An Air Quality Grade Forecasting Approach Based on Ensemble Learning

Weike Liu
Science and Technology on Information Systems Engineering Laboratory
National University of Defense Technology
Changsha China
liuwk15@hotmail.com

Hang Zhang
Science and Technology on Information Systems Engineering Laboratory
National University of Defense Technology
Changsha China
zhanghang13@nudt.edu.cn

Qingbao Liu
Science and Technology on Information Systems Engineering Laboratory
National University of Defense Technology
Changsha China
liuqingbao@nudt.edu.cn

***Abstract*—This paper proposes an air quality grade forecasting method based on ensemble learning. First, the training data sets are formed of the air quality data and related meteorological data crawled from air quality data website. After that, use the ensemble learning algorithm Leveraging Bagging to learn the training dataset and generate initial air quality grade forecasting model. And the initial forecasting model is used to make prediction on the prediction dataset. In total, the experiments test the learning algorithm both on the city scale and the station scale. Experimental results show that the proposed method has good prediction effect and good forecasting ability on the real forecast dataset.**

Keywords — ensemble learning, air quality grade forecasting, MOA

I. INTRODUCTION

Traditional air pollution modeling includes Gaussian models of different complexity, Lagrange models, chemical transport models and so on [1]. In spite of making use of the technological development of atmospheric science and computer science, these models are severely dependent on real-time updated meteorological data and a detailed list of emission sources. It imposes significant limitations on the use of these models. Therefore, some statistical models based on machine learning algorithms are used to predict air quality grade. Bougoudis et al. [2] aimed at finding the conditions of high-pollution and used a more generalized hybrid model which is based on unsupervised clustering. The hybrid model integrates artificial neural networks (ANN), random forests (RF) and fuzzy logic to predict the multi-index pollutants in Athens. Zhao et al. [3] proposed a Deep Recurrent Neural Network (DRNN) to predict daily air quality grade. Huang and Guo [4] also used the hybrid model of neural network and Long Short-Term Memory (LSTM) to predict the concentration of PM2.5. Liu et al. [5] obtained a reliable air quality prediction model using Support Vector Machines (SVM) by using monitoring data from three cities of Beijing, Tianjin and Shijiazhuang. Vong et al. [6] also used SVM to predict air quality (NO2, SO2, O3, SPM) from pollutants and meteorological data in Macao and China.

This paper proposes an air quality grade forecasting method based on ensemble learning algorithm. First, a web crawler is used to collect public air quality data from website. Then, the data is preprocessed to generate training datasets and prediction datasets. The ensemble learning algorithm Leveraging Bagging [7] is used to learn the training data set to get the forecasting model, which makes predictions on the prediction dataset. Experiments also compare the accuracy of the prediction models generated by Leveraging Bagging with other learning algorithms. The rest of this paper is organized as follows: Section 2 describes the impact factors for urban air quality prediction, followed by an introduction to data acquisition and data preprocessing. Section 3 introduces the ensemble learning methods and the process of the air quality grade forecasting method. Section 4 shows the experimental results. And the conclusion is shown in Section 5.

II. AIR QUALITY IMPACT FACTORS AND DATA ACQUISITION

A. Air quality impact factors

Most of the existing researches on air quality impact factors are based on the air quality data of a city or a province. Their findings indicate that the relationship between air quality and meteorological factors in different regions of the same country are usually different. This relationship is determined by the regional economic development, which influences the amount of fuel consumption, contaminant types, contaminant emission amount, pollution treatment technology, urban green area and other factors. These factors affect the adsorption and degradation of atmospheric pollutants within the city. In addition, regional topography also affects the diffusion conditions of atmospheric pollutants through atmospheric horizontal and vertical exchanges. Therefore, the level of economic development of the city, the pollution emissions of major industrial facilities, as well as the topography and climate of the region where the city is located will have an impact on local air quality impact factors.

Air pollution Index (API) combines several air contaminant concentrations and form a single index, which measures the air quality and is suitable for describing the short-term city air quality status and trends. Air pollutants include: soot, total suspended particulate matter, respirable Particulate Matter (PM10), Sulfur Dioxide (SO2), Nitrogen Dioxide (NO2), Ozone (O3), and Carbon monoxide (CO), volatile organic compounds, and so on. However, with the rapid development of economy and society, API can no longer meet the requirements of current

air pollution monitoring. For example, fine Particulate Matter (PM2.5), the main pollutant that frequently appears now, is not included in the API index. Therefore, since 2012, AQI has gradually replaced the past API to be the current air quality assessment indicators. Among the AQI indicators, the main pollutants involved in air quality assessment are PM2.5, PM10, SO2, NO2, O3, and CO. According to AQI, the urban air quality can be divided into 6 grades which is shown Table 1.

