Earth Science Informatics (2021) 14:1777–1786
https://doi.org/10.1007/s12145-021-00618-1

RESEARCH ARTICLE

An efficient correlation based adaptive LASSO regression method for air quality index prediction

Jasleen Kaur Sethi[1] · Mamta Mittal[2]

Received: 19 February 2021 / Accepted: 12 April 2021 / Published online: 29 April 2021

Abstract

One of the adverse effects of population growth and urbanization in developing countries is air pollution. Due to which more than 4.2 million deaths occur every year. Therefore, prediction of air quality is a subject worth in-depth research and has received substantial interest in the recent years from academic units and the government. Feature selection methods are applied before prediction to identify potentially significant predictors based on exploratory data analysis. In this research work, a feature selection method based on Least Absolute Selection and Shrinkage Operator (LASSO) named Correlation based Adaptive LASSO (CbAL) Regression method has been proposed for predicting the air quality. For the experimental evaluation, cross regional data, including the concentration of pollutants and the meteorological factors of Delhi and its surrounding cities, has been taken from the Central Pollution Control Board (CPCB) Website. Further, to validate this feature selection method, various machine learning techniques have been taken into consideration and some preventive measures have been suggested to enhance the air quality. Feature selection analysis reveals that carbon monoxide, sulphur dioxide, nitrogen dioxide and Ozone are the most important factors for forecasting the air quality and the pollutants found in the cities of Noida and Gurugram have a more substantial impact on the Air Quality Index of Delhi than other surrounding cities. The model evaluation depicts that the feature subset extracted by the proposed method performs better than the complete dataset and the subset extracted by LASSO Regression with an average classification accuracy of 78%. The findings of this study can help to identify important contributors of AQI so that viable measures to improve the air quality of Delhi can be carried out.

Keywords Air quality index · Feature selection · Correlation · Least absolute selection and shrinkage operator (LASSO) regression · Adaptive LASSO regression

Introduction

According to World Health Organization (WHO) (World Health Organisation n.d.) , 98% of the urban areas in the developing regions don't follow the air quality guidelines (https://www.who.int). The poor air quality is credited to several elements like the expansion in the number of motorized vehicles, the utilization of fuels that influence the environment and the inadequate environmental regulations (Singh et al. 2013). This leads to significant consequences on human health, the severity of which relies on air contamination concentration and the time of exposure. An Air Quality Index (AQI) is a tool employed to communicate about the status of air quality to the people. It converts the complex data of the concentration of various air pollutants into a single value. This index is used to warn the general public about the adverse health effects of high AQI (Wang et al. 2017). Many programmes for predicting air quality in various cities across India have been introduced, but their functionality is limited due to high costs and manpower requirements. Although in literature, various models, namely deterministic models, physical models, statistical models, and photochemical models, are available. However, in some of these models, real time forecast is not possible and has high storage overhead. Therefore machine learning techniques like classification (Tahir et al. 2019) and clustering (Mittal et al. 2014; Mittal et al. 2015) have been widely used. Additionally, AQI is also influenced by the cross regional transport of

Communicated by: H. Babaie

✉ Mamta Mittal
mittalmamta79@gmail.com

[1] Research Scholar, USICT, GGSIPU, New Delhi, India

[2] Department of CSE, G. B. Pant Government Engineering College, New Delhi, India

pollutants from the surrounding cities apart from many meteorological drivers and other pollutants like $PM_{2.5}$, CO, SO_2, ozone and NO_2 (Liu et al. 2018; Zhai and Chen 2018). Thus, the cross regional factors that include the concentration of pollutants from neighbouring cities of the study area and the meteorological factors must be considered for accurate prediction of AQI. Feature selection improves this prediction performance by extracting the features that efficiently describe the data and eliminate irrelevant ones (Chandrashekar and Sahin 2014). In this work, to predict the AQI of Delhi, the proposed feature selection method has been applied on the dataset comprising of meteorological factors and pollutant data of Delhi and surrounding cities: Faridabad, Gurugram, Noida and Ghaziabad. Further, these feature selection results have been validated using various machine learning techniques. The layout of the remaining paper has been organized as follows. In Section 2, related work has been presented. The following section presents the details about the study area. The proposed feature selection method for AQI prediction has been explained in Section 4. In Section 5, the experimental results with validation of the proposed method have been discussed. A brief conclusion has been presented in Section 6.

Related work

To monitor the relationship between various parameters that affect AQI concentration, feature selection is carried out that considers a set of features or attributes. This is required for the accurate prediction of AQI so that effective measures to control pollution could be undertaken. Feature selection is an important data preprocessing technique used to extract relevant features and eliminate redundant or noisy data used in number of applications like AQI prediction, diabetes prediction (Battineni et al. 2019), crop disease classification (Saeed et al. 2021; Tahir et al. 2021), human diseases prediction (Khan et al. 2020) and for medical imaging data (Naheed et al. 2020). With the increase in the domain of features, several methods have been proposed to reduce the irrelevant and redundant features. Feature Selection methods are generally categorized into: Filter methods, Wrapper methods and Embedded methods. These methods have been depicted in Fig. 1.

Filter Methods perform preprocessing by taking into account the properties of individual features and neglect the learning technique. These methods find a feature subset by ranking the individual features without considering the dependencies amongst them (Kumar and Minz 2014). Wrapper methods select the subset of features based on the interactions between them using a learning technique. The feature subset with the highest performance is selected. Whereas, Embedded methods take into consideration the merits of both Filter methods and Wrapper methods. They reduce the computation time of Wrapper methods by incorporating the selection of features during the training phase (Wang et al. 2016). The Least Absolute Selection and Shrinkage Operator (LASSO) (Tibshirani 1996) and Adaptive LASSO Regression (Zou 2006) are more efficient embedded methods for Feature Selection. Thus, they have been considered the focus of this research work and explained in more detail.

