Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

| p-ISSN: 0972-6268 e-ISSN: 2395-3454 | Vol. 18 | No. 3 | pp. 697-708 | 2019 |
|---|---|---|---|---|

Original Research Paper | Open Access

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

**Bingchun Liu*†, Hui Wang*, Arihant Binaykia**, Chuanchuan Fu* and Bingpeng Xiang***
*Research Institute of Circular Economy, Tianjin University of Technology, Tianjin, P.R. China
**Department of Industrial and Systems Engineering, Indian Institute of Technology, Kharagpur, India
†Corresponding author: Bingchun Liu

**Nat. Env. & Poll. Tech.**
**Website: www.neptjournal.com**

*Received:* 26-12-2018
*Accepted:* 27-02-2019

**Key Words:**
Machine learning
Air quality classification
Air quality index
Information gain
Support vector machine
Cross-validation

## 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 widely 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 new hybrid classification model based on information theory and support vector machine (SVM) using the air quality data of 4 cities in China namely Beijing, Guangzhou, Shanghai and Tianjin from January 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.

## 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, automobile emissions and some natural processes are called air pollutants. Some of the major air pollutants include $PM_{2.5}$, $PM_{10}$, $SO_2$, CO, $NO_2$ and $O_3$ (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 (Gurjar et al. 2008). Some harmful effects on human health include respiratory problems, bronchitis, cough, asthma, lung cancer and cardiovascular diseases, while the adverse effects on the environment include climate change, damaged vegetation, corrosion, acid rain and global warming. Apart from the air pollutants, some meteorological variables are also highly correlated with the air quality (Cogliani 2001).

Vehicle exhausts, industrial productions, coal burning and construction site 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 around 1.6 million people in China die each year from heart, lung and stroke problems due to air pollution (Robert & Richard 2015). The cities that are industrialized and developed suffer the most from air pollution (Guleda et al. 2004). Vehicle 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 (Chan & Yao 2008).

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 op-

erational point of view. Therefore, for the present study, we have used classification over regression to qualitatively predict the air quality levels in China.

## PAST STUDIES

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 air quality index (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 meteorological variables while forecasting AQI in many cities. Artificial neural network technique was used to forecast AQI (Jiang et al. 2004). To improve the forecast results, a GA-ANN approach was developed (Zhao et al. 2010). In Spain, air quality was predicted by an SVM-based regression model that captured the main insight of statistical learning theory in order to obtain a good prediction of the dependence among the main pollutants (Sánchez et al. 2011). A principal component regression model was developed for air quality forecasting in Delhi (Kumar & Goyal 2011). Recently a PCA-neural network model was developed to forecast AQI in Delhi, India (Kumar & Goyal 2013).

But a quantitative approach is not often the best for air quality forecast due to practical and operational reasons (Athanasiadis et al. 2006). In the present study, a qualitative approach is used for air quality classification. In recent years many researchers have started focusing on air quality classification. An online forecasting technique based on Hadoop was developed using SVM to predict air quality (Ghaemia et al. 2015). The feasibility of applying SVM was examined and performance comparison was done using 3 kernels: linear, polynomial and RBF (Bedoui et al. 2016). Different soft computing forecasting techniques have been presented to give a comprehensive review of air quality forecasting. An SVM predictive model was developed and the performance of the three kernels namely Gaussian, Polynomial and Spline were compared (Artemio et al. 2013). For predicting roadside fine particulate matter concentration level in Hong Kong Central, classification models were built based on artificial neural network (ANN) and SVM using R programming (Zhao & Yahya 2013). The feasibility of applying SVM to predict pollutant concentrations was also examined. Also, different ANN models have been used to forecast concentration levels of different air pollutants for the city Perugia (Viotti et al. 2002). 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 (Bordes et al. 2005). SVM has also been found to be inefficient in dealing with large training sets and its requirement of retraining of each new training set (Wang et al. 2008). 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 (Shannon 1948). The information theory is widely used for the construction of decision trees using ID3 and C4.5 algorithms (Hssina et al. 2014). 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 information gain (IG) and support vector machine (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 information gain 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 (Ross & Robert 1996) 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 4 cities in China. Some well-known R packages have been used for modelling and visualizing namely "nnet" package (Brian & William 2002), "e1071" package (Meyer et al. 2012), "kknn" package (Klaus & Klaus 2004) and "mlearning" package (Grosjean & Denis 2012).

