Modelling the Atmospheric Concentration of Carbon Monoxide by Using Ensemble Learning Algorithms

Adven Masih
Ural Federal University
Ekaterinburg, Russia
adven.masikh@urfu.ru

Abstract

Air quality monitoring is among several important tasks performed in environmental science and engineering. Photochemical reaction in troposphere is the major natural source of carbon monoxide production. Other significant portion of carbon monoxide in air is contributed due to anthropogenic activities such as vehicular emissions. The concentration of Carbon monoxide in atmosphere plays a vital role in the formation of ground level ozone which is highly harmful for human health, therefore a constant monitoring of carbon monoxide is essential. Similarly, development of air quality monitoring models is also vital because such models can efficiently provide warnings ahead of time when air pollution reaches to an unsatisfactory level. The study makes use of most advanced and widely popular machine learning techniques such as ensemble learning algorithms, artificial neural network (ANN) and support vector machine (SVM). The prediction of atmospheric carbon monoxide for this study is based on 5 atmospheric gases SO2, NO, No2, NOx, ozone directly associated with vehicular emissions, and 3 meteorological parameters temperature, wind speed and wind direction, which aid CO for photochemical reaction and transport it from one region to another. The literature review conducted for this paper revealed that, so far, a limited number of machine learning algorithms have been employed for modeling atmospheric gases e.g. carbon monoxide, nitrogen dioxide etc. On the other hand, recently, with the introduction of ensemble learning techniques, and deep neural networks, machine learning technology has become significantly advance. Given these observation, for this study a comparison was drawn between state-of-art prediction models and ensemble learning algorithms which indicates that ensemble classifiers such as Random Forest and Bagging perform better than ANN and SVM. It also discusses the effective use of ensemble learning algorithms to develop models that can efficiently predict the concentration of carbon monoxide in atmosphere.

1 Introduction

In recent years environmental risks caused by the rising level of carbon monoxide (CO) concentration in atmosphere due to stationary and mobile sources have significantly increased. CO is a colorless, odorless, tasteless and toxic air pollutant that forms due to the incomplete combustion of carbon-containing fuels such as oil, natural gas, gasoline, coal and wood. The photochemical reaction in troposphere and exhaust emissions from vehicles are the major sources of carbon monoxide production in atmosphere. The possible sources of CO gas at home include the hydrocarbon fuel appliances such as gas

fires, water heaters, cookers, central heating system etc. and open fire that uses oil, gas, wood and coal. The areas with high concentration of CO in atmosphere can endanger the local receptors. Several topographical and meteorological effects on the formation and transport of CO and the significant relationship between high concentration of CO and environmental risks have been listed in the work of [1].

The technique adopted for the second set of experiments is Bootstrap Aggregation or Bagging. It is an ensemble learning approach proposed by Breiman in 1996. It works on a principle of randomly resampling the original data with a replacement by using bootstrap method. It produces a dataset which is different from each other but with an equal sample size, prior to build a tree from each sample. Subsequently, a classification model from each sample is developed, results of such models are further combined to form a prediction model. Bagging uses weighted vote for classification problems whereas for a regression task it uses average vote. The processes discussed above carried out for bagging known for the fact that it resolve classifiers' most common problem of data over fitting.

The existing approaches of modelling CO concentrations to predict air pollution, in major, have employed traditional machine learning algorithms e.g. Artificial Neural Networks and Support Vector Machines. Although advance techniques – ensemble classifiers in data mining have successfully been applied in several fields such as bioinformatics, marketing and medicine [2-6], however, in environmental science, only few attempts [7, 8] have been made that employed ensemble learning algorithms as classifiers to predict the concentration of atmospheric pollutants. The literature review conducted in the context of this work revealed that the ensemble learning approaches, when used as predictive models, improve the accuracy of the model in comparison with the single base learning techniques such as ANN and SVM. There are limited studies making use of ensemble learning algorithms as classifiers, with no comparison trend drawn between the different techniques used for investigation. Therefore, the work aims at finding the most accurate machine learning models and algorithms to predict atmospheric CO, by using the concentrations of atmospheric gases and meteorological parameters.