LASSO regression

LASSO applies regularization where a portion of regression coefficients is decreased to zero. For feature selection, all the coefficients with non zero values are selected and the prediction error is minimized. When the value of this parameter is very high, then the coefficients of the regression variable become zero. This regression technique is widely accepted as it provides good prediction accuracy and reduces overfitting (Melkumova and Shatskikh 2017). Consider a data with M samples (a_j, b_j) where $a_j = (a_1, a_2, \ldots\ldots, a_p)^T$ are the predictor variables and b_j is the response. Suppose the regression coefficients are $\beta = (\beta_1, \beta_2, \ldots\beta_p)^T$. LASSO aims to solve

$$min_{\beta_0,\beta}\left\{\sum_{j=1}^{M}\left(b_j-\beta_0-a_j^T\beta\right)^2\right\}+\ \lambda\sum_{k=1}^{p}|\beta_k| \qquad (1)$$

Where $\lambda \geq 0$ is a tuning parameter. This tuning parameter λ controls the amount of shrinkage. Suppose $\lambda_0 = \sum_k \beta_0$. These values of parameters $\lambda < \lambda_0$ will produce shrinkage and some coefficients will exactly become zero.

Adaptive LASSO regression

Adaptive LASSO is a LASSO regression method with oracle properties. These properties enable a regression method to choose the correct subset of predictors and have optimal estimation rates. Adaptive LASSO is a regularization method that does not overfit large regression coefficients and also provides predictor subset selection by shrinking some coefficients to zero (Qian and Yang 2013).

From (2), using adaptive weights w_k, the Adaptive Lasso minimizes

$$min_{\beta_0,\beta}\left\{\sum_{j=1}^{M}\left(b_j-\beta_0-a_j^T\beta\right)^2\right\}+\ \lambda\sum_{k=1}^{p}w_k|\beta_k| \qquad (2)$$

The adaptive weight vector is usually found using Ridge regression. This regression method performs the ordinary least square method as it avoid overfitting. It includes all the predictors and shrinks the coefficients towards zero, but not exactly to zero. This shrinkage is achieved by using a penalty term called L2 norm. The predictors are scaled before applying Ridge regression and the adaptive weights are defined as:

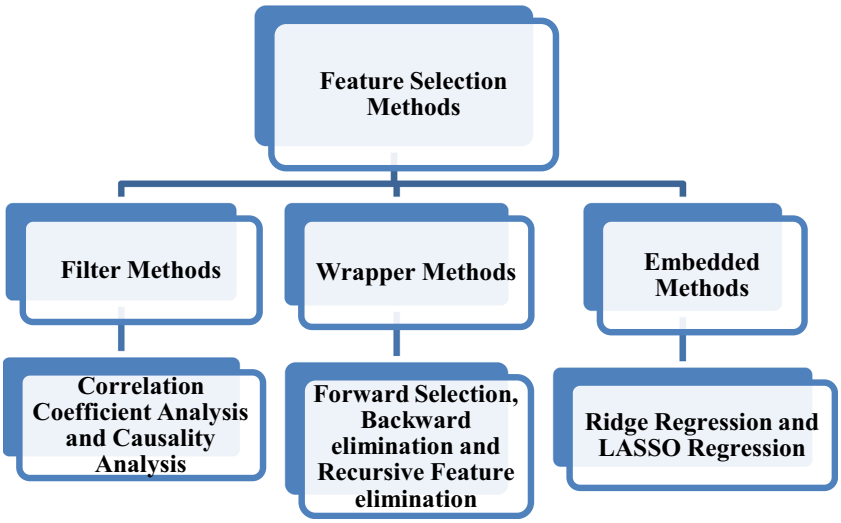

Fig. 1 Feature selection methods

$$w_k = \frac{1}{|\beta_k'|^q} \tag{3}$$

Where β_k' is the initial value of regression coefficients calculated usually using Ridge regression and q is a positive constant value used for the adjustment of adaptive weights taken as 0.5, 1 or 2. A plethora of information is available to study the association between AQI and meteorological parameters and other pollutants using embedded methods like LASSO and its variants, along with many other feature selection methods. Table 1 depicts the literature review of feature selection methods based on the study area, number of cities, sampling period, techniques and selected parameters used for predicting AQI.

From the above table, it has been observed that the meteorological parameters and the other pollutants like $PM_{2.5}$, CO, NO_2 and SO_2 are amongst the most influential factors that are used for the prediction of AQI. This work attempts to develop feature selection methods so that their anomalies in the existing techniques can be reduced and the accuracy and efficiency of air quality prediction can be enhanced. This can also be used to identify the parameters affecting AQI and thus air quality so that effective measures to improve air quality can be carried out.

Detailed description of study area

The National Capital Territory (NCT) of Delhi is located at 28.61°N 77.23°E in central India and has an extreme climate with very hot summers and very cold winters. The summer season extends from the month of May to July followed by the rainy season that extends from August till the end of September and winters occur from the month of November till February. The temperature in Delhi ranges from 25 °C to 45 °C and 5 °C to 22 °C in summers and winters, respectively (https://www.delhicapital.com/about-delhi/climate.html n.d.).

The daily air quality dataset of Delhi and its four surrounding cities: Faridabad, Gurugram, Noida and Ghaziabad, has been collected from the CPCB website (Central Pollution Control Board, Government of India n.d.)(http://cpcb.nic.in/) from January 1, 2018 to December 31, 2019 and consists of 730 samples. The study area is presented in Fig. 2 (http://www.maps-of-india.com/india-delhi-map/location-of-delhi/ n.d.).