## MATERIALS AND METHODS

### 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 information gain.

Information entropy is a concept from the information theory. 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 information gain. 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 $\{x_1 \ldots x_n\}$
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 $0*\log_b(0)$ limits to 0.

In a classification problem, the information gain measures the amount of information that can be characterised. The information gain 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.
$H(S_F)$ represents the Shannon entropy of feature F in set S.

For the continuous numeric features, the gain calculation is not completely the same. Hssina (2014) 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 information gain of that feature. After calculating the information gain (IG) values for continuous and categorical features we rank the features in descending order of their IG values.

**Support Vector Machine (SVM)**

Support Vector Machines (SVMs) are supervised machine learning algorithms to analyse data used for classification and regression analysis. SVM was first developed by Vapnik (1999). 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 (García Nieto et al. 2013). An SVM constructs hyperplanes that separate different classes. The optimal separation is achieved when the hyperplane has the largest functional margin. The larger the margin the more accurate the classifier. Here, we present the basic steps involved in SVM when the data is non-linearly separable.

Given there are n training sets $\{x_1, x_2, x_3 \ldots x_n\}$ and $\{y_1, y_2, y_3 \ldots y_n\}$

The hyperplanes are represented by the equation

$$w \cdot x + b = 0 \quad ...(3)$$

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

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

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 are separable. Therefore, it is important to choose a kernel and its appropriate parameters (Nieto et al. 2013). With the help of Lagrange multipliers we need to minimise ‖w‖ and maximise the distance d (Eq. 4).

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

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

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

The 3 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 & Yahya 2013). The RBF kernel on two

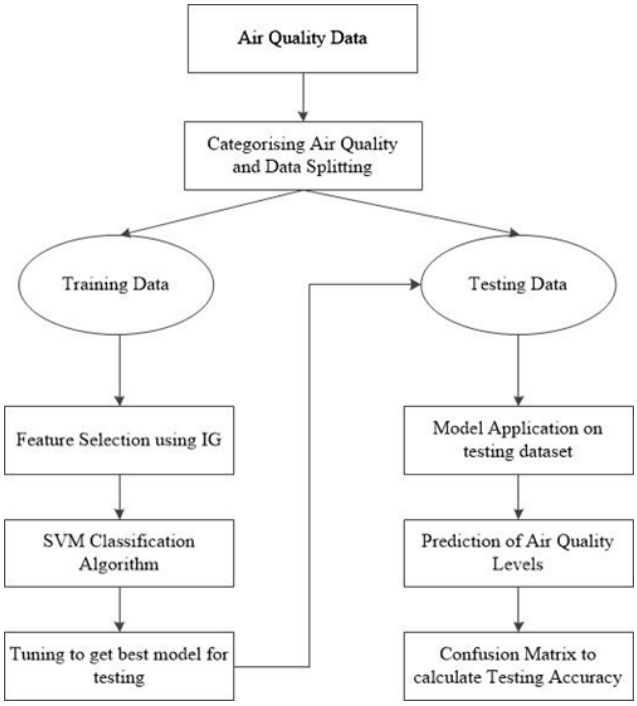



Fig. 1: Flowchart representation of IG and SVM Hybrid Model.

samples x and x', represented as feature vectors in some input space, is defined as:

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

$$K(x,x') = \exp\left(-\gamma\left|\left|x-x'\right|\right|^2\right) \quad ...(7)$$

Where, $\|x\text{-}x'\|^2$ may be recognized as the squared Euclidean distance between the two feature vectors x and x', $\sigma$ is a free parameter and $\gamma = 1/2\ s^2$.

### IG and SVM Hybrid Model

The objective of the present study is to present a new model of support vector machine based on the information gain for air quality classification and forecasting. The structured flowchart representation of the proposed IG and SVM Hybrid model is clearly shown in Fig. 1.

First, the air quality index (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 information gain 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.