2 Material and Methods

The dataset used for experiment contains meteorological and atmospheric gas concentrations data. The dataset were obtained from the official website of Department of Environment Food & Rural Affairs. It was recorded during January 1st, 2013 to 18th June, 2013 at a sampling rate of one hour near Marylebone road located in London, United Kingdom. A spatial prediction approach is adopted i.e. the time at which the concentrations were recorded is not considered, i.e. for modelling only meteorological parameters such as wind speed, wind direction, and temperature, along with the concentrations of other atmospheric gases e.g. No2, SO2, NO, NOx, carbon monoxide and ozone were considered. The analysis presented involves three main stages i.e. (1) data collection, (2) data preprocessing and (3) modelling as shown in figure 1. During data preprocessing several steps performed to clean dataset include raw data collection, removal of missing values and outliers, data transformation, and feature selection.

To author's best knowledge ensemble learning approach have not been applied in a comprehensive investigation for the prediction of atmospheric CO. Therefore, a thorough investigation comparing the modelling performance of machines learning algorithms have been carried out. Altogether a total of 11 predictive models were developed using both single based learning algorithms and meta-learning ensemble techniques by means of a well-known toolkit called WEKA (Waikato Environment for Knowledge Analysis) for the prediction of CO concentration. Furthermore, a comparative analysis was performed to figure out the algorithm that produces the best results.

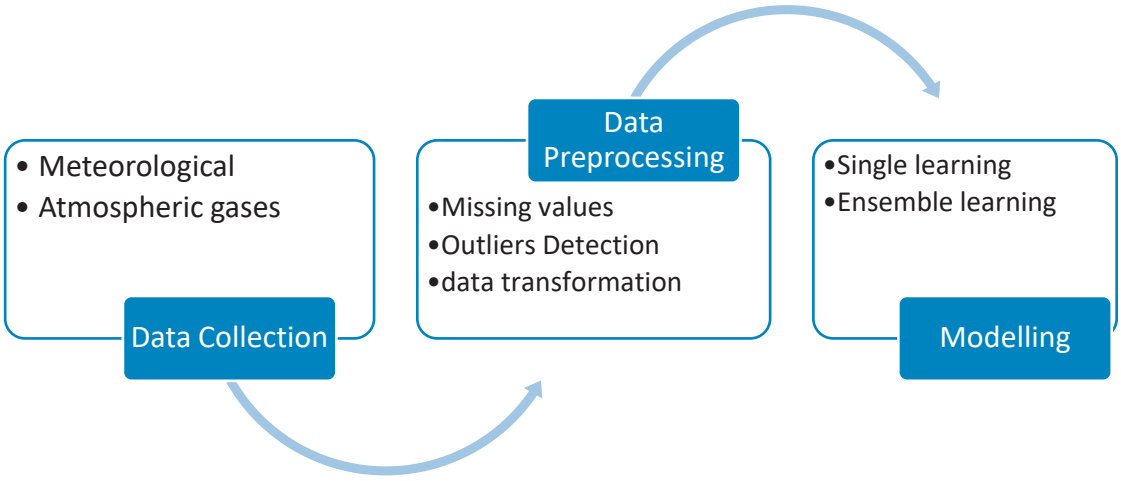

Figure 1: data processing scheme

To draw a comparison among state-of-art classification techniques such as meta learning (Additive Regression, Bagging, Random Subspace), Artificial Neural Network (Multilayer perceptron, Support Vector Machine), Lazy (KStar and IBk), Rules (M5Rules) and Tree classifiers (Random Forest, M5P, REPTree, and Random Tree) i.e. all possible classifiers available in WEKA classifier categories including Functions, Lazy, Meta, Rule and Tree were tried. The list of classification algorithms selected for detail analysis is presented in table-1. To evaluate individual models for testing and performance purposes, experimental design adopts the ten-fold cross validation for the implementation of all 11 machine learning algorithms. And to evaluate the accuracy of models four widely accepted evaluation measurements used are; Correlation Coefficient (CC), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Relative Absolute Error (RAE).

