Taylor & Francis
Taylor & Francis Group

Journal of Statistics and Management Systems

ISSN: 0972-0510 (Print) 2169-0014 (Online) Journal homepage: https://www.tandfonline.com/loi/tsms20

A new feature selection method based on machine learning technique for air quality dataset

Jasleen Kaur Sethi & Mamta Mittal

To cite this article: Jasleen Kaur Sethi & Mamta Mittal (2019) A new feature selection method based on machine learning technique for air quality dataset, Journal of Statistics and Management Systems, 22:4, 697-705, DOI: 10.1080/09720510.2019.1609726

To link to this article: https://doi.org/10.1080/09720510.2019.1609726

Published online: 25 Jun 2019.

Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 22 (2019), No. 4, pp. 697–705
DOI : 10.1080/09720510.2019.1609726

Taylor & Francis
Taylor & Francis Group

A new feature selection method based on machine learning technique for air quality dataset

Jasleen Kaur Sethi
Department of Computer Science and Engineering
University School of Information, Communication & Technology
Guru Gobind Singh Indraprastha University
Dwarka Sector-16C
New Delhi 110078
India

Mamta Mittal*
Department of Computer Science & Engineering
G.B. Pant Government Engineering College
Okhla Industrial Area
New Delhi 110020
India

Abstract

In the recent years, Air Pollution has become a matter of serious concern that leads to millions of premature deaths which motivates the researcher to predict the air quality in advance. Air Quality Index (AQI) always indicates the status of air quality. This index value is based on various pollutants like PM10, PM2.5, NO2, SO2, CO, O3, NH3, and Pb. Out of them PM2.5 nowadays is a major pollutant which is heavily affecting the quality of air. So, the major focus of this study is towards this parameter. PM2.5 is also dependent on various parameters, so in this study, a new feature selection method named as Causality Based Linear method has been proposed to select the most relevant parameters which affect the pollution. Experimental work has been carried out using existing machine learning techniques and proposed method on the air quality dataset of Delhi. It has been observed that the proposed method extracts wind speed, carbon monoxide and nitrogen dioxide as the key parameters and further accuracy of this method has been compared with the four existing methods where feature selection has not been considered. It has been found that the proposed method has given better accuracy with the key parameters.

Subject Classification: *(2010) 68M12*

Keywords: *Air Quality Index, Particulate Matter, Causality Based Linear method, Linear Regression, Decision Trees, Random Forest, Neural Networks.*

**E-mail:* mittalmamta79@gmail.com

1. Introduction

The traffic and the energy consumption continue to increase due to the urbanization and industrialization in developing countries. Number of contaminants like CO, CO_2, SO_2, NO_2 are released into atmosphere that lead to severe air pollution. Air Quality Index (AQI) is an assessment of the air quality that is closely related to human health [1, 16]. Therefore, prediction of Air Quality plays a vital role in management of air pollution. In literature, there are various Air Quality Prediction models: Deterministic Models, Statistical Models, Physical Models, Photochemical models and Machine Learning [18]. Although several models have been proposed, most of the models have high operational and storage overhead [2]. Machine learning approach overcomes these drawbacks and is an excellent tool for solving air pollution problems [15, 19]. In India, the air quality index is calculated based on the concentration of various pollutants. According to Liu et al [3], $PM_{2.5}$ is one of the most harmful air pollutants and has the highest influence on AQI out of all the pollutants Fine particulate matter ($PM_{2.5}$) consists of microscopic particles with diameter 2.5 μm or smaller. They can penetrate deeply into the lungs and cause number of health problems [11].

A plethora of research has been carried out for the prediction of $PM_{2.5}$. Ni et al [4] applied Autoregressive Integrated Moving Average (ARIMA) time series model for the short term prediction of $PM_{2.5}$ concentration for Beijing, China using data from multiple sources. Deters et al [2] used algorithms like Boosted Trees (BT), Linear Support Vector Machines (L-SVM), Rus Boosted Tree (RBT) to perform classification for the concentration of $PM_{2.5}$ for two stations in Quito, Ecuador. Regression analysis is performed using BT, L-SVM, Neural network and Convolutional Generalized Model (CGM). Using regression analysis, CGM gave the best results for the prediction of $PM_{2.5}$ concentration. Hourly spot concentration forecasting for Ozone, $PM_{2.5}$ and No_2 for six cities of Canada was performed using algorithms like Multiple Linear Regression (MLR), Online Sequential Multiple Linear Regression (OSMLR), Multilayer Perceptron Neural Network (MLPNN) and Online Sequential Extreme Learning Machine (OSELM) by Peng et al [5]. OSELM outperformed other techniques in the prediction of all three pollutants. Rybarczyk et al [6] performed the prediction of $PM_{2.5}$ for two stations of the city of Quito, Ecuator. To perform classification of $PM_{2.5}$, a number of algorithms like J48, ZeroR, Naïve Bayes were used. Out of all these algorithms the best accuracy for classification was given by J48 algorithm. The meteorological parameters along with the other pollutants