The parameters collected for each city are Absolute Temperature (AT), Relative Humidity (RH), Wind Speed (WS), and concentration of CO, SO_2, NO_2, ozone and $PM_{2.5}$. The missing values have been replaced by their mean in the air quality dataset. These predictors from all the cities have then been combined to form a single dataset with the AQI value of Delhi as the response. AQI has been calculated from the concentration of all pollutants by constructing their respective sub indices and then by applying the maximum operator (Sethi and Mittal 2019a, b). Further, based on the value, it has been converted into one of the six subclasses: Good (0–50), Satisfactory (51–100), Moderately Polluted (101–200), Poor (201–300), Very Poor (301–400) and Severe (401–500). The relationship between AQI and the various parameters has been discussed as follows:

- Absolute Temperature: Air Quality tends to deteriorate in the winter season because the air close to the ground is cooler and the turbulence is low, leading to less dispersion of pollutant particles. Overall a negative correlation of temperature with AQI has been found out (Guo et al. 2011).
- Relative Humidity: As the relative humidity increases, there is an increase in the number of pollutants as a higher humidity value results in more vapours being attached to various pollutants. Thus, a negative correlation of relative humidity with AQI has been found (Li et al. 2014).

Table 1 Literature Review of Feature Selection Methods for the Prediction of AQI

Reference	Study Area	Number of Cities	Sampling Period	Feature Selection Technique	Selected Parameters
Liu et al. 2018	China	4	Jan 1, 2014 to April 30, 2016	Information Gain and Support Vector Machines (SVM)	$PM_{2.5}$, PM_{10}, wind direction, wind power, CO, NO_2, SO_2 and ozone
Zhai and Chen 2017	China	1	Feb 1, 2013 to t July 31, 2016,	LASSO and Ridge Regression	NO_2, $PM_{2.5}$, and PM_{10} and meteorological factors such as minimum barometric
Hajek and Olej 2015	Czech Republic	3	Jan 1, 2009 to Dec 31, 2011	Correlation	Ozone and PM_{10}
Li et al. 2014	China	1	Jan 1, 2001 to Dec 31, 2011	Correlation	Temperature, relative humidity, precipitation and wind speed
Liu and Chen 2020	China	6	Jan 1, 2016 to Dec 21, 2018	Correlation and Mutual Information	$PM_{2.5}$, PM_{10}
Mahanta et al. 2019	India	1	Jan 1, 2015 to April 4, 2017	Regression Analysis	Meteorological data
Qi et al. 2018	China	1	Nov 1, 2015 to Dec 31, 2015	Neural Network	PM2.5, CO, PM10, Wind Strength (North), and Temperature
Zheng et al. 2013	China	1	Aug1, 2012 to Dec. 31, 2012	Decision Trees	PM_{10} and NO_2
Gu et al. 2019	China	1	Jan 1, 2018 to Dec 31, 2018,	Correlation	$PM_{2.5}$, SO_2 and Ozone
Zhang et al. 2020	China	1	Apr 1, 2018 to May 31. 2018	Correlation	CO and PM_{10}

- Wind Speed: High wind speed in the summer season blows away the air pollutants and improves air quality. Low wind speed in the winter season causes poor air quality (Li et al. 2014).
- Carbon Monoxide: AQI and CO are positively correlated as smoke from vehicular pollution and stubble burning are one reason for the poor air quality. Thus, CO is one of the crucial indicators that affect AQI (Liu et al. 2018).
- Sulphur Dioxide: It has been found out the emissions of SO_2 mainly occur with the emissions of PM and that gases like SO_2 contribute to the formation of these fine particles and thus increase AQI(Liu et al. 2018).
- Nitrogen Dioxide: The sources of NO_2 include emissions from motor vehicles and power plants. It also contributes to the formation of ozone. One of the reasons for the poor AQI has been attributed to the emissions of NO_2 from vehicles and industries (Zhai and Chen 2017).

Fig. 2 Study area

- Ozone: Secondary pollutant formed from nitrogen oxides in the presence of solar radiation. It causes adverse effects to human health and crops. It has been associated with cardio and respiratory problems. It plays a vital role in determining air quality (Hajek and Olej 2015).
- Particulate Matter: Combination of solid and liquid drops found in the air by coal combustions, vehicles and industries. $PM_{2.5}$ is one of the most harmful pollutants that have the strongest influence on the AQI (Sethi and Mittal 2019a, b).

The collected dataset has been analyzed using column charts. Column charts are means used to compare values across various categories. The column chart for the dataset of all the cities with the mean values of all predictors is depicted in Fig. 3.

The basic statistics: mean, median and standard deviation (SD) of all meteorological parameters and concentration of pollutants for various cities are presented in Table 2. These statistics are helpful to compare the concentration of predictors of all the cities.

From Table 2 and column charts, it can be observed that with respect to the pollutants, Faridabad has highest mean value of Ozone, Ghaziabad has highest mean of SO_2 and Delhi has highest mean values of three pollutants: CO, NO_2 and $PM_{2.5}$ compared to all other cities.