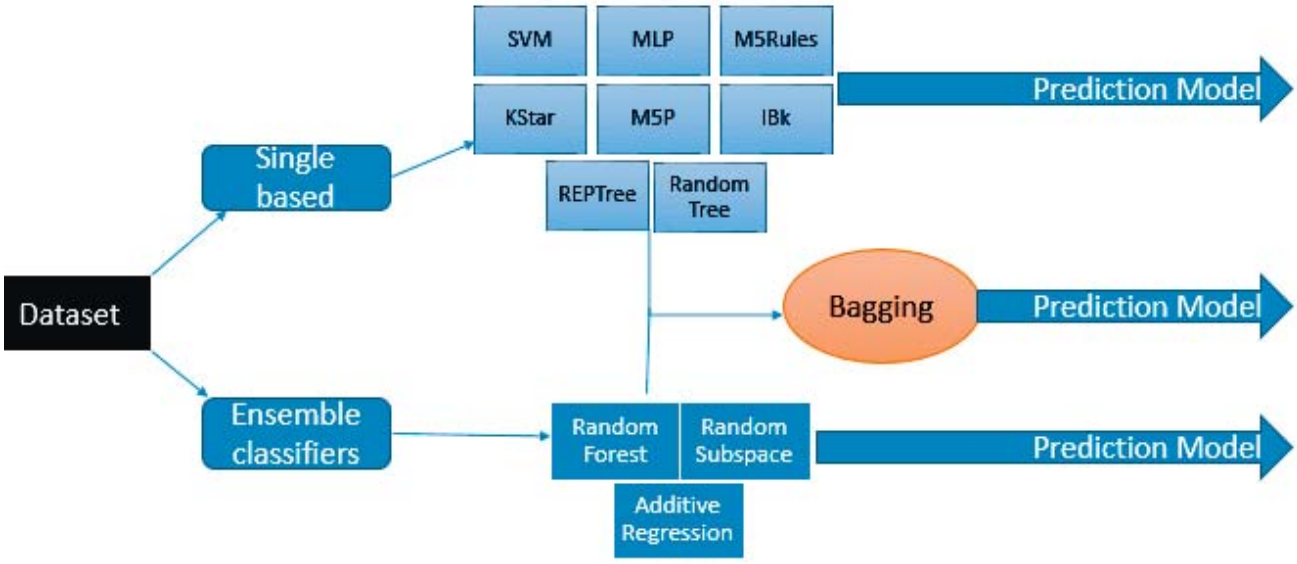

Figure 2: Experimental design

A careful analysis revealed some missing values and outliers, for which outlier removal and missing values filters were applied during data preparation. As the wind direction data is recorded in degrees and it ranges between 0-360, therefore to make sure that 0 and 360 are considered the same, wind data transformation was performed. For that the Wind Speed (WS) and Wind Direction (WD) were combined in the form of two new orthogonal components i.e. U=WS*cos(WD) and V= WS*sin(WD) to replace WS and WD. After data preparation, the original dataset of 4000 instances were used for modelling phase.

It is a said that the best results obtained from different classifiers largely depend on the type of dataset used. However experiments conducted during the first phase show that homogenous ensemble learner' i.e. Random Forest in particular outperformed single learning algorithms for the prediction of atmospheric carbon monoxide. Whereas in later phase the performance of single base learning algorithms and ensemble classifiers within Bagging were compared with the results obtained during the first phase.

3 Results and discussion

The dataset presented is a sequence of six atmospheric gases No2, SO2, NO, NOx, carbon monoxide, ozone and three meteorological parameters temperature, wind speed and direction recorded in a time series. The descriptive statistics of all 9 attributes used for data analysis is listed below in table-1.

Table-1 descriptive statistics of dataset attributes

Attribute	Unit	Min	Max	Mean	Standard Dev.
So2	μgm^3	-0.475	33.62	5.87	5.34
Ozone	μgm^3	0.35	82.95	22.3	18.63
NO	μgm^3	1.27	636.78	106.04	99.54
NO2	μgm^3	7.25	206.2	79.57	38.42
NOx	μgm^3	9.63	1153.37	242.17	187.64
CO	μgm^3	47.44	2012.6	457.15	235.65
Temperature	°C	-9.2	28.1	7.02	6.22
Wind Speed	ms^{-1}	0.1	10.1	3.83	1.79
Wind Direction	Degree	0.1	360	169.68	107.93

The results of experiments tabulated in table 2 (a) were aimed at comparing the performance of various single based algorithms against the three ensemble learning techniques i.e. Additive Regression, REPTree and Random Forest (RF) comprehensively. Table 2(a) clearly depicts that using homogenous ensemble learning approaches can significantly improve the prediction accuracy of atmospheric CO concentration. Whereas in table 2(b), the ability of ensemble classifier – Bagging when used as a classifier to reduce MAE, RMSE and RAE and improve the prediction accuracy can clearly be seen. Table 2 is evident that ensemble classifiers in terms of prediction accuracy can outperform the widely used single based learning algorithms – Artificial Neural networks (MLP) and Support Machine Vector (SVM).