Table 1: 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 2: 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 |

Table 3: 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 4: 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 |

## AIR QUALITY CLASSIFICATION IN CHINA

### Air Quality Data

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 $PM_{2.5}$, $PM_{10}$, $SO_2$, CO, $NO_2$, $O_3$, maximum temperature, minimum temperature, weather, wind direction and wind power. The unit of measurement of air pollution features $PM_{2.5}$, $PM_{10}$, $SO_2$, CO, $NO_2$ and $O_3$ is $mg/m^3$. The weather feature is classified as partly cloudy, sunny, rainy, cloudy, snow, dust, haze and fog. These are represented by values 0 to 7 respectively given in Table 1. Wind direction has been classified into nine types as north, northeast, east, south-east wind, southerly, north-west, west wind, south-west wind and no sustained wind. These are represented by values 0 to 8 respectively as given in Table 2. Wind power has been summarised into 5 levels: less than three, 3-4, 4-5, 5-6 and 6-7 grade as given in Table 3. 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 presented in Table 4.

### Data Distribution for Cross-Validation

Prediction leads us into unknown territory. During the development of a predictive model, we must assess its accuracy, reliability and credibility. Hence, it is very important to divide the available data into separate partitions, developing our models on one of these partitions and using the other for predictive model assessment and possibly model refinement. The model development is done on the training set and the prediction is carried 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 4-folds of training and testing datasets which are given in Table 5.

## RESULT ANALYSIS

### 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 is to select the best classification model based on accuracy testing. We have used Support Vector Machine (SVM), Artificial Neural Network (ANN) and K-Nearest Neighbours algorithm (KNN) for model training and accuracy testing. All the 11 features are selected for the input variables and air quality level is selected as an output variable as given in Table 6.

For developing the SVM model we have used the e1071 R Package in R programming language. The selection of the kernel is done by using the tuning function in the e1071 R Package. For our data set, it is observed that the Gaussian Radial Basis Function Kernel (RBF) gives the best results. For the RBF kernel, cost (C) and Gamma ($\gamma$) are the two important parameters involved (Eqs. 5, 6 and 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 (Yeganeh et al. 2012). The value of Gamma ($\gamma$) 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 $\gamma$ as it would unnecessarily increase the complexity. Hence, we tried 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.

For developing the ANN model we used the "nnet" R Package. 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. The "kknn" R Package was used for KNN model and the important parameters include number of neighbours considered (k) and the Minkowski distance (d). Again understanding the complexity and the accuracy we have selected k=7 and d=1.

Further, we used 4-fold cross-validation technique for examining the accuracy of our SVM, ANN and KNN models. The main reason is that the 4-fold cross-validation estimator has a lower variance than a single set estimator. If we take a single set, where 75% of data are used for training and

Table 5: 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 6: 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 | $SO_2$ | Numeric | Input Variable |
| 4 | CO | Numeric | Input Variable |
| 5 | $NO_2$ | Numeric | Input Variable |
| 6 | $O_3$ | 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 |

25% used for testing, the test set is very small, hence, there will be a lot of variation in the performance estimate for different samples or different partitions of the data to form training and testing sets. 4-fold validation reduces this variance by averaging over 4 different partitions, so the performance estimate is less sensitive to the partitioning of the data. All steps of the model fitting procedure (model selection, feature selection etc.) are performed independently in each fold of the cross-validation procedure so that the resulting performance estimate is not biased. The Tables 7, 8 and 9 give 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. 2 clearly indicates that the SVM model performs better than ANN and KNN models for all the cities. Thus, we selected SVM as the classification model for the prediction of air quality for all the 4 cities in China.

**Feature Selection Using Information Theory**

Next, in order to improve the testing accuracy as well as complexity of our SVM model, we performed the variable feature sensitivity analysis. First, we need to rank the input variables based on their information gain values. Table 6 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 information gains. 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 4 cities (Beijing, Guangzhou, Shanghai and Tianjin). Since for each city, the 4 training sets show the same feature ranking but slightly different information gain values, we have selected the "train1" gain set for each city. The information gain (IG) results for Beijing, Guangzhou, Shanghai and Tianjin are given in Table 10, Table 11, Table 12 and Table 13 respectively.