Proposed feature selection method

LASSO Regression method performs both regularization and feature selection where the total regression coefficients are constrained to be lower than a threshold fixed value. It reduces the dimensionality by using the tuning parameter, which signifies penalty. For large regression coefficients, LASSO has accurate subset selection, but lacks optimal prediction rates and in some cases, optimal estimation rates have incorrect predictor selection. Adaptive LASSO overcomes the drawbacks of LASSO regression as it has oracle properties. It has a good prediction rate and also computes the correct subset of predictors by using adaptive weights. The Adaptive LASSO regression method calculates the weights based on the initial regression coefficients calculated using Ridge regression with a positive constant q with value 0, 1 or 2, which is used to adjust these weights. To choose the value of this positive constant in Adaptive LASSO, a new method named Correlation based Adaptive LASSO Regression has been proposed. This method utilizes correlation to adjust the weights in the Adaptive LASSO method. The proposed methodology of feature selection for AQI prediction is depicted in Fig. 4. As shown in the figure, various predictors have been collected and the dataset has been pre-processed by replacing the missing values with their mean. Then, AQI has been calculated from the pollutants and is further converted into six classes. Next, the proposed CbAL Regression method has been applied for feature selection. The mathematical model for the proposed feature selection method has been shown in Fig. 5.

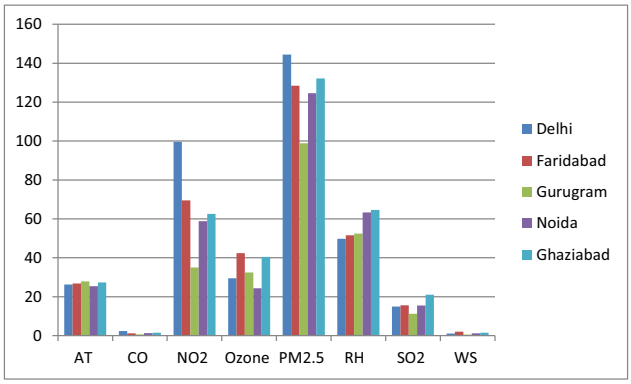

Fig. 3 Column Chart of Air Quality Dataset of Various Cities

For a dataset with M samples with predictors $a_j = (a_1, a_2, \ldots\ldots, a_p)$ and the response b_j, first, the correlation between all predictors and response is calculated using Pearson correlation coefficient r as:

$$r = \frac{\sum_{j=1}^{M}\left(a_j - a'\right)\left(b_j - b'\right)}{\sqrt{\sum_{j=1}^{M}\left(a_j - a'\right)^2 \sum_{j=1}^{M}\left(b_j - b'\right)^2}} \tag{4}$$

Where a' and b' are the mean values of each predictor a and response b respectively. Then, the adaptive weight vector is calculated based on the correlation coefficient as:

$$w_j = \frac{1}{\left|\beta'_j\right|^r} \tag{5}$$

Where the initial regression coefficients β'_j are calculated using ridge regression as:

$$\beta'_j = min_{\beta_0,\beta}\left\{\sum_{j=1}^{M}\left(b_j - \beta_0 - a_j^T\beta\right)^2\right\} + \lambda\sum_{k=1}^{p}|\beta_k^2| \tag{6}$$

Next, the regression coefficients are found using the Adaptive LASSO method based on the computed weights. Finally, the predictors with non zero values of regression coefficients are selected. The algorithm of the proposed CbAL Regression method is as follows:

Results and discussion

For the prediction of AQI, there is a need to select the relevant features and eliminate the noisy and irrelevant ones. Experimental work has been carried out on the air quality dataset of Delhi and its neighbouring cities.

Table 2 Statistics of the parameters in the study area

Parameter	Delhi			Faridabad			Gurugram			Noida			Ghaziabad		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
AT (°C)	26.21	26.35	6.92	26.83	28.49	7.28	27.91	27.91	4.81	25.35	27.05	7.16	27.33	28.99	7.14
CO (mg/m3)	2.32	2.18	0.94	1.13	1.13	0.80	0.60	0.46	0.54	1.33	1.09	0.90	1.52	1.21	0.90
NO2 (µg/m3)	99.57	96.24	48.91	69.52	69.52	32.75	34.99	25.38	40.41	58.82	54.54	29.15	62.52	61.92	26.70
Ozone (µg/m3)	29.47	27.40	14.05	42.33	39.19	21.89	32.46	31.85	16.61	24.31	19.39	19.57	40.48	36.23	21.89
PM2.5 (µg/m3)	144.40	118.69	103.30	128.30	94.86	113.40	98.88	92.17	65.90	124.50	90.95	106.50	132.17	97.64	107.21
RH (%)	49.71	50.90	18.81	51.56	51.90	15.61	52.38	52.38	17.87	63.25	65.56	16.10	64.53	65.51	15.87
SO2 (µg/m3)	14.95	14.94	7.71	15.60	15.24	10.90	11.16	9.49	10.49	15.49	13.39	9.19	21.08	17.83	15.32
WS (km/h)	1.04	0.88	0.71	2.05	1.98	1.11	0.58	0.58	0.32	1.14	1.08	0.45	1.48	1.033	1.25

Algorithm for Proposed CbAL Regression Method

1. Input the Dataset
2. for each a_j do
3. compute the Correlation $C(a_j, b_j)$
4. $r_j = C(a_j, b_j)$
5. end for
6. for each a_j do
7. compute w_j
8. compute initial β'_j
9. end for
10. for each a_j and b_j
11. compute λ
12. compute $\beta = (\beta_1, \beta_2, \ldots \beta_p)$
13. end for
14. for each β_j
15. if $\beta_j \neq 0$
16. select corresponding a_j
17. end for

Experimental results of proposed CbAL regression method

The correlation between all the predictors and AQI has been computed in the proposed CbAL Regression method using the Pearson correlation coefficient. These values are further assigned as a positive constant used to adjust the weight of the corresponding predictor. These absolute correlation values between all the predictors and the air quality index have been depicted in Table 3.