like CO, NO_2 and SO_2 are amongst the most influential factors that are used for the prediction of $PM_{2.5}$. Monitoring of relationship between these parameters and the concentration of $PM_{2.5}$ is required for the accurate prediction of the concentration of $PM_{2.5}$. Further, based on the relationship between parameters and the concentration of $PM_{2.5}$, parameter selection could be carried out. If the prediction of $PM_{2.5}$ is accurate, then effective measures to control the pollution due to $PM_{2.5}$ could be undertaken. In this paper, a new feature selection method called Causality Based Linear (CBL) model has been proposed for the prediction of $PM_{2.5}$ and further the analysis of air quality dataset has been performed using various machine learning techniques.

2. Proposed Method

Existing filter methods like Correlation analysis and Causality analysis ignore the dependencies amongst the features. However, dependencies amongst the features also play vital role in deciding the relevance of the feature in a dataset. Thus, a new method named as Causality Based Linear (CBL) has been proposed which is based on causal relationship among the features. Further, the relative importance of the individual feature has been determined using Linear Regression.

Interaction terms are used to compute the consequence of a predictor on the response based on another predictor variable. Consider a continuous response variable b with predictors α_1 and α_1. The expected value of response with the predictors based on linear regression is given by:

$$b = \beta_0 + \beta_1\alpha_1 + \beta_2\alpha_2 + \beta_{12}\alpha_1\alpha_2$$

Where βs are the regression coefficients [7, 12]. The interaction between the predictors α_1 and α_2 is analyzed by performing t-test on the regression coefficient β_{12}. For n predictors the total number of interactions when taking into consideration every order of interaction is 2^n which is exponentially large. Therefore, there is a need to compute a subset of significant interactions. CBL method is a useful tool that computes the significant two way interactions between predictors based on the results of Causality analysis and then uses linear regression to calculate the relative importance of each predictor. The steps of proposed methodology have been presented in Figure 1.

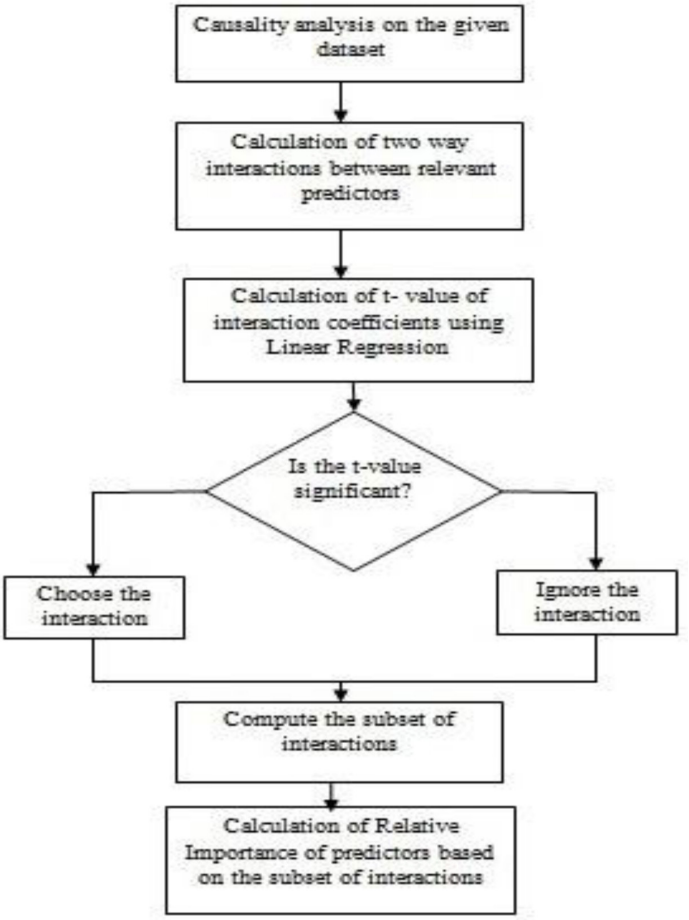

Figure 1

Steps of the Proposed Methodology

1. Causality analysis using Convergent Cross Mapping (CCM) technique [8] is applied to the given dataset and the relevant predictors are extracted.
2. Linear regression is carried out based on all the two way interactions between only the relevant predictors.
3. The relevant interactions are further chosen based on t-test on the regression coefficients.
4. Linear regression is used to find the relative importance of each predictor based on only the relevant interactions.