**IG + SVM Hybrid Classification Model for Air Quality Classification**

This study applies the Information Gain model and the SVM model together in order to predict the 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 introduced a new hybrid prediction model, information gain and support vector ma-

Table 7: Mean percentage accuracy comparison based on support vector machine 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 8: Mean percentage accuracy comparison based on artificial neural network 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 9: Mean percentage accuracy comparison based on k-nearest neighbors model.

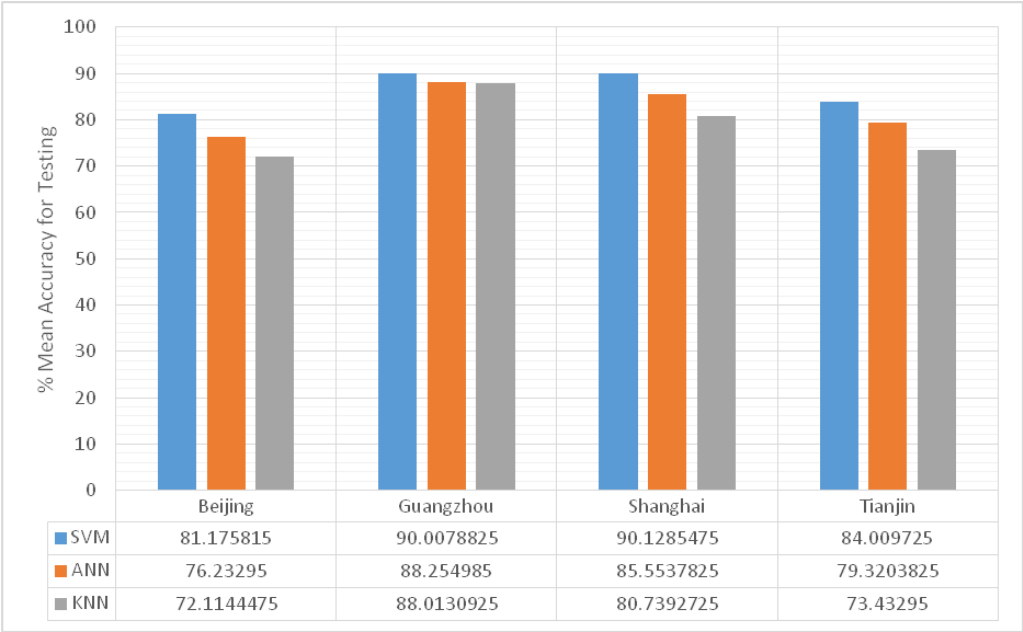
| KNN | Beijing | Guangzhou | Shanghai | Tianjin |
|---|---|---|---|---|
| train1 + test1 | 75.58685 | 88.31776 | 75.23364 | 76.63551 |
| train2 + test2 | 70.75472 | 85.37736 | 82.07547 | 70.75472 |
| train3 + test3 | 71.3615 | 88.26291 | 82.62911 | 75.58685 |
| train4 + test4 | 70.75472 | 90.09434 |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

chine. Fig. 3 shows the framework of this model. F1, F2…..F11 are the input variables. Firstly, the individual information gains are calculated for all input variable using the information theory. Then n selected output variables (IG_1, IG_2,….,IG_n) from the information gain model serve as input variables for the SVM model in order to predict the air quality levels.

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 information gain values from Tables 10, 11, 12 and 13 respectively. We first select the 4 top features based on their information gain value for each city. These 4 selected features are denoted by IG_4 features. Now we use IG_4 features 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_5 features, IG_7 features and IG_9 features. Fig. 4 compares the mean percentage testing accuracy based on 4-fold cross-validation for each city for variable input variables.

Fig. 4 clearly indicates that for all the 4 cities, Information Gain and SVM Hybrid model is performing better than SVM alone. For Beijing and Tianjin, IG_4 features+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_9 features+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_7 features+SVM is performing better than SVM alone.

For multivariable classification, confusion matrix is used to visualise 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. 5, 6, 7 and 8 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 |
| ANN | 76.23295 | 88.254985 | 85.5537825 | 79.3203825 |
| KNN | 72.1144475 | 88.0130925 | 80.7392725 | 73.43295 |

Fig. 2: Mean percentage accuracy comparison of SVM, ANN and KNN models for Beijing, Guangzhou, Shanghai and Tianjin.