From Table 3, the correlation values of the predictors are used to adjust the weights and then ridge regression has been carried out to find the initial value of regression coefficients. The weights have been computed based on initial coefficients and correlation. Next, adaptive lasso regression has been performed based on the calculated weights. The final regression coefficients computed by proposed CbAL regression method have been depicted in Table 4.

From the above table, it has deduced that only the predictors: OzoneGH, RHGH, SO_2GU, COGU, OzoneGU, ATGU, RHGU, WSGU, NO_2GU, CONO, SO_2NO, NO_2FA, ATDL,

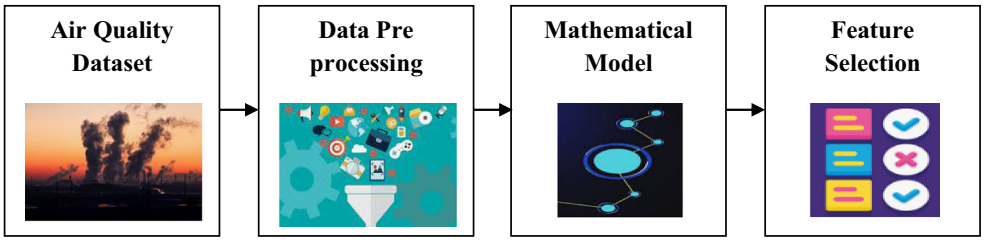

Fig. 4 Methodology for Feature Selection of AQI Prediction

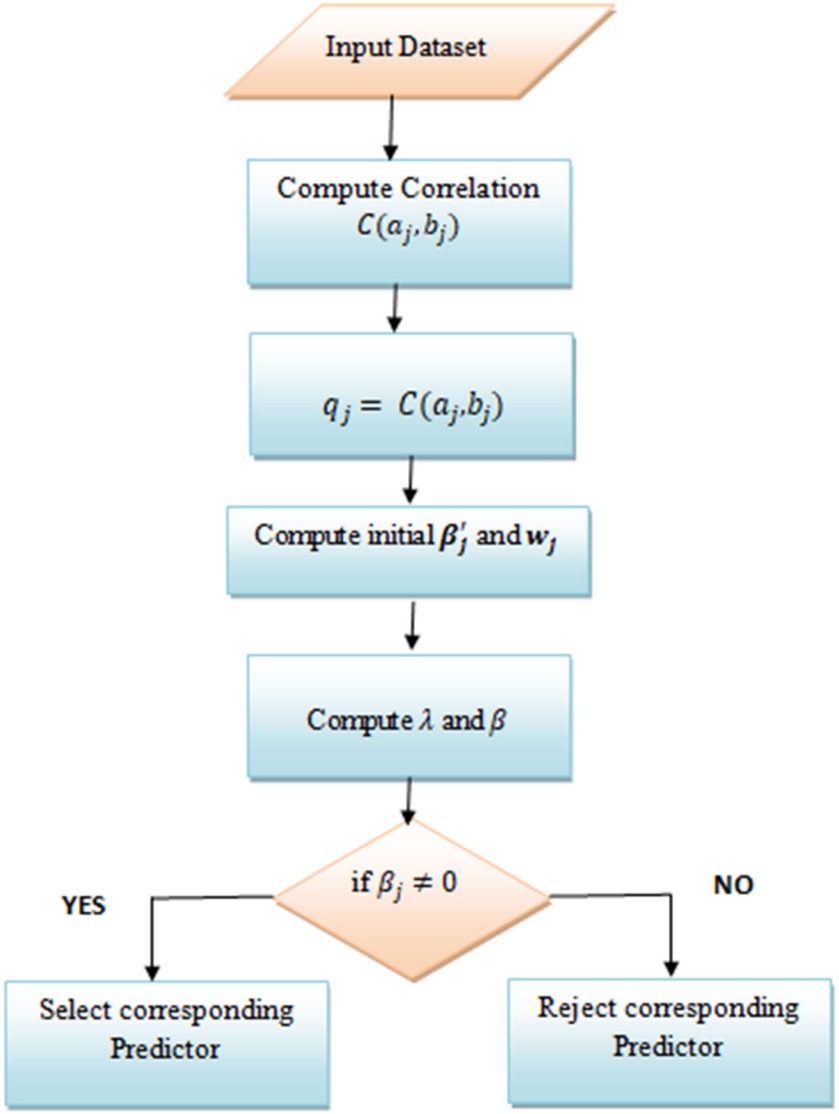

Fig. 5 Mathematical Model for CbAL Regression Method

$PM_{2.5}$DL, RHDL, WSDL, CODL, NO_2DL, OzoneDL, SO_2DL influence the AQI of Delhi. The plots of the coefficient estimate computed by LASSO and the proposed CbAL Regression method have been shown in Fig. 6.

Validation of proposed feature selection method

Updateable Naive Bayes, Decision Trees and Artificial Neural Network (ANN) have been used to validate the proposed CbAL Regression method. These results have been compared to the whole dataset without feature selection and with the whole dataset extracted by LASSO Regression. The parameters used for evaluation are weighted recall, weighted precision, weighted F1 score and accuracy. Weighted F1 score is calculated by multiplying the F1 score of each class by the number of samples from that class and dividing it by the total number of samples. In the similar manner weighted precision and weighted recall have been computed based on the precision and recall of each class and their respective number of samples. The validation results are summarized in Table 5. Naive Bayes is a probability based classifier which requires only a single scan of the dataset and is also useful for big data applications (Mittal et al. 2019). Updateable Naive Bayes is an incremental learning algorithm used to minimize the objective function. It uses a default precision of 0.1 as an incremental update for numeric valued predictors (Gladence et al. 2015). Decision Trees have nodes and classify the dataset based on the values of the various predictors. ANN considers the activation function, network architecture and weights of the input functions that are adjusted to reduce the difference between the predicted output and the actual output, so that the error rate is reduced (Kotsiantis 2007).

