Insights for Wellbeing: Predicting PM10 Values Using Stacking Ensemble Model

Huu-Vinh Nguyen[1, 2], Thi Thuy Nga Duong[2]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]University of Natural Resources and Environment, Ho Chi Minh City, Vietnam
nhvinh@hcmunre.edu.vn,dttnga_cntt@hcmunre.edu.vn

ABSTRACT

In this paper, we present our ISRS-HCMUNRE team's contribution to the task Insight for Wellbeing: Cross-Data Analytics for (transboundary) Haze Prediction at MediaEval 2021. We extracted different types of useful attributes for the problem: the weather data, the location features, the air pollution data on the data sets provided. We applied stacking method, deep learning models, machine learning model for prediction PM10 values at different locations for sub task 1.

1 INTRODUCTION

In many countries over the world, the prediction of air pollution is getting more attention. In this study, we aim to utilize deep learning and machine learning approach using insights from data provided by the organizer to predict the PM10 value, as given in the task 1 description of the competition MedialEval 2021 [3]. This task's primary motivation in to predict PM10 values for 3 days ahead. In this sub task, we explore the correlation between the PM10 value and the features we extracted from data set.

2 METHODOLOGY

2.1 Data Pre-processing

The dataset for task 1 includes weather data (temperature, humidity, wind speed…) and PM10 values for three countries: Thailand, Brunei and Singapore with many data points have zero value, unreasonably large or missing. They are called anomalies or outliers, which have to be preprocessed before extracting features. We calculated average values by day for each weather attributes if they collected by hour. We dealt with missing value by filling them with mean, zero [1].

We used Box plot method to determine the outliers. By using this method, we also found the max value and the min data range and moved outlier values into this range by setting them equals the max value or the min value.

2.2 Features Extraction

The number of weather attributes data provided is different for each country, they are important for predicting the PM10 values.

2.2.1 Timestamp features. For Brunei dataset, the weather attributes collected by day except wind speed values collected by hour. We calculated the average wind speed by day and put all weather attributes to train models.

MediaEval'21, 13-15 December 2021, Online

For Singapore dataset and Thailand, all the weather attributes collected by day. The PM10 values collected by hour. We calculated PM10 values average by day.

2.2.2 Location features. For station location data set, there is a weather station and an air quality station in a district. This is very helpful for prediction PM10 values by weather data in a district.

2.2.3 Weather features. The weather factors significantly affect the prediction of PM10 values. A show that the rain flow, wind speed affect the PM10 concentration in the air. We decided to use all weather attributes (e.g., temperature, humidity, wind speed, rain flow) for training the model.

3 MODELING

For training an appropriate model for the problem, we split raw dataset of each country into training and testing datasets following the time value. The training data contains 80 percent of data points and the testing dataset contains 20 percent.

For predicting PM10 values, we built two models called main-model and sub-model. We use weather data from three days before to predict three days ahead. The main-model is a stacked model [2] including three deep learning models: Long Short-Term Memory (LSTM)[4], Bi-directional (Bi-LSTM), Gate Recurrent Unit (GRU)[5] and a machine learning model: Linear Regression. The sub-model is a LSTM model that we used to predict PM10 values if there are missing weather data points in result data file. The data for training sub-model is only PM10 values from days before.

4 STACKING METHOD

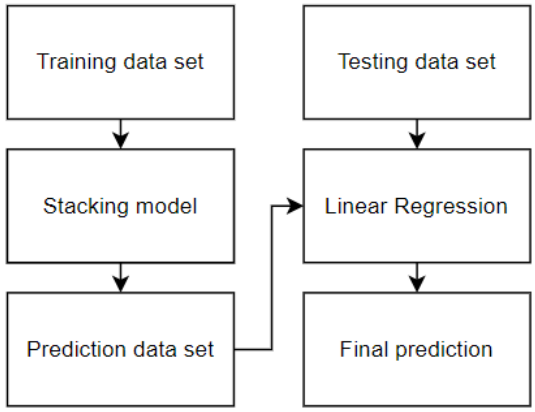

Figure 1: The illustration of stacking model

Table 1: Evaluation of the prediction PM10 values on Brunei training data set

	Single variable LSTM		Bi-LSTM		LSTM		GRU		Stacking	
Station ID	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
101B	10.8667	3.3	10.6673	10.812	10.9133	11.0315	10.8654	10.976	0.0425	0.21
201B	18.0297	4.25	20.6543	20.6661	21.1532	21.1606	20.6163	20.6284	0.0341	0.18
302B	31.77	5.64	23.4149	23.5046	23.7142	23.7927	23.3478	23.445	0.0362	0.19
401B	9.6541	4.11	13.9908	14.0457	14.4873	14.5318	14.4799	14.5192	0.0244	0.16