TABLE I. AQ GRADE

AQ Index	AQ Grade	AQ Description	Color
0~50	0	Excellent	Green
51~100	1	Good	Yellow
101~150	2	Mild pollution	Orange
151~200	3	Moderate pollution	Red
201~300	4	Severe pollution	Purple
>300	5	Serious pollution	Maroon

B. Data Acquisition and Preprocessing

The datasets used in experiments consists of time information, weather conditions, future air quality forecast results, and air quality at the city and monitoring sites. These data are from the public weather forecast websites PM2.5.in and PM.kksk.org. All data is automatically obtained by a web crawler. The main attributes of the datasets are shown in Table 2:

TABLE II. ATTRIBUTES OF DATASETS

Data Types	Data attributes
Time	year, month, day
Weather	Temperature, precipitation, wind direction, wind speed, relative humidity, comfort, body temperature
AQI	AQI value, AQ grade, Major pollutant, PM2.5, PM10, CO, NO2, O3_1, O3_8, So2

The preprocessing of the dataset mainly includes the following work: First is filling the missing items in the datasets. Then, for the anomaly data in the datasets, we used the adjacent normal data for overlay padding. Finally, the form of the datasets is transformed to make sure that the learning algorithm can address them.

III. AIR QUALITY GRADE FORECASTING BASED ON ENSEMBLE LEARNING

A. LeveragingBagging Ensemble Learning Algorithm

Online ensemble learning is an incremental learning method, which addresses the arriving instances one by one and gradually evolves the learning model. Oza and Russell [8] modified the traditional bagging to online learning condition and proposed Online Bagging (OB). OB contains an ensemble E that has M base model h_m (m=1, …, M). When a new instance arrives, the times that each base model trained for the instances are determined by *Poisson* (1). Therefore, the diversity of the base models is generated by different training times. At last, all the base models make joint prediction by voting.

Bifet et al. [7] improved OB and proposed online Leveraging Bagging (LB) algorithm. They leveraged the performance of bagging with two randomizations improvements: increasing resampling and using output detection codes. First, LB increases the weights of this resampling using a larger value λ to compute the value of the Poisson distribution. Using a value λ>1 will increase the diversity of the weights and modifying the input space of the classifiers inside the ensemble. Second, LB adds randomization at the output of the ensemble using output codes. In standard ensemble methods, all classifiers try to predict the same function. However, using output codes each classifier will predict a different function. This may reduce the effects of correlations between the classifiers, and increase diversity of the ensemble.

B. Forecasting Process

The downscaling air quality grade forecasting includes two stages. First, is the model initialization stage. Second, is the online forecasting and incremental learning stage.

In the model initialization stage, the ensemble learning algorithm will use the training samples to generate forecasting model. A training instance is the combination of the current urban air quality data with the measured values of the next hour urban air quality grade. The ensemble learning algorithm LB is used to train an initial forecasting model using the training datasets. Then, the performance of the initial forecasting model is evaluated by the test-then-train method and the results are also reported. Finally the model will be used as the initial forecasting model in next stage.

In the online forecasting and incremental learning stage, the web crawler keeps collecting data from the website and form the training instances. At T_0, web crawler gets the urban air quality data at T_0 and the initial forecasting model will use the data to predict the urban air quality grade of T_1. At T_1, web crawler gets the urban real air quality grade and the urban air quality data of T_1. And the urban air quality data at T_0 and the urban real air quality grade of T_1 are combined to generate a training instance. And the forecasting model will incrementally learn this training instance. Then, the updated forecasting model will predict the air quality grade of T_2 based on the urban air quality data of T_1. By analogy, the online prediction task and the incremental learning process of the model are completed. Algorithm 1 shows the process of the online forecasting and incremental learning algorithm.