From the above table, it has been observed that for the aforementioned machine learning techniques, the feature set extracted by the proposed CbAL Regression method depicts higher value of all performance evaluation parameters than the

Table 3 Correlation between Predictors and AQI of Delhi

Predictor	Correlation	Predictor	Correlation
$PM_{2.5}$GH	0.834631	$PM_{2.5}$NO	0.684137
ATGH	0.529538	ATNO	0.541071
SO_2GH	0.202368	RHNO	0.039822
NO_2GH	0.716162	WSNO	0.435244
COGH	0.731283	COFA	0.454239
OzoneGH	0.133825	SO_2FA	0.115971
RHGH	0.094663	NO2FA	0.184773
WSGH	0.264966	OzoneFA	0.107208
SO_2GU	0.09532	WSFA	0.437636
COGU	0.504674	RHFA	0.201921
OzoneGU	0.112366	$PM_{2.5}$FA	0.598772
$PM_{2.5}$GU	0.625593	ATFA	0.522114
ATGU	0.235356	ATDL	0.539391
RHGU	0.093515	$PM_{2.5}$DL	0.940196
WSGU	0.384794	RHDL	0.075104
NO_2GU	0.411947	WSDL	0.415687
CONO	0.634667	CODL	0.562552
OzoneNO	0.020022	NO_2DL	0.634017
NO_2NO	0.540441	OzoneDL	0.099795
SO_2NO	0.284255	SO_2DL	0.345301

GH Predictors of Ghaziabad, *GU* Predictors of Gurugram, *NO* Predictors of Noida, *FA* Predictors of Faridabad, *DL* Predictors of Delhi.

Table 4 Coefficients Computed by Proposed CbAL Regression Method

Predictor	Coefficient	Predictor	Coefficient
$PM_{2.5}$GH	.	$PM_{2.5}$NO	.
ATGH	.	ATNO	.
SO_2GH	.	RHNO	.
NO_2GH	.	WSNO	.
COGH	.	COFA	.
OzoneGH	0.034392	SO_2FA	.
RHGH	−0.46233	NO2FA	−0.21034
WSGH	.	OzoneFA	.
SO_2GU	−0.20418	WSFA	.
COGU	6.933239	RHFA	.
OzoneGU	−0.06239	$PM_{2.5}$FA	.
$PM_{2.5}$GU	.	ATFA	.
ATGU	1.008346	ATDL	−2.20638
RHGU	−0.03386	$PM_{2.5}$DL	1.062738
WSGU	3.99696	RHDL	−0.10128
NO_2GU	0.012798	WSDL	−11.4404
CONO	−7.74861	CODL	3.522296
OzoneNO	.	NO_2DL	0.165982
NO_2NO	.	OzoneDL	0.109845
SO_2NO	1.010614	SO_2DL	0.683592

complete dataset without feature selection and the dataset extracted by LASSO regression method.

Discussion

Environmental Pollution has led to both acute and chronic consequences on human health, the severity of which depends on the concentration of air pollutants and the time of exposure. AQI is a mathematical tool that defines pollution and is influenced by various pollutants, meteorological parameters, and the cross regional transport of pollutants from surrounding cities. Thus, there is a need to develop efficient models for the air quality prediction capable of computing the seasonal changes in the air quality of a particular region. The prediction model's performance deteriorates when irrelevant and noisy features are used to develop such a model. Therefore, a feature selection method named CbAL Regression method has been proposed for AQI prediction of Delhi based on the air quality dataset of four neighbouring cities collected from the CPCB website. It has been found that the proposed method emphasized on OzoneGH, RHGH, SO_2GU, COGU, OzoneGU, ATGU, RHGU, WSGU, NO_2GU, CONO, SO_2NO, NO_2FA, ATDL, $PM_{2.5}$DL, RHDL, WSDL, CODL, NO_2DL, OzoneDL, SO_2DL parameters. The highest value of regression coefficients as computed by the proposed algorithm are CO for Delhi, Gurugram and Ozone, SO_2 and NO_2 for Ghaziabad, Noida and Faridabad respectively. To validate the proposed method, various machine learning algorithms have been used. It has been deduced that the feature subset extracted by the proposed method outperformed the complete dataset and the one extracted by LASSO Regression. From the results, it has been concluded that the pollutants that affect the AQI are CO, Ozone, NO_2, and SO_2. Thus, some of the preventive measures to curb air pollution due to the aforementioned pollutants have been discussed as follows:

- One of the significant sources of air contamination is waste burning. Thus, there is a need to prioritize regulatory focus on waste burning and use more structured waste management methods so that pollution due to CO can be avoided (Wiedinmyer et al. 2014).
- Emission from crop residue burning is another source of various air pollutants. There is a need to generate awareness amongst farmers about the negative effects of residue burning and proper management of these residues through its use for bio-energy to maintain soil productivity (Jain et al. 2014).
- Use of solar energy and geothermal energy as an alternative to coal-based thermal power plants is a significant source of SO_2 (Streets et al. 2007).
- Due to the CNG conversion of public transit vehicles, an increase of 13.7% in the concentration of NO_2 was