Although the performance of Additive Regression on its own is worse when compared directly with other two ensemble techniques such as Random Forest and Random Subspace in table 2(a), however, it is worth noting that when Bagging used with other ensemble base classifiers such as Additive Regression, Random Subspace and Random Forest have significantly enhanced its prediction accuracy, which confirms the superior prediction performance of ensemble learning approaches when adopted within Bagging.

From tables 2(a, b), it is inferred that Random Forest overall have performed the best i.e. either when employed as a homogenous ensemble classifier or as a base classifier within Bagging with a highest correlation coefficient equal to 0.86 and a least Relative Absolute Error (RAE) equal to 46.57%. Among single base algorithms, the performance of SMOreg and Lazy.KStar was the best, followed by MLP, M5P, M5Rules and REPTree. The poor performance of Random Tree have made it put at the bottom of the table, however, interestingly, its performance within Bagging (2b) have remarkably improved by 11%, which puts it to top 3 classifiers' list. In fact, the prediction accuracies of Bagged Random Tree was found almost equal to that of Random Forest when used on its own (independent of Bagging). KStar and SMOreg (SVM) were the only classifiers which stayed unaffected due to stable SVM algorithms [**9**] when employed as a base classifiers in

Bagging, hence the correlation coefficient for both remain unchanged, however, a study show that SVM within Bagging performs better [10]. Apart from that all single base as well as ensemble classifiers when employed as a base classifier in Bagging have significantly improved the prediction accuracy with a higher correlation coefficient and lower error.

It is a fact that Artificial Neural Networks especially MLP is the most commonly used machine learning technique for atmospheric pollution prediction and it suffers from problems related to over fitting and local minima. Therefore, the study shows the ability of ensemble classifier – Bagging which cannot just resolve MLP's problem of local minima and overfitting but also results in an enhanced accuracy.

Table-2: Single based and ensemble learning classifiers

Classifiers	Algorithms	Independent of Bagging 2(a)				With Bagging 2(b)			
		CC	MAE	RMSE	RAE (%)	CC	MAE	RMSE	RAE (%)
Ensemble classifiers	Random Forest	0.87	0.08	0.12	46.48	0.87	0.08	0.11	46.45
	Random Subspace (REPTree)	0.83	0.09	0.13	53.13	0.84	0.09	0.13	51.95
	Additive Regression	0.78	0.11	0.146	60.47	0.80	0.11	0.14	58.71
Single-based algorithms	Lazy.Kstar	0.83	0.10	0.14	49..53	0.84	0.09	0.13	47.84
	M5P	0.84	0.09	0.13	50.49	0.85	0.09	0.12	49.08
	M5Rules	0.81	0.10	0.14	55.14	0.81	0.10	0.14	54.54
	REPTree	0.81	0.10	0.14	55.34	0.85	0.089	0.123	49.13
	Lazy.Ibk	0.79	0.104	0.16	57.76	0.82	0.093	0.136	51.43
	SMOreg (SVM)	0.85	0.09	0.12	47.17	0.85	0.09	0.12	47.1
	Multilayer Perceptron	0.82	0.11	0.14	59.64	0.83	0.098	0.13	54.2
	Random Tree	0.73	0.12	0.17	66.38	0.85	0.088	0.122	48.71

As the results discussed in table 2 do not involve statistical significance of classifiers therefore, to further evaluate the performances of predictive models, a comparison was drawn by using WEKA implemented "Experimenter" tab and is shown in table 3. To evaluate the statistical significance of different predictive classifiers, a statistical test named T-tester (corrected) was performed with a confidence interval of 5% by using 10-fold cross validation, and the results were compared based on correlation coefficient obtained. With selected classifiers the focus of the experiment was to compare the performance of ensemble learning algorithms against the most widely used classification algorithms. The results of the experiment presented in table-3 include two characters (v and *) beside the correlation coefficient figure indicate the level of significance. The experiment performed is based on the comparison with the first classifier, in which "v" besides correlation coefficient indicates that the classifier has performed significantly better than the base classifier, whereas the other character is a symbol of poor performance as compared to the baseline classifier. Meanwhile, in case none of the character appears is an indication of neither better nor worse performance of the classifier against the baseline classifier.