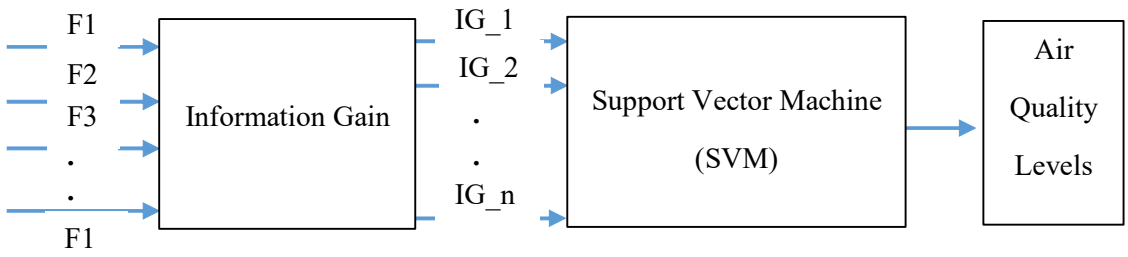



Fig. 3: Framework of IG + SVM hybrid classification model for air quality classification.

## CONCLUSION

The Information Gain and Support Vector Machine Hybrid Model presented in this paper has enabled us to achieve better predictions of air quality levels in 4 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 was selected as the classification model for the prediction of air quality for the 4 cities in China. Secondly, for cities Beijing and Tianjin, $PM_{2.5}$, $PM_{10}$, weather and wind directions are the 4 important and more relevant input features, whereas for cities Guangzhou and Shanghai $PM_{2.5}$, $PM_{10}$, weather, wind directions, wind power, CO, $NO_2$, $SO_2$ and $O_3$ are the 9 most important and relevant input features out of the total 11 features that have a greater impact on prediction of air quality levels. Thirdly, Information Gain and SVM Hybrid model gave better accuracy for all the cities when compared to SVM alone. Also for all the cities, the model IG_7 features+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 opens the doors for further research and exploration of different forecasting methods and machine learning techniques in order to achieve better accuracy of air quality prediction in classification mode and also to reduce the complexity of the previous models.

## ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China [grant number 71503180]; Major

Table 11: Feature ranking based on information gain values for Guangzhou.

| Sl. No. | Feature Name | Information Gain | 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 12: Feature ranking based on information gain values for Shanghai.

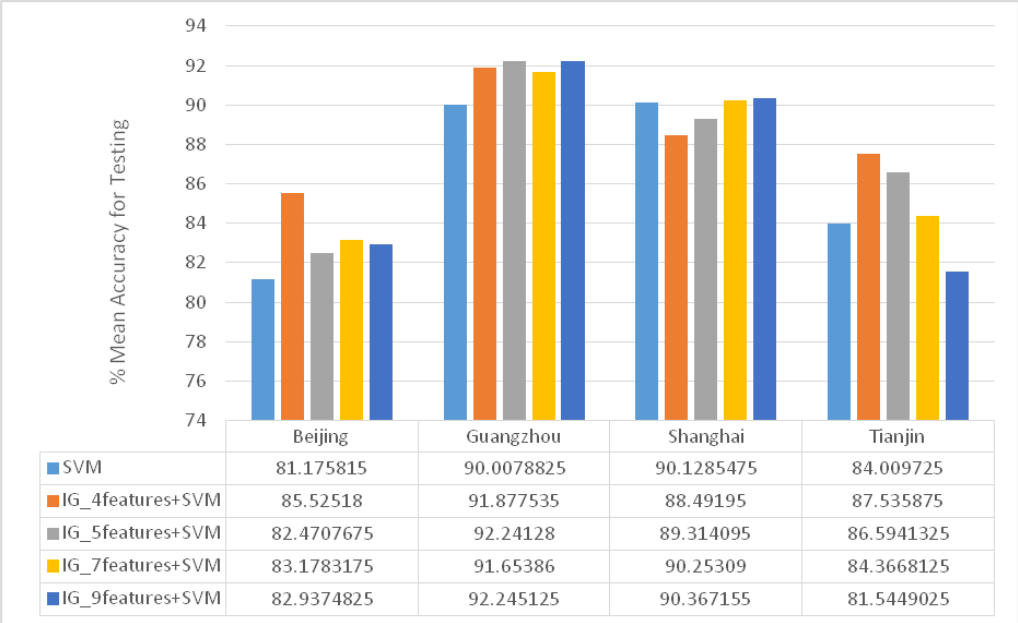
| Sl. No. | Feature Name | Information Gain | Rank |
|---|---|---|---|
| 1 | PM2.5 | 0.8021741 | 1 |
| 2 | PM10 | 0.5501355 | 3 |
| 3 | $SO_2$ | 0.1658154 | 8 |
| 4 | CO | 0.4503882 | 6 |
| 5 | $NO_2$ | 0.26437 | 7 |
| 6 | $O_3$ | 0.1061854 | 9 |
| 7 | Max. Temperature | 0.05342147 | 10 |
| 8 | Min.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

