World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:12, 2018

Multi-Level Air Quality Classification in China Using Information Gain and Support Vector Machine

Bingchun Liu, Pei-Chann Chang, Natasha Huang, Dun Li

***Abstract*—**Machine Learning and Data Mining are the two important tools for extracting useful information and knowledge from large datasets. In machine learning, classification is a wildly used technique to predict qualitative variables and is generally preferred over regression from an operational point of view. Due to the enormous increase in air pollution in various countries especially China, Air Quality Classification has become one of the most important topics in air quality research and modelling. This study aims at introducing a hybrid classification model based on information theory and Support Vector Machine (SVM) using the air quality data of four cities in China namely Beijing, Guangzhou, Shanghai and Tianjin from Jan 1, 2014 to April 30, 2016. China's Ministry of Environmental Protection has classified the daily air quality into 6 levels namely Serious Pollution, Severe Pollution, Moderate Pollution, Light Pollution, Good and Excellent based on their respective Air Quality Index (AQI) values. Using the information theory, information gain (IG) is calculated and feature selection is done for both categorical features and continuous numeric features. Then SVM Machine Learning algorithm is implemented on the selected features with cross-validation. The final evaluation reveals that the IG and SVM hybrid model performs better than SVM (alone), Artificial Neural Network (ANN) and K-Nearest Neighbours (KNN) models in terms of accuracy as well as complexity.

***Keywords*—**Machine learning, air quality classification, air quality index, information gain, support vector machine, cross-validation.

I. Introduction

AIR Pollution has become one of the most serious environmental concerns for many countries. It serves as a hindrance to the social and economic development of any nation of the world. This is simply because of the introduction of the different harmful pollutants in the air that causes discomfort to the living species and damages our environment and the climate. These unwanted substances that are added to the atmosphere through industrial and manufacturing operations, burning of fossil fuels, automobiles emissions and some natural processes are called air pollutants. Some of the major air pollutants include PM2.5, PM10, SO2, CO, NO2 and O3 (China's Ministry of Environmental Protection). An abnormal amount of these pollutants can have harmful effect on the human health as well as the environment [1]. Some harmful effects on human health include respiratory problems, bronchitis, cough, asthma, lung cancer and cardiovascular disease. Some adverse effects on the environment include climate change, damaged vegetation, corrosion, acid rain and global warming [2]. Apart from the air pollutants, some meteorological variables are also highly correlated with the air quality [3].

Vehicles exhausts, industrial productions, coal burning, and construction sites dust are the key polluters contributing to 85%-90% of pollution woes (China's Ministry of Environmental Protection). This has a direct impact on the people's health. A report from the University of California showed that that around 1.6 million people in China die each year from heart, lung and stroke problems due to air pollution [4]. The cities that are industrialized and developed suffer the most from air pollution [5]. Vehicles exhausts are the main culprit for pollution in Beijing and Guangzhou whereas construction site dust, transport of polluted items and industrial production add to air pollution in coastal cities of Tianjin and Shanghai [6].

These drastic consequences and adverse effects of air pollution have turned the attention of the authorities, researchers and the general public towards the area of air quality. Hence there is an urgent need for modelling, planning and forecasting Air Quality. Prediction and Classification serve as two major components that provide the authorities with air quality information in advance in order to come up with the necessary measures soon enough for the well-being of the public. The qualitative air quality levels (serious pollution, severe pollution, moderate pollution, light pollution, good and excellent) are more practical from an operational point of view. Therefore, for the present study, we have used classification over regression to qualitatively predict the air quality levels in China.

II. Literature Review

Researchers in the past have worked on developing various mathematical and statistical tools to forecast air quality and take preventive measures to avoid any crisis. A lot of research has been done on developing regression models that give a quantitative prediction of air quality based on the AQI. AQI can be defined as an index that gives the daily estimate of air quality due to the various air pollutants and weather conditions. A high correlation was observed with metrological variables while forecasting AQI in many cities [3]. ANN technique was used to forecast AQI [7]. To improve the forecast results, a GA-ANN approach was developed [8]. In Spain, air quality was predicted by an SVM-based regression model that captured the main

Bingchun Liu is with the Research Institute of Circular Economy & Management School of Tianjin University of Technology, Tianjin, 300384 P.R. China (e-mail: tjutlbc@ tjut.edu.cn).

Pei-Chann Chang & Dun Li are with the Industrial Automation School, Zhuhai college of Beijing Institute of Technology, Zhuhai, Guangdong, 519088 P.R. China (corresponding author, e-mail: 3248068337@qq.com).

Natasha Huang is with the English Language Center, Sino-US School, Zhuhai College of Beijing Institute of Technology, Zhuhai, Guangdong, 519088 P.R. China (e-mail: yuhsin.huang@suc.bitzh.edu.cn).

insight of statistical learning theory in order to obtain a good prediction of the dependence among the main pollutants [9]. A principal component regression model was developed for Air Quality forecasting in Delhi [10]. Recently a PCA-neural network model was developed to forecast AQI in Delhi, India [11].

But a quantitative approach is not often the best for Air Quality forecast due to practical and operational reasons [12]. In the present study, a qualitative approach is used for Air Quality classification. In recent years many researchers have started focussing on air quality classification. An online forecasting technique based on Hadoop was developed using SVM to predict air quality [13]. The feasibility of applying SVM was examined and performance comparison was done using three kernels: linear, polynomial, and RBF [14]. An SVM predictive model was developed and the performance of the three kernels namely Gaussian, polynomial and spline were compared [15]. For predicting roadside fine particulate matter concentration level in Hong Kong Central, classification models were built based on ANN and SVM using R programming [16]. The feasibility of applying SVM to predict pollutant concentrations was also examined [17]. Also, different ANN models have been used to forecast concentration levels of different air pollutants for the city Perugia [18]. From the above literature, it is evident that SVM and ANN are the two important forecasting methods for classification in air quality research.