Algorithm1: *OnlineForecastingAndIncrementalLearning*

Input:
1. *Model*: the initial forecasting model
2. *Stream*: the air quality dataset
3. *Resfile*: output file of prediction result

Output: *Model*: the updated forecasting model

Process:
1. **while** (*Stream*.hasNext()) **do**
2. *testInst* = *Stream*. nextInstance()
3. **if** *testInst.*classIsMissing() **then**
4. *prediction* = *Model.*getPrediction(*testInst*)
5. *Resfile.*println(*prediction*)
6. **else**
7. *Model.*TrainOnInstance(*testInst*)
8. **end if**

9. **end while**
10. **Return** *Model*

When the air quality dataset still has instances, the model will keeps get instances from the dataset. If an instance has the urban air quality grade, the model will incrementally train on it. Otherwise, the model will make prediction on it and the results will be stored into a file to make analysis. Once all the instances in the dataset are addressed, the updated model will be returned.

C. *Downscaling Forecasting*

By change the scales of the urban air quality grade in the training dataset during the initial model training stage, the model can achieve the downscaling prediction of the urban air quality grade in the time domain and space domain. If the downscaling is in the time domain, the air quality grade can be released from daily to hourly level. When the city scale is changed from urban scale to station scale, the air quality grade forecasting will get a regional downscaling.

IV. EXPERIMENTS

A. *Settings and Datasets*

In the comparison experiments, the ensemble learning algorithm Leveraging Bagging is applied. The base classifier of the ensemble is the Hoeffding Tree with default parameters. The ensemble classifier consists of 10 base classifiers. All the experiments are carried out based on the Massive Online Analysis (MOA) [9], which is designed for online data stream learning. And all the experiments are carried out on the machine with i7-6700HQ 2.60GHz, 8 GB RAM and Windows 10.

As for the datasets, the urban air quality data in Beijing and Changsha are used. The training datasets contains 3165 hours from 20:00 on November 1, 2018, to 21:00 on March 27, 2019. The predicting dataset contains 136 hours from 22:00 on March 27, 2019 to 14:00 on April 2, 2019. Training dataset of Beijing have 3029 instances and the training dataset of Changsha has 3030 instances. Both the predicting dataset of Beijing and Changsha have 136 instances.

B. *Comparison of Initial Forecasting Model*

In this section, four representative online learning methods are compared, which include LB, OB, HoeffdingTree (HT), and Naïve Bayes (NB) algorithm. Both the OB and LB use 10 HT as base models. All the base classifiers are initialized by 50 instances. All the comparison experiments are carried out 10 times and calculate the average results. At the same time, using two different single classifier algorithms HT and NB to verify that of the ensemble algorithms have a greater advantage than the single model algorithms. By comparing the prediction results of the two base classifier algorithms on the same dataset, the optimal base classifier algorithm is discriminated. As the algorithms learns the datasets in a test-then-train manner, the accuracy of the algorithms can be reported.

First, the compared algorithms are conducted on the datasets of the city scale, Beijing and Changsha. Figure 1 shows the accuracy curves of Beijing and Changsha. The accuracy of LB algorithm is much higher than the other three algorithms. In the comparison of the single classifier algorithm, the effect of the HT is better than the NB, which shows that the LB using the HT as base classifier is optimal. At the same time, the accuracy of the training model has reached a high level when the learning procedure ends. The accuracy of the models in both cities is close to 90%. This shows that the training model can reach the stage of online prediction.