For the first experiment four classifiers named M5P, M5Rules and KStar were picked and compared against Random Forest from table-2(a). The performance of classifiers using statistical significance and correlation coefficient revealed that the

accuracy of Random Forest is far better than other three classifiers. In a similar manner Random Forest was tested against widely popular algorithms MLP and SVM for atmospheric pollution concentrations during second experiment has also proved the superiority of Random Forest over the single based classifiers.

In experiment number three, all ensemble classifiers such as Random Subspace, Random Tree, and Random Forest with and without Bagging were compared, where Random Forest yet again outperformed the others. Lastly, two most popular classification techniques MLP and SVM and an ensemble classifier Random Forest were put together for a comparison proved the superiority of the ensemble learning classification techniques.

Table 3: Prediction model comparison

Experiment	Classifiers	Correlation Coefficient
1	Random Forest	0.87
	M5P	0.84*
	M5Rules	0.81*
	KStar	0.84*
2	SMOreg (SVM)	0.85
	Multilayer Perceptron	0.82*
	Random Forest	0.87v
3	REPTree	0.81
	Bagged REPTree	0.85v
	Random Subspace	0.83v
	Bagged Random Subspace	0.84v
	Random Forest	0.87v
	Bagged Random Forest	0.87v
4	SMOreg	0.85
	Multilayer perceptron	0.82
	Random Forest	0.87v
	Bagged SMOreg	0.85
	Bagged Multilayer perceptron	0.83*
	Bagged Random Forest	0.87v

4 Conclusion and future direction

For this paper 11 machine learning approaches including single learning and ensemble learning algorithms were tested and compared to predict the atmospheric carbon monoxide concentration. The study makes use of five concentration of atmospheric gases (SO2, NO, No2, NOx, ozone) and three environmental parameters (temperature, wind speed, wind direction) for the prediction of atmospheric concentration of carbon monoxide. The results obtained suggest that ensemble learning classifiers cannot just solve the problem of over fitting data and local minima which MLP and SVM suffer from, and perform better than state of the art traditional algorithms, but can also improve the performance of traditional classifiers when used as a base classifiers in Bagging.

5 References

1. The ongoing Challenge of managing carbon monoxide pollution in Fairbanks Alaska, (Interim report), National Academy Press, Washington D.C. pp 19-25, 2002
2. van Loon M, Vautard R, Schaap R, Bergström R, Bessagnet B, Brandt J, Builtjes P. J. H, Christensen J. H., Cuvelier C, Graff A, Jonson J. E, Krol M, Langner J, Roberts P, Rouil L, Stern R, Tarrasón L, Thunis P, Vignati E, White L, and Wind P, "Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble," Atmospheric Environment, vol. 41, no. 10, pp. 2083– 2097, 2007.

3. E. Alfaro, N. García, M. Gámez, and D. Elizondo, “Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks” Decis. Support Syst., vol. 45, no. 1, pp. 110–122, 2008.
4. W. Wang, C. Men, and W. Lu, “Online prediction model based on support vector machine” Neurocomputing, vol. 71, no. 4–6, pp. 550–558, 2008.
5. S. Abdul-Wahab and S. Al-Alawi, “Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks” Environmental Modelling & Software, vol. 17, no. 3, pp. 219–228, 2002.
6. U. Schlink, S. Dorling, E. Pelikan, G. Nunnari, G. Cawley, H. Junninen, A. Greig, R. Foxall, K. Eben, T. Chatterton, J. Vondracek, M. Richter, M. Dostal, L. Bertucco, M. Kolehmainen, and M. Doyle, “A rigorous inter-comparison of ground-level ozone predictions” Atmospheric Environment, vol. 37, no. 23, pp. 3237–3253 2003.
7. T. G. Dietterich, “Ensemble Methods in Machine Learning,” Lecture Notes in Computer Science, vol. 1857, pp. 1–15, 2000.
8. F. C. Morabito and M. Versaci, “Fuzzy neural identification and forecasting techniques to process experimental urban air pollution data” Neural networks : the official journal of the International Neural Network Society, vol. 16, no. 3–4, pp. 493–506, 2003.
9. “Stability (learning theory) - Wikipedia.” Available: https://en.wikipedia.org/wiki/Stability [Accessed: Aug 25, 2018].
10. G. Valentini, M. Muselli, and F. Ruffino, “Cancer recognition with bagged ensembles of support vector machines,” Neurocomputing, vol. 56, pp. 461–466, 2004.