Table 13: Feature Ranking based on Information Gain values for Tianjin.

| Sl. No. | Feature Name | Information Gain | 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 |

Table 10: Feature ranking based on information gain values for Beijing.

| Sl. No. | Feature Name | Information Gain | 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 |




| | 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. 4: Mean percentage accuracy comparison of variable input feature SVM models for Beijing, Guangzhou, Shanghai and Tianjin.

Project of Tianjin Education Committee and Social Science [grant number 2017JWZD16]; Tianjin Science and Technology project [18ZLZDZF00040].

## REFERENCES

Artemio, S.O., Marco A.A., Efrén G.H., Carlos, P.O., Juan, M.R.A. and J. Emilio, V.S. 2013. Forecast urban air pollution in Mexico city by using support vector machines: A kernel performance approach. International Journal of Intelligence Science, 3: 126-135.

Athanasiadis, I.N., Karatzas, K.D. and Mitkas, P.A. 2006. 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.

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

Bordes, A., Ertekin, S., Weston, J. and Bottou, L. 2005. Fast kernel classifiers with online and active learning. Journal of Machine Learning Research, 6: 1579-1619.

Brian, R. and William, V. 2002. 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.

Chan, C.K. and Yao, X.H. 2008. Air pollution in mega cities in China. Atmospheric Environment, 42: 1-42

Cogliani, E. 2001. Air pollution forecast in cities by an air pollution index highly correlated with meteorological variables. Atmospheric Environment, 35: 2871-2877.

García Nieto, P.J., Combarro, E.F., del Coz Díaz, J.J. and Montañés, E. 2013. A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): A case study. Applied Mathematics and Computation, 219: 8923-8937.

Ghaemia, Z., Farnaghi, M. and Alimohammadi, A. 2015. 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, International Conference on Sensors & Models in Remote Sensing & Photogrammetry, Kish Island, Iran. XL-1(W5).

Grosjean, P. and Denis, K. 2012. Package "mlearning"; Machine learning algorithms with unified interface and confusion matrices. R Package Version, 1-0.

Guleda, O.E., Ibrahim, D. and Halil, H. 2004. Assessment of urban air quality in Istanbul using fuzzy synthetic evaluation. Atmospheric Environment, 38: 3809-3815

Gurjar, B.R., Butler, T.M., Lawrence, M.G. and Lelieveld, J. 2008. Evaluation of emissions and air quality in megacities. Atmospheric Environment, 42: 1593-1606.

Hssina, B., Merbouha, A., Ezzikouri, H. and Erritali, M. 2014. A comparative study of decision tree ID3 and C4.5. IJACSA, 4(2).

Jiang, D.H., Zhang, Y., Hu, X. Zeng, Y., Tan, J. and Shao, D. 2004. Progress in developing an ANN model for air pollution index forecast. Atmospheric Environment, 38: 7055-7064.

Klaus, S. and Klaus, H. 2004. Package "kknn"; weighted k-nearest-neighbor techniques and ordinal classification. Discussion Paper 399, SFB 386, Ludwig-Maximilians University Munich.