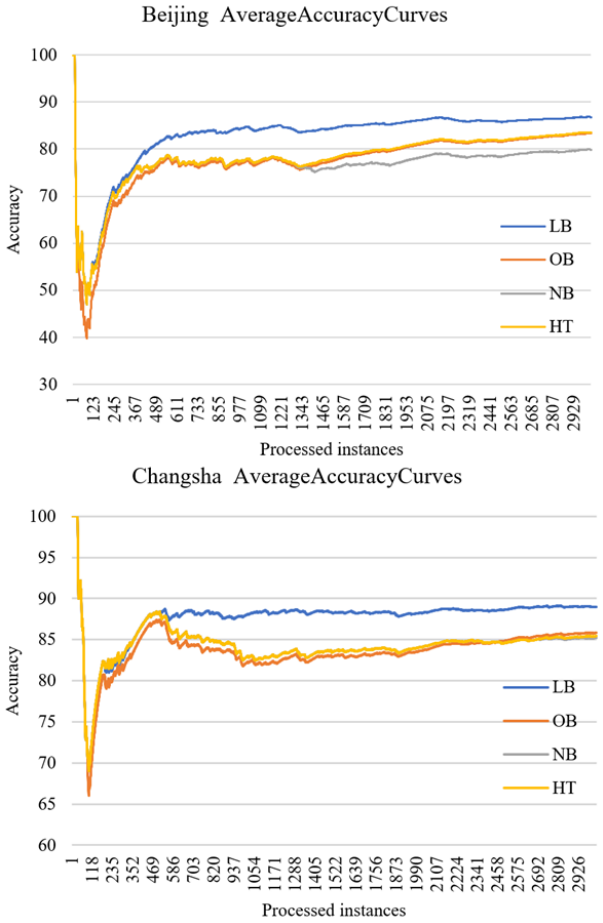

Fig. 1. accuracy curves of Beijing and Changsha

Second, the compared algorithms are conducted on the datasets of the station scale, Beijing Olympic Center and Changsha Southern Train Station which is shown in Figure 2. In sum, LB still achieves a better performance than other methods.

In the comparison of initial forecasting model on city scale, all the algorithms obtain good forecasting performance. In the training data set of Changsha City, the average prediction accuracy of the four algorithms reach more than 80%, and the trend accuracy is continuously improved with the increase of training instances. The LB algorithm performs best, and its prediction accuracy for the last set of instances in the training set reaches 86%. The result is higher than 85% and meets the air quality model prediction accuracy requirements. In the comparison on the station scale, the forecast accuracy of all algorithms is generally lower than the forecast accuracy of the city scale. From the trend of the prediction accuracy of the training model, the accuracy increases with the increase of the instance, so when the training data set continues to expand, the initial training model with better performance may be obtained.

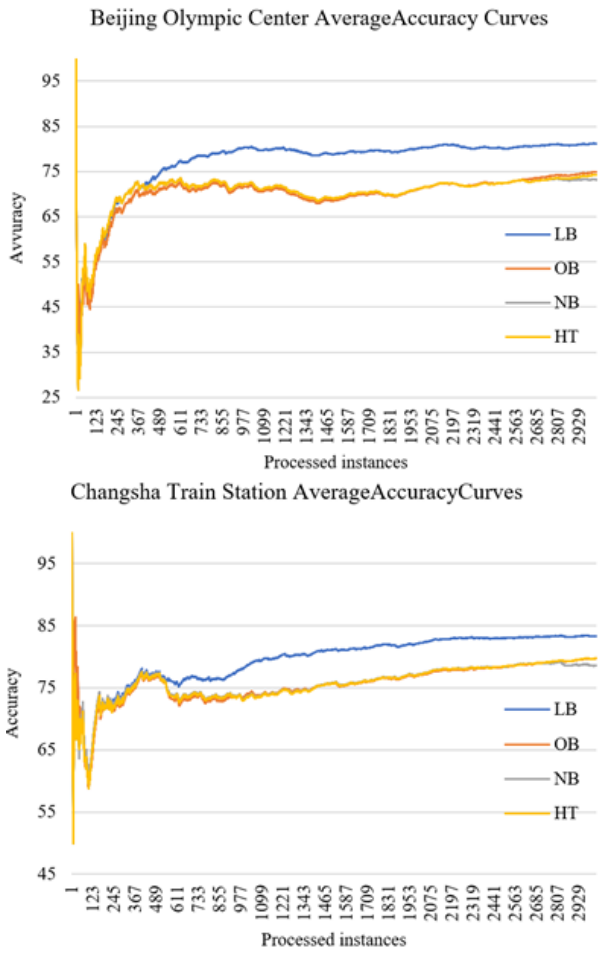

Fig. 2. accuracy curves of Beijing Olympic Center and Changsha Southern Train Station

C. *ForecastingResults*

In this section, the performance of the initial forecasting models is tested on the predicting datasets. The initial forecasting models first make predictions on a instances without the real air quality grade. Then, the real air quality grade is used to test the predictive accuracy of the ensemble algorithms. Table 3 shows the predicting air quality grade and the real grade on April 1st. And the comparison results in the table is presented using the predictive/real format. For example, the value 0/1 means that the predictive air quality grade is 0 and the real grade is 1. If the predictive value is different from the real value, the cell will be filled with yellow.