Despite its good classification accuracy, the SVM model has been criticized in the past for large computation time due to the use and re-computation of large scale kernel matrices [19]. SVM has also been found to be inefficient in dealing with large training sets and its requirement of retraining of each new training set [20]. For enhancing the classification results and improving the complexity of the model, it is important to select the important and significant input variables for Air Quality forecast. Input Variables or Feature selection is done using the information theory [21]. The information theory is wildly used for the construction of decision trees using ID3 and C4.5 algorithms [22]. But the use and application of the information theory have been left unexplored with the other efficient machine learning algorithms like SVM, ANN, KNN, etc.

The present study aims to use the air pollutant concentrations and weather conditions as input variables to predict Air quality (classified as Serious Pollution, Severe Pollution, Moderate Pollution, Light Pollution, Good and Excellent) based on IG and SVM hybrid model. The proposed model is compared to SVM (alone), Neural Network and KNN models to access its efficiency and also to justify the inclusion of IG with SVM machine learning algorithm. The objective of this study is not limited to improving the accuracy but also to reduce the number of input features thus improving the overall complexity of the model. R [23] is widely used open-source software environment for statistical analysis of data and visualising the results. We have chosen R Programming language as a tool to analyse the air quality data for four cities in China. Some well-known R Packages have been used for modelling and visualizing namely "nnet" package [24].

The structure of the paper is: Section III describes the methodologies used. Section IV describes the air quality data and data distribution for air quality classification in China. Section V gives the result analysis for selection of classification model, feature selection and IG and SVM hybrid model implementation. Finally, Section VI sums up the final conclusions of this study.

III. Methodology

A. Information Gain (IG)

The contribution of each feature towards the air quality classification differs. Some have a significant contribution whereas some are not significant enough to predict the air quality. Hence, we need to select only the most significant features and neglect the feature that does not contribute to the air quality classification. This can be achieved by ranking the features with IG.

Information entropy is a concept from the information theory [21]. Entropy is the degree of randomness hence the more uncertain or random the event is, the more information it will contain. Shannon Entropy is the most applied technique for calculating the IG. It is defined as the amount of information that an event provides. Shannon represented Entropy as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_b P(x_i) \quad (1)$$

where, X is a discrete random variable with values {x1… xn}, b is the base of the logarithm used, P(X) is the probability mass function of X and H(X) is the Shannon Entropy of X.

Here we have used the most common value of b, i.e. 2. In the case of P(xi) = 0 for some i, the value of the corresponding sum 0*logb(0) limits to 0.

In a classification problem, the IG measures the amount of information that can be characterised. The IG is calculated differently for the numeric/continuous and categorical features. For the categorical features: the gain information is the difference between the Shannon entropy of the entire set and the Shannon entropy of that feature. It is given by the equation:

$$Gain(S, F) = H(S) - H(S_F) \quad (2)$$

where, Gain(S, F) represents the gain information for feature F in set S. H(S) represents the total Shannon Entropy of set S. H(SF) represents the Shannon Entropy of feature F in set S.

For the continuous numeric features, the gain calculation is not completely the same. [22] presented the following steps to calculate gain for continuous numeric features: 1) Sort the continuous numeric values of the feature into ascending order. 2) Remove the values that are repeated. 3) Divide the unique remaining values into greater than and less than intervals and find the number of values that the interval contains and group them into their respective output classifications. 4) Finally, calculate the gains for each interval as we did for the categorical features and select the maximum value as the IG of that feature. After calculating the IG values for continuous and categorical features, we rank the features in descending order of their IG

values.

B. Support Vector Machine (SVM)

SVM are supervised machine learning algorithms to analyse data used for classification and regression analysis. SVM was first developed by [25]. It was originally developed for solving classification problems but later it was also applied in many other machine learning applications like image processing, categorising text, face recognition, time series analysis and regression analysis [15]. An SVM constructs hyperplanes that separate different classes. The optimal separation is achieved when the hyperplane has the largest functional margin [17]. The larger the margin the more accurate the classifier. Here we present the basic steps involved in SVM when the data are non-linearly separable.

Given there are n training sets {x1, x2, x3… xn} and {y1, y2, y3… yn}, the hyperplanes are represented by the equation

$$w \cdot x + b = 0 \tag{3}$$

that is parameterised by vector w and constant b. The distance between the hyperplane and the input points is simply given by the equation:

$$d(x) = \frac{|(w \cdot x_i + b)|}{||w||} \tag{4}$$

where, i varies from 1 to n.

For a larger functional margin and better accuracy, the distance needs to be maximised. But in order to maximise the distance we need to minimise the value of ||w||. Most of the times, the original data are not linearly separable hence the first task of SVM is to map the data into richer feature space where the data are separable. Hence, a kernel trick is useful here to map the old non-linearly separable data into a new higher dimensional space where the data is separable. Therefore, it is important to choose a kernel and its appropriate parameters. With the help of Lagrange multipliers, we need to minimise ||w|| and maximise the distance d (4).

$$\text{Maximize: } \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{5}$$

$$\text{Constraints: } \sum_{i=1}^{n} \alpha_i y_i = 0 \, , 0 \leq \alpha_i \leq C. \tag{6}$$

where, α is a vector which represents n Lagrange multipliers that are needed to be found, C represents the cost parameter and K represents the kernel function.

The three popular kernels for classification are linear, polynomial and radial kernels. Kernel selection is not an exact science and can be done by using trial and error. The Gaussian radial basis function kernel, or RBF kernel, is a popular kernel function used in various learning algorithms but most commonly in SVM classification. If prior information is not there about the data, then RBF kernel is generally used (Zhao and Hasan, 2013). The RBF kernel on two samples x and x', represented as feature vectors in some input space, is defined as

$$K(x, x') = \exp\left(-\frac{\left||x - x'|\right|^2}{2\sigma^2}\right)$$

or

$$K(x, x') = \exp\left(-\gamma \left||x - x'|\right|^2\right) \tag{7}$$

where, ||x-x'||2 may be recognized as the squared Euclidean distance between the two feature vectors x and x', σ is a free parameter and γ = 1/2 σ2.