Kumar, A. and Goyal, P. 2011. Forecasting of air quality in Delhi using principal component regression technique. Atmospheric Pollution Research, 2: 436-444.

Kumar, A. and Goyal, P. 2013. Forecasting of air quality index in Delhi using neural network based on principal component analysis. Pure and Applied Geophysics, 170(4): 711-722.

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

Ross, I. and Robert, G.R. 1996. A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5(3): 299-314

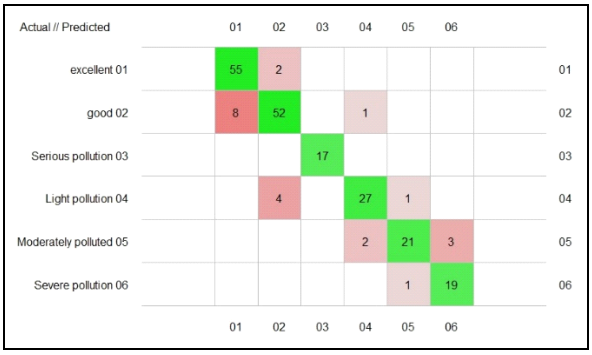



Fig. 5a: IG_4 features+SVM on train1+test1 for Beijing.

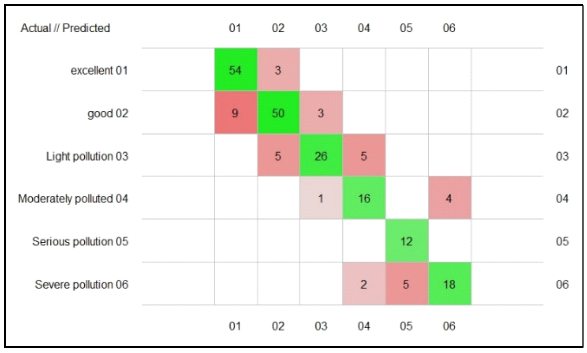



Fig. 5b: SVM on train1+test1 for Beijing.

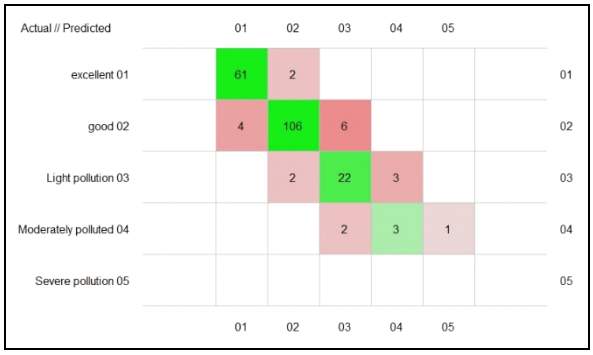



Fig. 6a: IG_9 features+SVM on train2+test2 for Guangzhou.

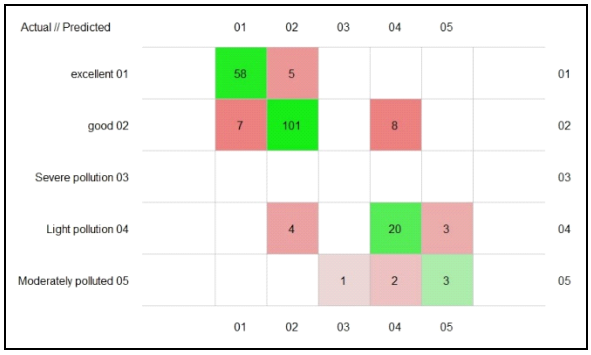



Fig. 6b: SVM on train2+test2 for Guangzhou.

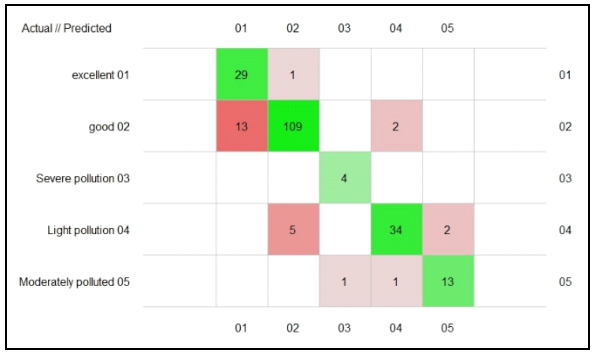



Fig. 7a: IG_9 features+SVM on train1+test1 for Shanghai.