Table 2: Evaluation of the prediction PM10 values on Thailand training data set

	Single variable LSTM		Bi-LSTM		LSTM		GRU		Stacking	
Station ID	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
42T	7.8234	10.31	25.7271	25.745	25.3611	25.3749	26.0117	26.038	0.108	0.14
43T	7.43	10.19	8.784	8.8256	9.3066	9.3318	9.2756	9.3076	0.0988	0.13
44T	6.9928	8.98	21.4329	21.4416	21.2444	21.2527	21.436	21.4452	0.1161	0.15
62T	6.343	8.68	20.3824	20.4103	19.9103	19.9332	19.6263	19.6539	0.119	0.15
63T	6.4491	8.32	9.4611	9.5014	9.975	9.9999	9.9127	9.952	0.1095	0.15
80T	5.7	7.4	16.9081	17.0003	17.0171	17.0769	17.042	17.1219	0.154	0.2

Table 3: Evaluation of the prediction PM10 values on Singapore training data set

	Single variable LSTM		Bi-LSTM		LSTM		GRU		Stacking	
Station ID	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
1WS	5.1162	6.21	26.7096	26.778	26.8381	26.8885	26.6675	26.7233	0.1537	0.19
2ES	5.9605	7.63	24.1252	24.1544	24.2381	24.2608	24.1295	24.1519	0.2189	0.26
3CS	4.7474	5.81	21.3828	21.4401	21.612	21.6594	21.4659	21.5168	0.17	0.21
4SS	5.1535	6.34	29.3604	29.4337	29.4885	29.5365	29.637	29.7021	0.13	0.16
5NS	5.2575	6.68	30.2583	30.3607	30.6443	30.7297	30.5918	30.6796	0.179	0.22

To enhance the prediction result, we employed the stacked generalization technique. The stacked model has two levels: level 0 and level 1. The level 0 data is the training dataset inputs and level 0 models learn to make predictions from this data. The level 1 input data is the output of level 0 models and the single level 1 model, or meta-learner to make predictions form this data.
We utilized three models: LSTM, Bi-LSTM, GRU, as level 0 models. We used Linear Regression model as level 1 model for final prediction.

5 PERFORMANCE METRICS

For evaluating the performance on the proposed methods, we use root mean square error (RMSE), mean absolute error (MAE), as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$MAE = (\frac{1}{N})\sum_{i=1}^{N}|y_i - \hat{y}|$$

Where $\hat{y}$ is the i^{th} predicted value from model, $\bar{y}$ is the average of observed values, and y_i is i^{th} observed value, (i = 1, ..., N).

6 RESULTS AND DISCUSSION

After extracting the necessary information, we evaluated four deep learning models: LSTM, Bi-LSTM, GRU, single variable LSTM and stacking model on testing data set. The table 1,2,3 show the RMSE, MAE score for each air quality station.

There are a lot of data points missing in Thailand and Singapore data sets leads to the testing result in poor performance.

REFERENCES

[1] Denis Cousineau and Sylvain Chartier. 2010. Outlier detection and treatment: a review. *International Journal of Psychological Research* (2010), 58–67.

[2] Dat Q. Duong, Quang M. Le, Tan-Loc Nguyen-Tai, Hien D. Nguyen, Minh-Son Dao, and Binh T. Nguyen. 2021. An Effective AQI Estimation Using Sensor Data and Stacking Mechanism. In *New Trends in Intelligent Software Methodologies, Tools and Techniques - Proceedings of the 20th International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques, SoMeT 202, Cancun, Mexico, 21-23 September, 2021 (Frontiers in Artificial Intelligence and Applications)*, Hamido Fujita and Héctor Pérez-Meana (Eds.), Vol. 337. IOS Press, 405–418. https://doi.org/10.3233/FAIA210040

[3] Asem Kasem, Minh-Son Dao, Effa Nabilla Aziz, Duc-Tien Dang-Nguyen, Cathal Gurrin , Minh-Triet Tran, Thanh-Binh Nguyen, and Wida Suhaili. Overview of MediaEval 2021: Insights for Wellbeing Task Cross-Data Analytics for Transboundary Haze Prediction.

[4] Jung-Hwan Park, Seong-Joon Yoo, Kyung-Joong Kim, Yeong-Hyeon Gu, Keon-Hoon Lee, and U-Hyon Son. 2017. PM10 density forecast model using long short term memory. In *2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*. 576–581. https://doi.org/10.1109/ICUFN.2017.7993855

[5] Guang Yang, HwaMin Lee, and Giyeol Lee. 2020. A Hybrid Deep Learning Model to Forecast Particulate Matter Concentration Levels in Seoul, South Korea. *Atmosphere* 11, 4 (2020). https://doi.org/10.3390/atmos11040348