3. Performance Analysis of Proposed and Existing Methods

The analysis of air quality dataset of Delhi has been performed using Decision Trees, Random Forest, Linear Regression and Neural Networks. This analysis has been performed on two datasets: first is the complete dataset and second dataset has the features extracted by the CBL method. Decision Trees algorithms are based on recursive partitioning and perform classification on the basis of a given condition. Random forest recursively divide the dataset into collection of decision trees. Linear Regression fits a predictor as a linear function of responses [14]. Neural Networks are used to create a model based on human brain structured processing elements called neurons [9, 13, 17]. The air quality dataset of Delhi has been collected from Central Pollution Control Board (CPCB) website [10] from April 2015 to June 2018.

3.1 *Coefficient Estimates for Linear Regression*

A Causality analysis performed using Convergent Cross Mapping (CCM) technique on the given dataset results in bidirectional causality

Table 1

Coefficient estimates for Linear Regression

Parameter	Estimate	Std. Error	t value	Pr(> | t |)
AT	–1.9491	0.85646	–2.276	0.0231*
RH	0.14009	0.03139	4.462	0.00000913***
WS	–1.4192	0.56194	–2.525	0.0117*
CO	78.5856	12.3454	6.366	0.000000000309***
NO2	1.98317	0.31737	6.249	0.000000000637***
SO2	0.03459	0.04232	0.817	0.414
Ozone	0.08748	0.11595	0.754	0.4508
WS:CO	–0.8195	0.14746	–5.557	0.0000000361***
WS:NO2	–0.1838	0.03965	–4.635	0.0000041***
AT:WS	0.10103	0.01889	5.35	0.000000112***
CO:NO2	–0.1501	0.05863	–2.559	0.0106*
AT:CO	–1.3366	0.33888	–3.944	0.0000862***
AT:NO2	–0.0025	0.00955	–0.258	0.7968

between wind speed, absolute temperature and $PM_{2.5}$ and also between carbon monoxide and $PM_{2.5}$. Causality also exists between $PM_{2.5}$ and nitrogen dioxide. Thus, CBL method takes into account the interaction amongst wind speed, absolute temperature and the concentration of carbon monoxide and nitrogen dioxide to perform linear regression. The result of linear regression coefficients with their corresponding t- value and its associated probability is depicted in Table 1.

From the above table it is observed that only the interaction between wind speed and carbon monoxide, nitrogen dioxide and absolute temperature and between nitrogen dioxide and carbon monoxide and between absolute temperature and carbon monoxide are significant.

3.2 *Relative Importance of Parameters using Linear Regression*

Linear regression is performed using only the significant interactions and the relative importance of each parameter in percentage is summarized in Table 2.

From the above table, it has been found that the most relevant parameters chosen by the CBL method are the concentration of nitrogen dioxide, carbon monoxide and wind speed.

Table 2

Relative Importance of Parameters using Linear Regression

Parameter	Relative Importance (in percentage)
NO_2	41.2255154
CO	21.2922478
WS	18.3452669
AT	5.3834217
RH	3.0737116
SO_2	2.8377518
AT:WS	2.5199141
WS:CO	1.7881085
WS:NO_2	1.3162923
AT:CO	0.9862647
Ozone	0.6639625
CO:NO_2	0.5675425

Table 3
Accuracy Prediction of Proposed and Existing Methods on Various Parameters

	Complete Dataset (No feature Selection)				Proposed Method (With Feature Selection)			
Machine Learning Technique	r	R^2	AME	ACC	r	R^2	AME	ACC
Decision Tree	0.693	0.481	48.97	35.50	0.806	0.651	36.33	45.65
Random Forest	0.847	0.717	37.07	36.95	0.851	0.725	32.32	54.34
Linear Regression	0.713	0.509	51.08	30.43	0.839	0.705	35.73	41.30
Neural Network	0.730	0.533	49.82	30.43	0.856	0.734	32.44	45.65

3.3 *Accuracy Prediction of Proposed and Existing Methods on Various Parameters*

The air quality dataset of Delhi has been analyzed using machine learning techniques. Initially, the complete dataset has been analyzed using these techniques. Further, the analysis has been carried out on the features extracted by CBL method. In each case, the correlation coefficient (r), coefficient of determination (R^2), absolute mean error (AME) and accuracy (ACC) have been found out. The results in both the cases have been shown in Table 3.