C. IG and SVM Hybrid Model

The objective of the present study is to present a new model of SVM based on IG for air quality classification and forecasting. First, the AQI is categorised into different air quality levels. Then, the data are split into training and testing sets. The features are ranked based on their IG values. Then the model is developed based on SVM machine learning algorithm using the selected significant features on the training dataset. We then tune the data to get the best model for testing. Classification carried out on the testing dataset based on the developed model on the training set. Finally, the air quality levels are predicted, and confusion matrices are generated to visualise the results as well as calculate the testing accuracies.

IV. Air Quality Classification in China

A. Air Quality Data

TABLE I
FACTORS OF INPUT FEATURE: WEATHER

Weather	Factors
Partly Cloudy	0
Sunny	1
Rainy	2
Cloudy	3
Snow	4
Dust	5
Haze	6
Fog	7

TABLE II
FACTORS OF INPUT FEATURE: WIND DIRECTION

Wind Direction	Factors
North wind	0
North-east wind	1
east wind	2
South-east wind	3
south wind	4
North-west wind	5
west wind	6
South-west wind	7
No sustained wind	8

The data set used for the study of air quality classification for Beijing, Guangzhou, Shanghai and Tianjin is the live daily air quality data released by the China Environmental Monitoring Station and the meteorological data from the China Meteorological Administration for the time period January 1, 2014 to April 30, 2016. Air pollution levels are obtained from actual observation, issued by the China National

Environmental Monitoring Center. The features selected for the study includes PM2.5, PM10, SO_2, CO, NO_2, O_3, Maximum temperature, Minimum Temperature, weather, wind direction and wind power. The unit of measurement of air pollution features PM2.5, PM10, SO_2, CO, NO_2 and O_3 is mg/m_3. The features weather is classified as partly cloudy, sunny, rainy, cloudy, snow, dust, haze and fog. These are represented by values 0 to 7 respectively shown in Table I. Wind direction has been classified as north, northeast, east, south-east wind, southerly, north-west, west wind, south-west wind and no sustained wind nine types. These are represented by values 0 to 8 respectively as shown in Table II. Wind Power has been summarised into 5 levels: less than three, 3-4, 4-5, 5-6 and 6-7 grade as shown in Table III. Finally, Air quality is classified into six air pollution levels: serious pollution and severe pollution, moderate pollution, light pollution, good and excellent. This is in accordance with national air pollution levels depending on their AQI values as shown in Table IV.

TABLE III
FACTORS OF INPUT FEATURE: WIND POWER

Wind Power	Factors
<3 level	0
3-4 level	1
4-5level	2
5-6 level	3
6-7 level	4

TABLE IV
AIR QUALITY LEVELS (OUTPUT VARIABLE)

AQI	Air Quality Levels
0-50	Excellent
51-100	Good
101-150	Lightly Polluted
151-200	Moderately Polluted
201-300	Severe Pollution
300+	Serious Pollution

B. Data Distribution for Cross-Validation

Prediction in research is to discern the future; therefore, the development of our predictive models is evaluated by accuracy, reliability and credibility. We divide the available data into separate partitions, develop the models on one of the partitions and use the other partition to assess and even to refine the predictive model. The model development is completed on the training set, while the prediction is carried out on the testing set. For the present study, we split (3/4th) of the data into training and the rest (1/4th) into testing data and used 4-fold cross-validation technique in order to get accurate and credible results. The daily air quality dataset for Beijing, Guangzhou, Shanghai and Tianjin consists of 851 days of data starting from January 1, 2014 to April 30, 2016. For each city, the data are divided into four folds of training and testing datasets which are shown in Table V.

C. Selection of Classification Model: Comparison between SVM, ANN & KNN

There are various Soft-computing forecasting techniques available for the classification prediction in machine learning and artificial intelligence. Hence the first step of our experiment must be to select the best classification model based on accuracy testing. We are going to use SVM, ANN and KNN algorithm for model training and accuracy testing. All the 11 features are selected for the input variables and Air Quality levels is selected as an output variable as shown in Table VI.

TABLE V
DATA DISTRIBUTION FOR AIR QUALITY DATASETS OF BEIJING, GUANGZHOU, SHANGHAI AND TIANJIN

Sl. No.	Data (training/testing)	Duration	No. of data points	Total no. of data points
1	Train 1	01.01.2014 - 30.09.2015	638	851
	Test1	01.10.2015 - 30.04.2016	213	
2	Train 2	01.08.2014 - 30.04.2016	639	851
	Test 2	01.01.2014 - 31.07.2014	212	
3	Train 3	02.03.2015 - 31.07.2014 *	638	851
	Test 3	01.08.2014 - 01.03.2015	213	
4	Train 4	30.09.2015 - 01.03.2015**	639	851
	Test 4	02.03.2015 - 29.09.2015	212	

* 02.03.2015 - 31.07.2014 represents data from 2nd March 2015 to 30th April 2016 and from 1st January 2014 to 31st July 2014.

** 30.09.2015 - 01.03.2015 represents data from 30th September 2015 to 30th April 2016 and from 1st January 2014 to 1st March 2015.

TABLE VI
FEATURE CLASSIFICATION BASED ON NUMERIC / CATEGORICAL AND INPUT / OUTPUT

Sl. No.	Feature Name	Feature Type	Input / Output Variable
1	PM2.5	Numeric	Input Variable
2	PM10	Numeric	Input Variable
3	SO2	Numeric	Input Variable
4	CO	Numeric	Input Variable
5	NO2	Numeric	Input Variable
6	O3	Numeric	Input Variable
7	Max Temperature	Numeric	Input Variable
8	Min Temperature	Numeric	Input Variable
9	Weather	Categorical	Input Variable
10	Wind Direction	Categorical	Input Variable
11	Wind Power	Categorical	Input Variable
12	Air Quality Levels	Categorical	Output Variable