TABLE III. PREDICTIVE AND REAL AQ GRADE ON APRIL 1ST

	Beijing	**Beijing Olympic Centre**	**Changsha**	**Changsha Southern Train Station**
0	0/0	0/0	1/1	1/1
1	0/0	0/0	1/1	1/1
2	0/0	0/0	1/1	1/1
3	0/0	0/0	1/1	1/1
4	0/0	0/0	1/1	1/1
5	0/0	0/0	1/1	1/1
6	0/1	0/0	1/1	1/3
7	1/0	0/0	1/1	1/1
8	0/0	0/0	1/1	1/1
9	0/0	0/0	1/1	1/1
10	0/0	0/0	1/1	1/1
11	0/0	0/0	1/1	1/1
12	0/0	0/0	1/0	1/1
13	0/0	0/0	1/0	1/1
14	0/0	0/0	0/0	1/1
15	0/0	0/0	0/0	1/0
16	0/0	0/0	0/1	0/0
17	0/0	0/0	1/1	0/1
18	0/0	0/0	1/1	0/1
19	0/1	0/1	1/1	1/1
20	0/1	1/1	1/1	1/1
21	1/1	1/1	1/1	1/1
22	1/1	1/1	1/1	1/1
23	1/1	1/1	1/1	1/1
Num. error	4	1	3	4
Accuracy (%)	83.33	95.83	87.5	83.33

As it shown in Table 3, the predictive accuracy is best at the Beijing Olympic center. And most of the wrong prediction happens from 12:00 to 20:00. In general, the inversion layer is prone to occur at night, in the morning, and in the evening. During these times, various pollutants in the air are not easily diffused. Therefore, the air quality grade is likely to rise during these periods. After the sun came out, the ground temperature rise rapidly and the inversion layer begin to dissipate. Therefore, after 10 am, the air quality will generally turn better. Theoretical analysis is also confirmed in the results. First, air quality grades tend to rise between 7 and 8 in the morning. Similarly, during the period from 16:00 to 8:00, the air quality grade tends to rise or fluctuate at each monitoring station in each city. In the above table, it can be found that after 12 o'clock, the average air quality grade of the city of Changsha has dropped from 1 to 0, and it did not rise until 16 pm.

At the same time, by analyzing the causes of the errors, we can find:

- The forecasting models respond slowly to sudden changes in air quality grades. This is mainly reflected in the comparison of the predicted results at 6 o'clock in the morning at Changsha Southern Train Station.
- Forecasting models tend to produce hysteresis during periods when air quality grades are prone to change. This shows that the forecasting model needs to learn incrementally.
- Although the model is not sensitive enough to predict the sudden change of air quality grade, from the 24-hour continuous forecast, the prediction accuracy of the method is more than 83%. It shows that this method still has great application value in real life.

Next, the results on the whole prediction dataset are presented (From 22:00 on March 27, 2019 to 14:00 on April 2, 2019.). Figure 3 and 4 shows the predictive AQ grade and real AQ grade on the four whole prediction datasets. As for Beijing and Changsha, the prediction AQ grade is consistent with the overall trend of the real grade. In general, there is a problem that the forecast value lags behind the true value. It can be found from the figures that in Beijing and Changsha, the forecast model is sensitive to the situation that the air quality begins to turn better.