From the above table, it has been observed that the experimental results improve when feature selection methods are used. It is further observed all the four models produce good results with the proposed CBL method. Out of all the machine learning techniques, Random Forest performed significantly better results compared to other techniques.

4. Conclusion

In this study, a new feature selection method called Causality Based Linear (CBL) model has been proposed for the prediction of $PM_{2.5}$. Further, the analysis of air quality dataset has been carried out using various machine learning techniques. The analysis has been performed for the whole dataset and for the subset of features extracted by CBL method. The performance of the models based on all the algorithms have been evaluated based on the correlation coefficient (r), coefficient of determination (R^2), absolute mean error (AME) and accuracy (ACC). Wind

speed, the concentration of carbon monoxide and nitrogen dioxide were found as the relevant features by the proposed CBL method. It has been observed that Random Forest has the highest accuracy in comparison with other techniques. It has been further observed that the experimental results improve when feature selection methods are used and all the four models produce good results with the proposed CBL method.

References

[1] Wang, D., Wei, S., Luo, H., Yue, C., & Grunder, O. A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. Science of The Total Environment, 580, 719-733 (2017).

[2] Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., Rybarczyk, Y. Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters. *Journal of Electrical and Computer Engineering*, 2017 (2017).

[3] Liu, C. M. Effect of PM2. 5 on AQI in Taiwan. Environmental Modelling & Software, 17(1), 29-37 (2002).

[4] Ni, X. Y., Huang, H., & Du, W. P. Relevance analysis and short-term prediction of PM $_{2.5}$ concentrations in Beijing based on multi-source data. Atmospheric Environment, 150, 146-161 (2017).

[5] Peng, H., Lima, A. R., Teakles, A., Jin, J., Cannon, A. J., & Hsieh, W.W. Evaluating hourly air quality forecasting in Canada with nonlinear updatable machine learning methods. Air Quality, Atmosphere & Health, 10(2), 195-211 (2017).

[6] Rybarczyk, Y., & Zalakeviciute, R. Machine learning approach to forecasting urban pollution. In Ecuador Technical Chapters Meeting (ETCM), (pp. 1-6) (2016).

[7] Norton, E. C., Wang, H., & Ai, C. Computing interaction effects and standard errors in logit and probit models. *Stata Journal*, *4*, 154-167 (2004).

[8] Sugihara, G., May, R., Ye, H., Hsieh, C. H., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. science, 1227079.

[9] Mittal, M., Goyal, L. M., Sethi, J. K., & Hemanth, D. J. Monitoring the Impact of Economic Crisis on Crime in India Using Machine Learning. *Computational Economics*, 1-19 (2018).

[10] http://cpcb.nic.in/

[11] Brokamp, C., Jandarov, R., Rao, M. B., LeMasters, G., & Ryan, P. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. Atmospheric Environment, 151, 1-11 (2017).

[12] Ai, C., & Norton, E. C. Interaction terms in logit and probit models. *Economics letters*, *80*(1), 123-129 (2003).

[13] Hemanth, D. J., Anitha, J., & Mittal, M. Diabetic Retinopathy Diagnosis from Retinal Images Using Modified Hopfield Neural Network. *Journal of medical systems*, *42*(12), 247 (2018).

[14] Walid Moudani, Ahmad Shahin, Fadi Shakik & Félix Mora-Camino, Dynamic programming applied to rough sets attribute reduction, *Journal of Information and Optimization Sciences,* 32:6, 1371-1397 (2011)

[15] Philibert, A., Loyce, C., & Makowski, D. Prediction of N 2 O emission from local information with random forest. Environmental pollution, 177, 156-163 (2013).

[16] Singh, K. P., Gupta, S., & Rai, P. Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmospheric Environment, 80, 426-437 (2013).

[17] Shaban, K. B., Kadri, A., & Rezk, E. Urban air pollution monitoring system with forecasting models. *IEEE Sensors Journal,* 16(8), 2598-2606 (2016).

[18] Jasleen Kaur Sethi, Mamta Mittal A Study of Various Air Quality Prediction Models. Circulation in Computer Science, ICIC 2017, 128-131 (2018).

[19] Prableen Kaur, Manik Sharma, Mamta Mittal, "Big Data and Machine Learning Based Secure Healthcare Framework", Procedia Computer Science, Elsevier, Volume 132, pp. 1049-1059 (2018).