For developing the SVM model we have used the e1071 R Package [26] in R Programming Language. The selection of the kernel is done by using the tuning function in the e1071 R Package. For our dataset, it is observed that Gaussian Radial Basis Function Kernel (RBF) gives the best results. For the RBF kernel, Cost (C) and Gamma (γ) are the two important parameters involved (see (5)-(7)). C is defined as a regularization constant that highly influences the performance of the SVM on a dataset by controlling the trade-off between the errors and maximizing the distance between classes. The value of Gamma (γ) determines the lower bound for the RBF Kernel. The cost parameter can be adjusted to avoid overfitting. The process of choosing these parameters is called Hyperparameter optimization. Hyperparameter Optimization or model selection is a method to determine the best parameters to optimize the performance of the algorithm. Hyperparameter optimization also ensures that after tuning there is no problem

of overfitting the data. Generally, the quality of the tuning parameters can be improved by running k-fold cross-validation: the training data set is split into k groups of nearly equal size, then iteratively training the SVM using k-1 groups and make predictions on the group that remains. It is undesirable to find the exact values of C and γ as it would unnecessarily increase the complexity. Hence, we try to find the approximate values of the parameters by tuning the model on the training datasets with the gamma values varying from 10^-6: 10^-1 and cost varying from 10: 10^4. This tuning uses 10 fold cross validation sampling method to select the best parameters that correspond to maximum accuracy of the training data set. Finally, we used the best model parameters to make predictions on the testing data set.

TABLE VII
MEAN PERCENTAGE ACCURACY COMPARISON BASED ON SVM MODEL

SVM	Beijing	Guangzhou	Shanghai	Tianjin
train1 + test1	85.44601	91.58879	89.25234	86.4486
train2 + test2	81.13208	85.84906	90.09434	83.96226
train3 + test3	77.46479	90.14085	92.48826	86.85446
train4 + test4	80.66038	92.45283	88.67925	78.77358
Mean % Accuracy	**81.175815**	**90.0078825**	**90.1285475**	**84.009725**

TABLE VIII
MEAN PERCENTAGE ACCURACY COMPARISON BASED ON ANN MODEL

ANN	Beijing	Guangzhou	Shanghai	Tianjin
train1 + test1	81.22066	84.11215	83.17757	82.24299
train2 + test2	75.4717	84.43396	89.15094	82.07547
train3 + test3	73.23944	91.5493	84.03756	71.83099
train4 + test4	75	92.92453	85.84906	81.13208
Mean % Accuracy	**76.23295**	**88.254985**	**85.5537825**	**79.3203825**

TABLE IX
MEAN PERCENTAGE ACCURACY COMPARISON BASED ON KNN MODEL

KNN	Beijing	Guangzhou	Shanghai	Tianjin
train1 + test1	75.58685	88.31776	75.23364	76.63551
train2 + test2	70.75472	85.37736	82.07547	70.75472
train3 + test3	71.3615	88.26291	82.62911	75.58685
train4 + test4	70.75472	90.09434	83.01887	70.75472
Mean % Accuracy	**72.1144475**	**88.0130925**	**80.7392725**	**73.43295**

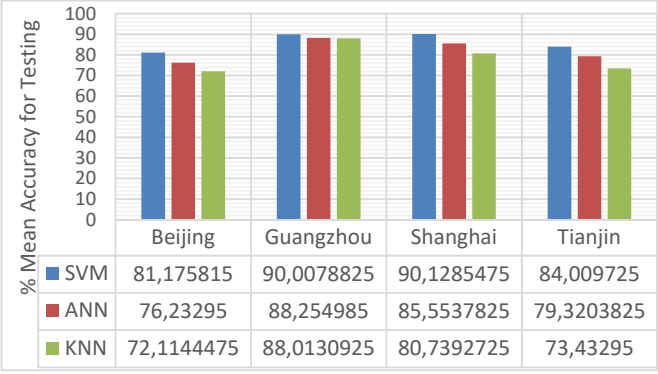


Fig. 1 Mean Percentage Accuracy Comparison of SVM, ANN and KNN models for Beijing, Guangzhou, Shanghai and Tianjin

For developing the ANN model, we use the "nnet" R Package [24]. The main parameters are number of hidden layers and maximum number of iterations. Keeping into account the complexity and the testing accuracy, the number of hidden layers were taken to be 12 and the maximum number of iterations were taken to be 500. "kknn" R Package was used for KNN model and the important parameters include number of neighbours considered (k) and the Minkowski distance (d). Again considering the complexity and the accuracy we have selected k=7 and d=1. Furthermore, we use 4-fold cross-validation technique to examine the accuracy of our SVM, ANN and KNN models. The choice for the 4-fold cross-validation estimator is justified for its lower variance than a single set estimator. If a single set is adopted, where 75% of data is used for training and 25% is used for testing, the test set is considerably small, leading to a large variation in the performance estimate on the basis of different partitions of the data. 4-fold validation, however, greatly reduces the variance by averaging four different partitions, and therefore the performance estimate becomes less sensitive to different data partitions. All steps of the model fitting procedure (model selection, feature selection etc.) is performed independently in each fold of the cross-validation procedure so that the resulting performance estimate is not biased. Tables VII-IX show the 4-fold cross validation accuracy testing results for Beijing, Guangzhou, Shanghai and Tianjin for SVM, ANN and KNN models.

The comparison of mean model testing accuracy in Fig. 1 clearly indicates that the SVM model performs better than ANN and KNN models for all the cities. Thus, we select SVM as the classification model for the prediction of Air Quality for all the 4 cities in China.

D.Feature Selection Using Information Theory

TABLE X
FEATURE RANKING BASED ON IG VALUES FOR BEIJING

Sl. No.	Feature Name	IG	Rank
1	PM2.5	0.791580274	1
2	PM10	0.59728128	3
3	SO_2	0.165815442	8
4	CO	0.450388179	6
5	NO_2	0.264369962	7
6	O_3	0.106185425	9
7	Max Temperature	0.053421471	10
8	Min Temperature	0.05037903	11
9	Weather	0.627976719	2
10	Wind Direction	0.488244197	4
11	Wind Power	0.479676359	5

Next, in order to improve the testing accuracy as well as complexity of our SVM model, we will perform variable feature sensitivity analysis. First, we need to rank the input variables based on their IG values. Table VI shows that presently the model has 11 input variables and 1 output variable. In order to calculate the individual input variable contributions to Air Quality levels, we find their IG. We have ranked the features according to their individual contribution towards the Air Quality levels for training sets for each city, i.e. a total of 16 gain value sets for four cities (Beijing, Guangzhou, Shanghai and Tianjin). Since for each city, the four training sets

show the same feature ranking but slightly different IG values, we have selected the "train1" gain set for each city. The IG results for Beijing, Guangzhou, Shanghai and Tianjin are shown in Tables X-XIII, respectively.

