zhou, H. Jiang, J. Wang, and J. Zhou, "A hybrid model for PM2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network," Sci. Total Environ., vol. 496, pp. 264–274, Oct. 2014.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[5] A. C. Cosma and R. Simha, "Machine learning method for real-time noninvasive prediction of individual thermal preference in transient conditions," Building Environ., vol. 148, pp. 372–383, Jan. 2019.

[6] D. Zhu, C. Cai, T. Yang, and X. Zhou, "A machine learning approach for air quality prediction: Model regularization and optimization," Big Data Cogn. Comput., vol. 2, no. 1, p. 5, 2018.

[7] D. Wang, S. Wei, H. Luo, C. Yue, and O. Grunder, "A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine," Sci. Total Environ., vol. 580, pp. 719–733, Feb. 2017.

[8] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 3149–3157.

[9] W. Sun et al., "Intelligent in-vehicle air quality management: A smart mobility application dealing with air pollution in the traffic," in Proc. 23rd World Congr. Intell. Transp. Syst., Melbourne, Victoria, Australia, 2015, pp. 1–12.

[10] C. Ma et al., "Reducing air pollution exposure in a road trip," in Proc. 24th World Congr. Intell. Transp. Syst., Montreal, Canada, 2017, pp. 1–12.

[11] Y.Cheng, S.Zhang, C.Huan, M.O.Oladokun, and Z.Lin, "Optimization on fresh outdoor air ratio of air conditioning system with stratum ventilation for both targeted indoor air quality and maximal energy saving," Building Environ., vol. 147, pp. 11–22, Jan. 2019.

[12] W. Sun et al., "Moving object map analytics: A framework enabling contextual spatial-temporal analytics of Internet of Things applications," inProc. IEEEInt. Conf.ServiceOper .Logistics, Inform.(SOLI),Jul.2016, pp. 101–106.

[13] S. S. Roy, C. Pratyush, and C. Barna, "Predicting ozone layer concentration using multivariate adaptive regression splines, random forest and classification and regression tree," in Proc. Int. Workshop Soft Comput. Appl., 2016, pp. 140–152.
[14] J.C.Chang and S.R.Hanna, "Air quality model performance evaluation," Meteorol. Atmos. Phys., vol. 87, nos. 1–3, pp. 167–196, 2004.
[15] E. Meijering, ''A chronology of interpolation: From ancient astronomy to modern signal and image processing,'' Proc. IEEE, vol. 90, no. 3, pp. 319–342, Mar. 2002.
[16] S. Mahajan, H.-M. Liu, T.-C. Tsai, and L.-J. Chen, "Improving the accuracy and efficiency of PM2.5 forecast service using cluster-based hybrid neural network model," IEEE Access, vol. 6, pp. 19193–19204, 2018.
[17] Y.Zhengetal., ''Forecasting fine-grained air quality based on bigdata,' 'in Proc. ACMSIGKDDInt. Conf. Knowl. Discovery Data Mining.NewYork, NY, USA: ACM, 2015, pp. 2267–2276.
[18] C.ZhangandD.Yuan,''Fastfine-grainedairqualityindexlevelprediction using random forest algorithm on cluster computing of spark,'' in Proc. IEEE12thInt.Conf. Ubiquitous Intell. Comput. IEEE 12thInt.Conf.Autonomic Trusted Comput. IEEE 15th Int. Conf. Scalable Comput. Commun. Associated Workshops, Aug. 2015, pp. 929–934.
[19] M. Gao, L. Yin, and J. Ning, ''Artificial neural network model for ozone concentration estimation and Monte Carlo analysis,'' Atmos. Environ., vol. 184, pp. 129–139, Jul. 2018.
[20] Y.Zheng,F.Liu, and H.-P.Hsieh,''U-air:When urban airquality inference meetsbig data,''inProc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining. New York, NY, USA: ACM, 2013, pp. 1436–1444.
[21] H.-P. Hsieh, S.-D. Lin, and Y. Zheng, ''Inferring air quality for station location recommendation based on urban big data,'' in Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining. New York, NY, USA: ACM, 2015, pp. 437–446.
[22] Wang and G. Song, "A deep spatial-temporal ensemble model for air quality prediction,'' Neurocomputing, vol. 314, pp. 198–206, Nov. 2018.
[23] C. J. Huang and P.-H. Kuo, ''A deep CNN-LSTM model for particulate matter(PM2.5)forecastinginsmartcities,''Sensors,vol.18, no.7,p.2220, 2018.
[24] H. J. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Statist., vol. 29, no. 5, pp. 1189–1232, 2001.