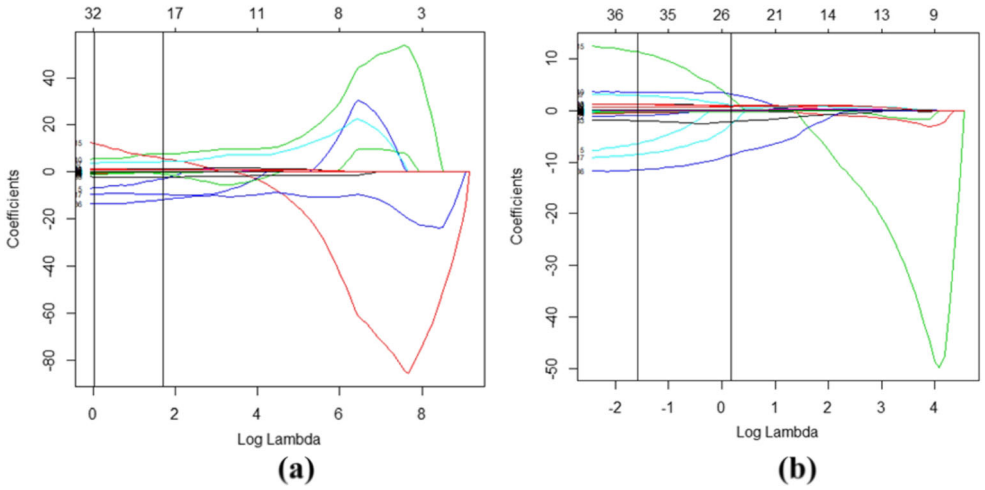

Fig. 6 Coefficient Estimates of (**a**) LASSO Regression (**b**) Proposed CbAL Regression Method

observed (Chelani and Devotta 2007). This concentration of NO_2 increases with the age of the CNG engine. Therefore, maintenance of public vehicles with CNG engines is necessary to control pollution.

- Volatile Organic Compounds from vehicles and the burning of other fuels are a significant source of Ozone. Thus, cleaner fuels and batteries in vehicles are crucial to curb Ozone pollution.

Conclusions and future work

The adequacy for the prediction of the Air Quality can be improved by utilizing feature selection methods. In this research, a feature selection method named CbAL Regression has been proposed for the AQI prediction. Experimental work has been performed on the dataset of Delhi and significantly, eight parameters have been contemplated by the authors to demonstrate air contamination levels, also taking into account the cross regional transport of pollutants from its neighboring cities: Faridabad, Gurugram, Noida and Ghaziabad. Further, the proposed method has been applied to consider the most relevant features that affect the prediction model's performance abruptly. It has been inferred that the number of parameters selected by the proposed CbAL Regression method is: eight, seven, two, two and one from the cities of Delhi, Gurugram, Ghaziabad, Noida and Faridabad, respectively. The regression coefficients computed by the proposed method are 3.522 for CO in Delhi, 6.933 for CO in Gurugram and 0.034 for Ozone, 1.010 for SO_2 and − 0.210 for NO_2 in Ghaziabad, Noida and Faridabad, respectively. Thus, the concentration of Ozone, CO, SO_2 and NO_2 are the most directly applicable influencing factors for the prediction of AQI of Delhi identified by the proposed method. Further, the air quality in Delhi is susceptible to surrounding pollutants from Gurugram and Noida. Results of experiments validate that our proposed CbAL Regression method has acquired the best performance among all the machine learning methods considered in terms of various performance metrics. The model evaluation demonstrates that the subset extracted by the proposed model performs about 5% better than the overall dataset and 10% better than the subset extracted by LASSO regression method in determining the average classification accuracy. The limitation of this study is that the AQI may be influenced by other temporal factors like hour of the day or day of the week which are not considered in this work. In future, more predictors may be added based on the temporal factors for further analysis.

Table 5 Validation of the Proposed CbAL Regression Method

Machine Learning Technique	Updateable Naive Bayes				Decision Tree				ANN			
	W. Recall	W. Precision	W. F1 Score	Acc	W. Recall	W. Precision	W. F1 Score	Acc	W. Recall	W. Precision	W. F1 Score	Acc
No Feature Selection	0.54	0.55	0.54	0.69	0.57	0.57	0.56	0.70	0.49	0.31	0.25	0.65
LASSO Regression	0.57	0.57	0.56	0.70	0.67	0.67	0.67	0.80	0.50	0.38	0.29	0.69
Proposed Method	0.63	0.60	0.61	0.74	0.70	0.70	0.70	0.83	0.63	0.45	0.33	0.77

W.Recall Weighted Recall, W.Precision: Weighted Precision, *W.F1 Score* Weighted F1 score, *Acc* Accuracy.

References

Battineni G, Sagaro GG, Nalini C, Amenta F, Tayebati SK (2019) Comparative machine-learning approach: a follow-up study on type 2 diabetes predictions by cross-validation methods. Machines 7(4):74

Central Pollution Control Board (CPCB), Government of India (n.d.) . http://cpcb.nic.in/. (accessed 20th January, 2021)

Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28

Chelani AB, Devotta S (2007) Air quality assessment in Delhi: before and after CNG as fuel. Environ Monit Assess 125(1–3):257–263

Gladence LM, Karthi M, Anu VM (2015) A statistical comparison of logistic regression and different Bayes classification methods for machine learning. ARPN J Eng Appl Sci 10(14):5947–5953

Gu K, Zhou Y, Sun H, Zhao L, Liu S (2019) Prediction of air quality in Shenzhen based on neural network algorithm. Neural Comput & Applic 32:1879–1892. https://doi.org/10.1007/s00521-019-04492-3

Guo Y, She F, Wang S, Liu B, Li J, Wang J (2011) Assessment on air quality in Lanzhou and its relation with meteorological conditions. J Arid Land Resour Environ 25(11):100–105