TABLE XI
FEATURE RANKING BASED ON IG VALUES FOR GUANGZHOU

Sl. No.	Feature Name	IG	Rank
1	PM2.5	0.680815	2
2	PM10	0.7426638	1
3	SO_2	0.2762303	6
4	CO	0.165248	8
5	NO_2	0.2376274	7
6	O_3	0.118727	9
7	Max Temperature	0.05999035	11
8	Min Temperature	0.1012421	10
9	Weather	0.628029	3
10	Wind Direction	0.5807055	4
11	Wind Power	0.5349926	5

TABLE XII
FEATURE RANKING BASED ON IG VALUES FOR SHANGHAI

Sl. No.	Feature Name	IG	Rank
1	PM2.5	0.8021741	1
2	PM10	0.5501355	3
3	SO_2	0.1658154	8
4	CO	0.4503882	6
5	NO_2	0.26437	7
6	O_3	0.1061854	9
7	Max Temperature	0.05342147	10
8	Min Temperature	0.05037903	11
9	Weather	0.6279767	2
10	Wind Direction	0.4882442	4
11	Wind Power	0.4796764	5

TABLE XIII
FEATURE RANKING BASED ON IG VALUES FOR TIANJIN

Sl. No.	Feature Name	IG	Rank
1	PM2.5	0.7685662	1
2	PM10	0.6102263	2
3	SO_2	0.15698	8
4	CO	0.2404346	7
5	NO_2	0.2812715	6
6	O_3	0.1010213	10
7	Max Temperature	0.09354128	11
8	Min Temperature	0.110348	9
9	Weather	0.4072525	3
10	Wind Direction	0.3673369	4
11	Wind Power	0.334871	5

E. IG + SVM Hybrid Classification Model for Air Quality Classification

This study applies the IG model and the SVM model together in order to predict Air Quality Levels. Although the SVM model produces satisfactory results for Air Quality Levels Prediction Classification problem but in order to achieve better results we introduce a new hybrid prediction model, IG and SVM. Fig. 2 shows the framework of this model. F1, F2…..F11 are the input variables. Firstly, the individual IG are calculated for all input variable using the information theory. Then n selected output variables (IG_1, IG_2,….,IG_n) from the IG n model serve as input variables for the SVM model in order to predict the Air Quality Levels.

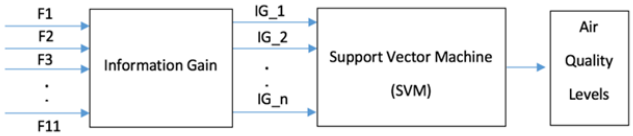
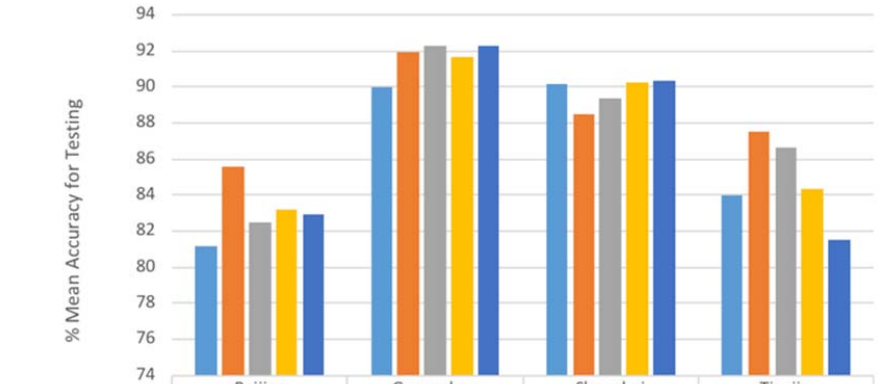


Fig. 2 Framework of IG + SVM Hybrid Classification Model for Air Quality Classification

For variable feature sensitivity analysis, we need to select a variable number of input features for the SVM model. For our study, we select 4, 5, 7 and 9 top ranking features based on their IG values from Tables X-XIII. We first select the 4 top features based on their IG value for each city. These 4 selected features are denoted by IG_4features. Now we use IG_4features as an input variable for our SVM model. This model is now trained on all the training sets for each city and best model parameters are selected respectively for testing. This is done on similar terms for IG_5features, IG_7features and IG_9features. Fig. 3 compares the Mean Percentage Testing Accuracy based on 4-fold cross-validation for each city for variable input variables.

Fig. 3 clearly indicates that for all the 4 cities, IG and SVM Hybrid model is performing better than SVM alone. For Beijing and Tianjin, IG_4features+SVM is the best prediction model with 85.525% and 87.535% accurate prediction of the unseen testing data respectively. Whereas for Guangzhou and Shanghai, IG_9features+SVM comes out to be the best prediction model with 92.245% and 90.367% accurate prediction of the unseen testing data respectively. It is also important to note that for all the cities the model, IG_7features+SVM is performing better than SVM alone.

For multivariable classification, confusion matrix is used to visualize the actual and predicted values of each class. Each diagonal element represents the number of exact matches (correct predictions) for each class whereas the non-diagonal elements represent mismatches (incorrect predictions). The sum of the diagonal elements in the confusion matrix determines the total number of correct predictions of the entire testing data set. Testing Accuracy for classification is calculated by dividing the total number of correct prediction by the total number of observations in the testing set. Figs. 4-7 represent confusion matrices that show the comparison between the best IG+SVM model and the SVM model for one set of training and testing data for Beijing, Guangzhou, Shanghai and Tianjin respectively.