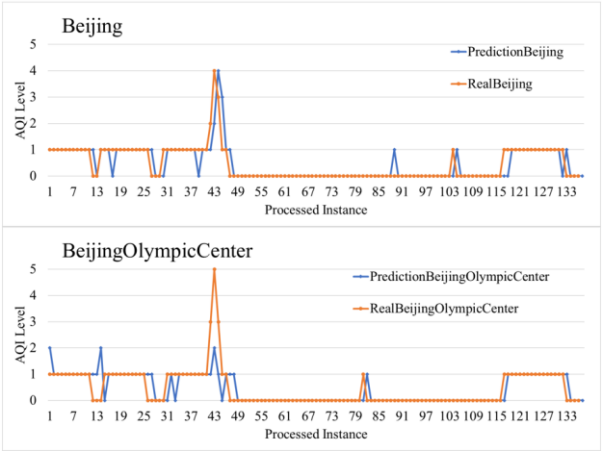

Fig. 3. Forecasting results of LB prediction model of Beijing

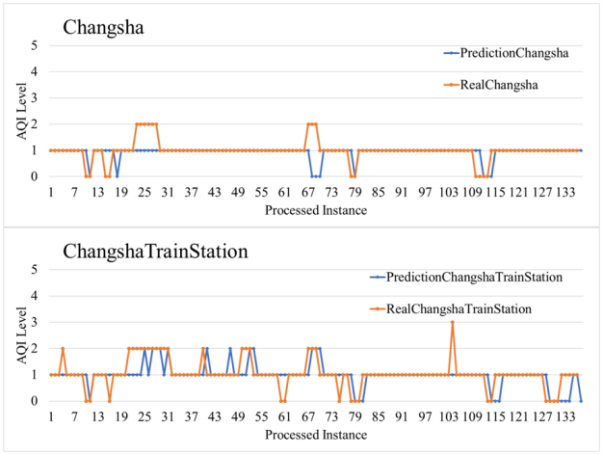

Fig. 4. Forecasting results of LB prediction model of Changsha

Since the air quality changes are relatively frequent on a small scale, it can be clearly seen that the prediction accuracy of the station scale is lower than the city scale. From the perspective of the overall trend of AQ grades, the forecast results are basically consistent with the real situation. When the AQ grades of some stations changes, the prediction model does not necessarily predict the moment when the change occurs. From 17:00 to 19:00 pm on March 29, the actual AQ grade of Beijing Olympic Center increased from grade 1 to 4 and then back to 1, while the model only predicted that the peak value of the change was 2. Similar to the prediction of the city scale, the prediction of the forecast model on the station scale is also more sensitive to the change of the grade.

From the above results, it can be found that the prediction model has a good effect on both the city scale and station scale air quality grade prediction. The accuracy of Changsha are higher than Beijing. This may be because the urban area of Changsha City is smaller than that of Beijing, and the stations are densely distributed. The correlation factors between the stations in the city are strong, so the prediction accuracy is high. To improve the prediction accuracy, the air quality monitoring station can be added in the actual situation, or the air quality monitoring data can be collected for a longer time as a training set to train the model.

V. Conclusion

This paper proposes a new forecasting method for air quality grade prediction. The main contribution is the use of the ensemble learning algorithm to predict air quality grade. We collect real-world data sets through web crawlers and train the models. Experimental results show that the proposed method has good prediction effect and good forecasting ability on the real forecast dataset.

References

[1] Grange, S.K.; Carslaw, D.C.; Lewis, A.; Boleti, E.; Heuglin, C. Random forest meteorological normalisation models for swiss pm10 trend analysis. *Atmospheric Chemistry and Physics Discussions* **2018**.

[2] Bougoudis, I.; Demertzis, K.; Iliadis, L.S. Hisycol a hybrid computational intelligence system for combined machine learning: The case of air pollution modeling in athens. *Neural Computing and Applications* **2016**, *27*, 1191-1206.

[3] Zhao, X.; Zhang, R.; Wu, J.L.; Chang, P.C. *A deep recurrent neural network for air quality classification*. 2018; Vol. 9, p 346-354.

[4] Huang, C.-J.; Kuo, P.-H. A deep cnn-lstm model for particulate matter (pm2.5) forecasting in smart cities. 2018; Vol. 18, p 2220.

[5] Liu, B.C.; Binaykia, A.; Chang, P.C.; Tiwari, M.K.; Tsao, C.C. Urban air quality forecasting based on multi-dimensional collaborative support vector regression (svr): A case study of beijing-tianjin-shijiazhuang. *Plos One* **2017**, *12*, e0179763-.

[6] Vong, C.-M.; Ip, W.-F.; Wong, P.-K.; Yang, J.-y. Short-term prediction of air pollution in macau using support vector machines. 2012; Vol. 2012.

[7] Bifet, A.; Holmes, G.; Pfahringer, B. In *Leveraging bagging for evolving data streams*, European Conference on Machine Learning & Knowledge Discovery in Databases, 2010.

[8] Oza, N.C. In *Online bagging and boosting*, 2005 IEEE International Conference on Systems, Man and Cybernetics, 12-12 Oct. 2005, 2005; pp. 2340-2345 Vol. 2343.

[9] Bifet, A.; Holmes, G.; Kirkby, R.; Pfahringer, B. Moa: Massive online analysis. *J. Mach. Learn. Res.* **2010**, *11*, 1601-1604.