Hajek P, Olej V (2015) Predicting common air quality index–the case of Czech microregions. Aerosol Air Qual Res 15(2):544–555

http://www.maps-of-india.com/india-delhi-map/location-of-delhi/ (n.d.) (accessed 20th January 2021)

https://www.delhicapital.com/about-delhi/climate.html (n.d.) (accessed 20th January 2021)

Jain N, Bhatia A, Pathak H (2014) Emission of air pollutants from crop residue burning in India. Aerosol Air Qual Res 14(1):422–430

Khan MA, Qasim M, Lodhi HMJ, Nazir M, Javed K, Rubab S, Din A, Habib U (2020) Automated design for recognition of blood cells diseases from hematopathology using classical features selection and ELM. Microsc Res Tech 1–15. https://doi.org/10.1002/jemt.23578

Kotsiantis SB (2007) Supervised machine learning: a review of classification techniques. Informatica 31:249–268

Kumar V, Minz S (2014) Feature selection: a literature review. SmartCR 4(3):211–229

Li L, Qian J, Ou CQ, Zhou YX, Guo C, Guo Y (2014) Spatial and temporal analysis of air pollution index and its timescale-dependent relationship with meteorological factors in Guangzhou, China, 2001–2011. Environ Pollut 190:75–81

Liu H, Chen C (2020) Spatial air quality index prediction model based on decomposition, adaptive boosting, and three-stage feature selection: a case study in China. J Clean Prod 265:121777

Liu B, Chang PC, Huang N, Li D (2018) Multi-level air quality classification in China using information gain and support vector machine. Int J Comput Inf Eng 12(12):1092–1101

Mahanta S, Ramakrishnudu T, Jha RR, Tailor N (2019) Urban air quality prediction using regression analysis. In TENCON 2019-2019 IEEE region 10 conference (TENCON) (pp. 1118-1123). IEEE, https://doi.org/10.1109/TENCON.2019.8929517

Melkumova LE, Shatskikh SY (2017) Comparing ridge and LASSO estimators for data analysis. Procedia Eng 201:746–755

Mittal M, Sharma RK, Singh VP (2014) Validation of k-means and threshold based clustering method. Int J Advance Technol 5(2):153–160

Mittal M, Sharma RK, Singh VP (2015) Modified single pass clustering with variable threshold approach. Int J Innov Comput Inf Control 11(1):375–386

Mittal M, Balas VE, Goyal LM, Kumar R (Eds.) (2019) Big data processing using spark in cloud. Springer. https://doi.org/10.1007/978-981-13-05

Naheed N, Shaheen M, Khan SA, Alawairdhi M, Khan MA (2020) Importance of features selection, attributes selection, challenges and future directions for medical imaging data: a review. Comput Model Eng Sci 125(1):314–344

Qi Z, Wang T, Song G, Hu W, Li X, Zhang Z (2018) Deep air learning: interpolation, prediction, and feature analysis of fine-grained air quality. IEEE Trans Knowl Data Eng 30(12):2285–2297

Qian W, Yang Y (2013) Model selection via standard error adjusted adaptive lasso. Ann Inst Stat Math 65(2):295–318

Saeed F, Khan MA, Sharif M, Mittal M, Goyal LM, Roy S (2021) Deep neural network features fusion and selection based on PLS regression with an application for crops diseases classification. Appl Soft Comput 103:107164

Sethi JK, Mittal M (2019a) A new feature selection method based on machine learning technique for air quality dataset. J Stat Manage Syst 22(4):697–705

Sethi J, Mittal M (2019b) Ambient air quality estimation using supervised learning techniques. EAI Endorsed Trans Scalable Inf Syst 6(22). https://doi.org/10.4108/eai.13-7-2018.159406

Singh KP, Gupta S, Rai P (2013) Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmos Environ 80:426–437

Streets DG, Fu JS, Jang CJ, Hao J, He K, Tang X et al (2007) Air quality during the 2008 Beijing Olympic games. Atmos Environ 41(3):480–492

Tahir MAUH, Asghar S, Manzoor A, Noor MA (2019) A classification model for class imbalance dataset using genetic programming. IEEE Access 7:71013–71037

Tahir M B, Khan M A, Javed K, Kadry S, Zhang Y D, Akram T, Nazir M (2021) Recognition of apple leaf diseases using deep learning and variances-controlled features reduction Microprocessors and Microsystems, 104027. https://doi.org/10.1016/j.micpro.2021.104027

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol 58(1):267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Wang L, Wang Y, Chang Q (2016) Feature selection methods for big data bioinformatics: a survey from the search perspective. Methods 111:21–31

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

Wiedinmyer C, Yokelson RJ, Gullett BK (2014) Global emissions of trace gases, particulate matter, and hazardous air pollutants from open burning of domestic waste. Environ Sci Technol 48(16):9523–9530

World Health Organisation. (n.d.) https://www.who.int (accessed 20th January 2021)

Zhai B, Chen J (2017) Research on the forecasting of air quality index (AQI) based on FS-GA-BPNN: a case study of Beijing, China. Proceedings of the 14th ISCRAM Conference – Albi, France

Zhai B, Chen J (2018) Development of a stacked ensemble model for forecasting and analyzing daily average PM2. 5 concentrations in Beijing, China. Sci Total Environ 635:644–658

Zhang Y, Zhang R, Ma Q, Wang Y, Wang Q, Huang Z, Huang L (2020) A feature selection and multi-model fusion-based approach of predicting air quality. ISA Trans 100:210–220

Zheng Y, Liu F, Hsieh HP (2013) U-air: when urban air quality inference meets big data. In proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, KDD 2013 (pp. 1436-1444). Association for Computing Machinery

Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101(476):1418–1429