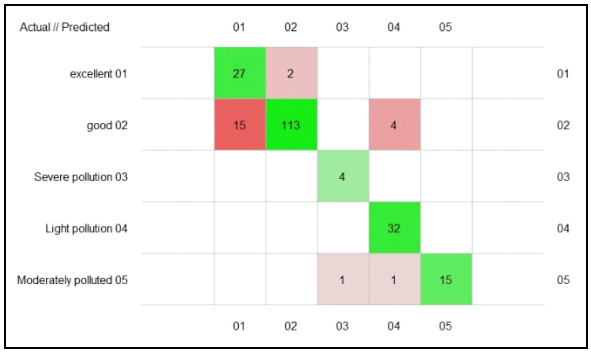



Fig. 7b: SVM on train1+test1 for Shanghai.

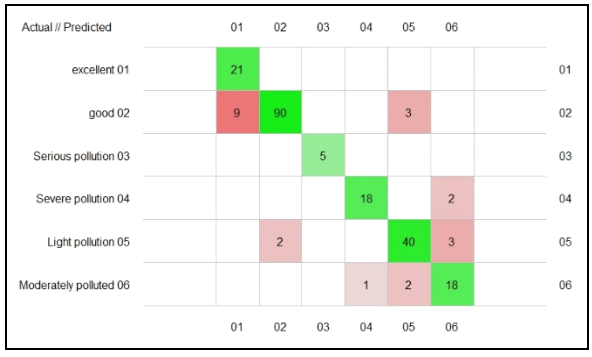



Fig. 8a: IG_4 features+SVM on train1+test1 for Tianjin.

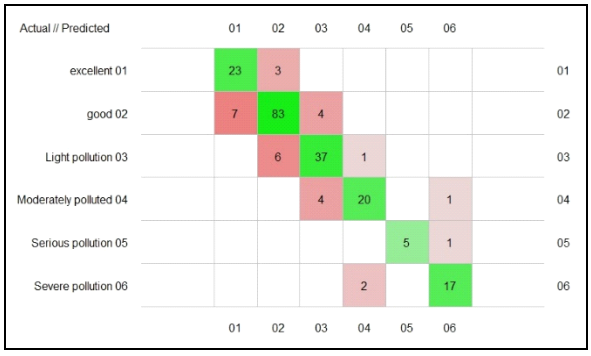



Fig. 8b: SVM on train1+test1 for Tianjin.

Robert, A.R. and Richard, A.M. 2015. Air pollution in China: Mapping of concentrations and sources. PLoS ONE, 10(8): e0135749.

Sánchez, A.S., García Nieto, P.J., Fernández, P.R., Coz Díaz, J. J. and Iglesias-Rodríguez, F.J. 2011. 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: 1453-1466.

Shannon, E. 1948. A mathematical theory of communication. The Bell System Technical Journal, 27: 379-423, 623-656.

Vapnik, V.N. 1999. An Overview of statistical learning theory. IEEE Transactions of Neural Networks, 10(5).

Viotti, P., Liuti, G. and Di, G. 2002. Atmospheric urban pollution: Applications of an artificial neural network (ANN) to the city of Perugia. Ecological Modelling, 148: 27-46

Wang, W.J., Men, C.Q. and Lu, W.Z. 2008. Online prediction model based on support vector machine. Neurocomputing, 71: 550-558.

Yeganeh, B., Shafie Pour Motlagh, M., Rashidi, Y. and Kamalan, H. 2012. Prediction of CO concentrations based on a hybrid partial least square and support vector machine model. Atmospheric Environment, 55: 357-365.

Zhao, H., Zhang, J., Wang, K., Bai, Z., and Liu, A. 2010. A GA-ANN model for air quality predicting. Computer Symposium (ICS), International, 693-699.

Zhao, Y. and Yahya, A.H. 2013. Machine learning algorithms for predicting roadside fine particulate matter concentration level in Hong Kong Central. Computational Ecology and Software, 3(3): 61-73.