	Beijing	Guangzhou	Shanghai	Tianjin
SVM	81.175815	90.0078825	90.1285475	84.009725
IG_4features+SVM	85.52518	91.877535	88.49195	87.535875
IG_5features+SVM	82.4707675	92.24128	89.314095	86.5941325
IG_7features+SVM	83.1783175	91.65386	90.25309	84.3668125
IG_9features+SVM	82.9374825	92.245125	90.367155	81.5449025

Fig. 3 Mean Percentage Accuracy Comparison of variable input feature SVM models for Beijing, Guangzhou, Shanghai and Tianjin

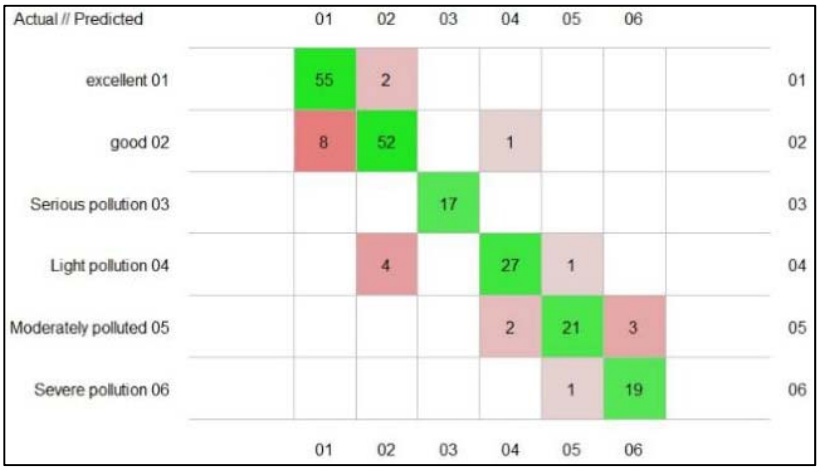


Fig. 4 (a) IG_4features+SVM on train1+test1 for Beijing

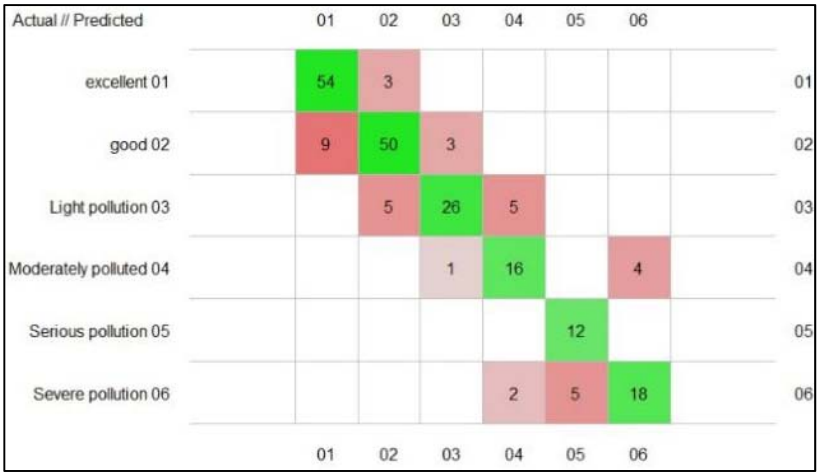


Fig. 4 (b) SVM on train1+test1 for Beijing

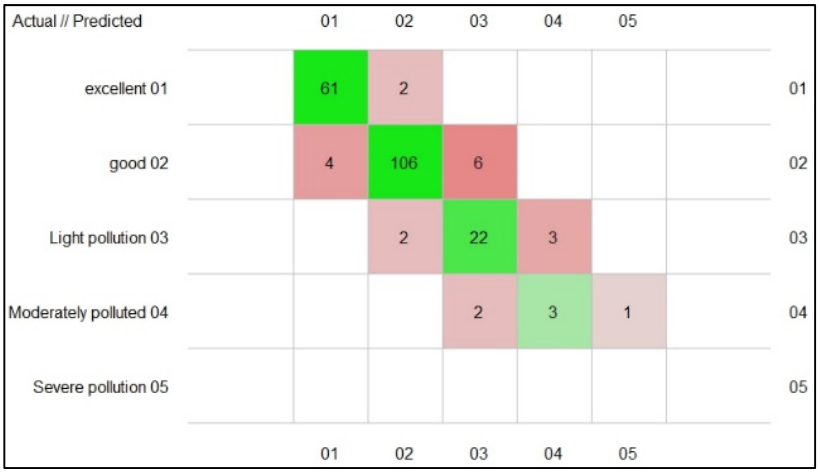


Fig. 5 (a) IG_9features+SVM on train2+test2 for Guangzhou

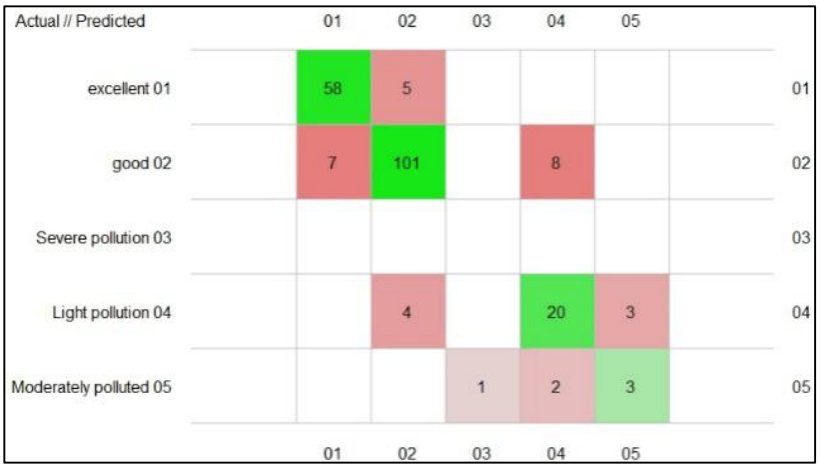


Fig. 5 (b) SVM on train2+test2 for Guangzhou

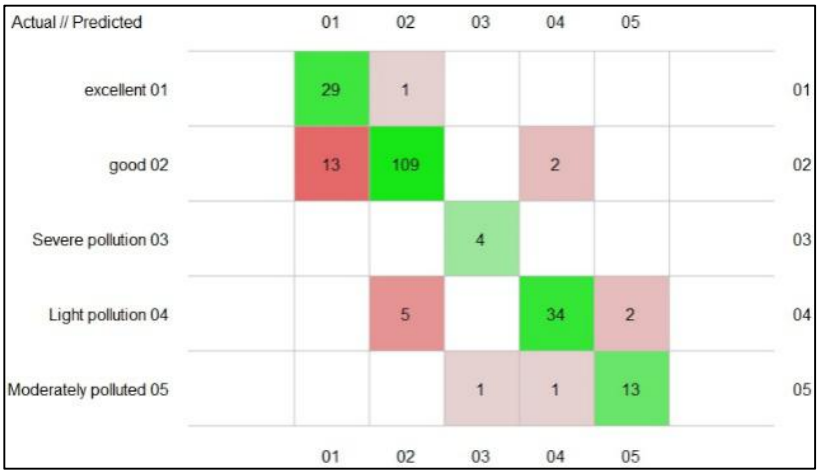


Fig. 6 (a) IG_9features+SVM on train1+test1 for Shanghai

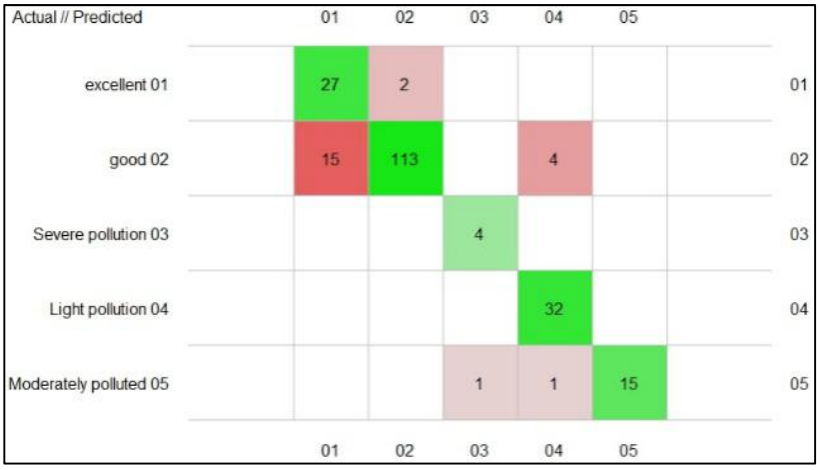


Fig. 6 (b) SVM on train1+test1 for Shanghai

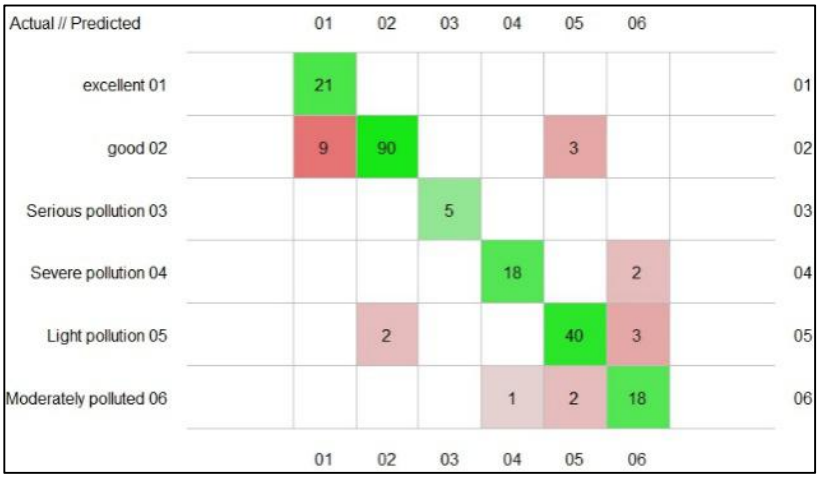


Fig. 7 (a) IG_4features+SVM on train1+test1 for Tianjin

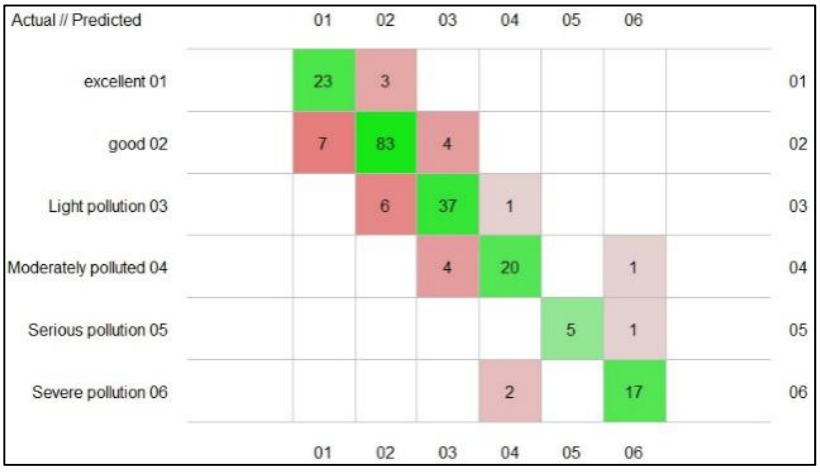


Fig. 7 (b) SVM on train1+test1 for Tianjin

V. Conclusion

The IG and SVM Hybrid Model presented in this paper has enabled us to achieve better predictions of Air Quality levels in four cities of China. After analysing the results of the experiment, we obtained the following conclusions. Firstly, the SVM model with RBF Kernel performs better for classifying non-linear data than ANN and KNN models for all the cities hence SVM Model is selected as the classification model for the prediction of Air Quality for the four cities in China. Secondly, for cities Beijing and Tianjin, “PM2.5”, “PM10”,

"weather" and "wind directions" are the four important and more relevant input features whereas for cities Guangzhou and Shanghai "PM2.5", "PM10", "weather", "wind directions", "Wind Power", "CO", "NO_2", "SO_2" and "O_3" are the nine most important and relevant input features out of the total 11 features that have a greater impact on prediction of air quality levels. Thirdly, IG and SVM Hybrid Model give better accuracy for all the cities when compared to SVM alone. Also, for all the cities the model, IG_7features+SVM gave better classification results than SVM alone. Also, the proposed model reduces the number of input variables thus reducing the complexity of the model.

This paper attempts to achieve the goal of more accurate predictions of air quality in classification mode and the goal of reducing complexity of previous models. The proposed techniques can be further extended to other applications in areas of classifications and machine learning. It is hoped that this research makes contribution to forecasting methods and machine learning techniques.

ACKNOWLEDGMENTS

The authors would like to appreciate the funding support from the National Natural Science Foundation of China (71503180).

REFERENCES

[1] B. R. Gurjar, T. M. Butler, M. G. Lawrence, J. Lelieveld. Evaluation of emissions and air quality in megacities. Atmospheric Environment 42 (2008) 1593–1606.

[2] Niharika, Venkatadri M, Padma S. Rao. A survey on Air Quality forecasting Techniques. International Journal of Computer Science and Information Technologies, Vol. 5 (1) (2014) 103-107.

[3] Euro Cogliani. Air pollution forecast in cities by an air pollution index highly correlated with meteorological variables. Atmospheric Environment 35 (2001) 2871-2877.

[4] Robert A. Rohde, Richard A. Muller. Air Pollution in China: Mapping of Concentrations and Sources. PLoS ONE 10(8): e0135749 (2015).

[5] Guleda Onkal-Engin, Ibrahim Demir, Halil Hiz. Assessment of urban air quality in Istanbul using fuzzy synthetic evaluation. Atmospheric Environment 38 (2004) 3809–3815.

[6] Chak K. Chan, Xiaohong Yao. Air pollution in mega cities in China. Atmospheric Environment 42 (2008) 1–42.

[7] Dahe Jiang, Yang Zhang, Xiang Hu, Yun Zeng, Jianguo Tan, Demin Shao. Progress in developing an ANN model for air pollution index forecast. Atmospheric Environment 38 (2004) 7055–7064.

[8] Hong Zhao, Jie Zhang, Kai Wang, Zhi peng Bai, Aixie Liu. A GA-ANN Model for Air Quality Predicting. Computer Symposium (ICS), International (2010) 693 – 699.

[9] A. Suárez Sánchez, P. J. García Nieto, P. Riesgo Fernández, J. J. del Coz Díaz, F. J. Iglesias-Rodríguez. Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). Mathematical and Computer Modelling 54 (2011) 1453 –1466.

[10] Anikender Kumar, Pramila Goyal. Forecasting of air quality in Delhi using principal component regression technique. Atmospheric Pollution Research 2 (2011) 436 – 444.

[11] Anikender Kumar, Piyush Goyal. Forecasting of Air Quality Index in Delhi Using Neural Network Based on Principal Component Analysis. Pure and Applied Geophysics 170 (4) (2013) 711-722.

[12] Ioannis N. Athanasiadis, Kostas D. Karatzas, Pericles A. Mitkas. Classification techniques for air quality forecasting. Fifth ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence, 17th European Conference on Artificial Intelligence, Riva del Garda, Italy, (2006).

[13] Z. Ghaemia, M. Farnaghi, A. Alimohammadi. Hadoop-based Distribution System for Online Prediction of Air Pollution based on Support Vector Machine. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-1/W5, 2015. International Conference on Sensors & Models in Remote Sensing & Photogrammetry, Kish Island, Iran. (2015).

[14] S. Bedoui, S. Gomri, H. Samet, A. Kachouri. A prediction distribution of atmospheric pollutants using support vector machines, discriminant analysis and mapping tools (Case study: Tunisia). Pollution, 2(1) (2016) 11-23.

[15] Artemio Sotomayor-Olmedo, Marco A. Aceves-Fernández, Efrén Gorrostieta-Hurtado, Carlos Pedraza-Ortega, Juan M. Ramos-Arreguín, J. Emilio Vargas-Soto. Forecast Urban Air Pollution in Mexico City by Using Support Vector Machines: A Kernel Performance Approach. International Journal of Intelligence Science, 3 (2013) 126-135.

[16] Yin Zhao, Yahya Abu Hasan. Machine learning algorithms for predicting roadside fine particulate matter concentration level in Hong Kong Central. Computational Ecology and Software, 3(3) (2013) 61-73.

[17] Weizhen Lu, Wenjian Wang, A. Y. T. Leung, Siu-Ming Lo, R. K. K. Yuen, Zongben Xu, Huiyuan Fan. Air Pollutant Parameter Forecasting Using Support Vector Machines. IJCNN, (Volume: 1), (2002).

[18] P. Viotti, G. Liuti, P. Di Genova. Atmospheric urban pollution: applications of an artificial neural network (ANN) to the city of Perugia. Ecological Modelling 148 (2002) 27–46.

[19] Antoine Bordes, Seyda Ertekin, Jason Weston, Leon Bottou. Fast Kernel Classifiers with Online and Active Learning. Journal of Machine Learning Research 6 (2005) 1579–1619.

[20] Wenjian Wang, Changqian Men, Weizhen Lu. Online prediction model based on support vector machine. Neurocomputing 71 (2008) 550–558.

[21] C. E. Shannon. A Mathematical Theory of Communication. The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, (1948).

[22] Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri, Mohammed Erritali. A comparative study of decision tree ID3 and C4.5. IJACSA, (2014).

[23] Ross Ihaka, Robert Gentleman. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics, Volume 5, Issue 3, (1996), 299-314.

[24] Brian Ripley, William Venables. Package "nnet": Feed-Forward Neural Networks and Multinomial Log-Linear Models. Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0, (2002).

[25] Vladimir N. Vapnik. An Overview of Statistical Learning Theory. IEEE Transactions of Neural Networks, Vol. 10, No. 5, (1999).

[26] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch. Package "e1071". Misc. functions of the Department of Statistics (e1071), TU Wien. The comprehensive R archive network (2012).