Predictive Analysis of Air Pollution using Collaborative Filtering Prediction Algorithm

Samieksha Sharma, Akanksha Gupta
MBS College
India
swnky.samiee@gmail.com, akgupta1809@gmail.com

Monia Digra
SMVD University
India
monia13ite@gmail.com

ABSTRACT: *Growing trends in Air pollution is possessing threat to environment. Various Researchers have extended their work in predicting air pollution using various predictive analytics. In this paper, we are implementing a predictive model for monitoring air pollution level in different cities of India and publishing it as a web service .The algorithm being used is Collaborative Filtering Prediction Algorithm. A comparison has also been carried out in different predictive analytics mainly using Machine Learning techniques such as regression and Deep Learning Technique and Collaborative filtering technique.*

Keywords: Collaborative filtering, Pearson Coefficient, Heuristics, Deep learning

Received: 10 September 2018, Revised 15 December 2018, Accepted 7 January 2019

DOI: 10.6025/jic/2019/10/2/66-76

1. Introduction

With the ever growing development, air pollution is becoming a serious threat to environment. Government and Environment bodies are taking necessary steps to monitor the air pollution levels and take precautionary measures. Various Researchers have applied predictive modelling techniques to develop air pollution models which can be used to monitor air pollution level in different regions. In this paper also, we had tried to develop a prediction model for air pollution. The model is published as a web service, so that it can be used by Third party vendors or government agencies to use air pollution predictive data for monitoring and regulating the levels of air pollution. For this project, we are capturing 15 days data for different cities in India and monitoring parameters like Nitric Oxide, Nitrogen Dioxide, Oxides of Nitrogen, Sulfur Dioxide, Carbon Monoxide, Ozone, PM2.5, Relative Humidity, Wind Speed and Wind Direction. Dataset has been obtained from the web repository of Central board of Air pollution, India. The methodology which is being applied to develop predictive model had been developed by using collaborative filtering approach. Mostly three Types of predictive Modelling Techniques are used for predicting Clustering Models, Propensity Models and Collaborative Filtering.

In Clustering Models, data is clustered in groups based on some grouping variables, Any Test data is matched with the variables of Clusters DNA and categorized in respective cluster as the properties of variable. Behavioral Clustering and product Based clustering are examples of clustering Models. Propensity models are another type of predictive models which predict the future

pattern based on analysis of existing Data. Collaborative Filtering are another type of Predictive analysis model which can predict the property of data based on similar existing data items. Collaborative filtering approach assumes the basic principal, that if item A is similar to item B in terms of properties, then response of Item A to event X will be similar to response of Item B. Collaborative filtering Approach can be of two types Based on Memory and Based on Item data .The memory Based model initially analyses the similarities among users and then select most similar users as the neighbor of active user. Memory Based model suffers shortcomings of large space and time complexity as the number of dataset increases. Model based approach tends to be faster as it possess less space and time complexity, but for better results memory Based model is preferred.

The Accuracy of prediction results by Collaborative filtering approach largely depends on similarity coefficient used to calculate similarity between two items in dataset. Multiple Similarity Coefficient functions have been suggested by various Research works like cosine similarity, means square difference, Pearson Coefficient Similarity. Most effective similarity coefficient used is Pearson coefficient similarity.

2. Literature Survey

Various literature surveys and paper has been published regarding Predictive Analysis of Air pollution. Most of the previous work is done on data set related to similar parameters of air pollution. However different techniques have been implemented by various authors to make models to predict air pollution. Boja and Karumari [1] have proposed the use of Feed-forward Back Propagation network (BPN) model and Mamdani Fuzzy Inference model for predicting SO2 and PM10 concentrations in Indian cities like Hyderabad, vishakapatnam, kurnool. E Martinez, MA Aceves in their paper [2] have used Anfis, a Neuro-Fuzzy system to make a prediction model for Mexico city pollution. However to improve the results and obtain better prediction, author have proposed to used Ant Colony Optimization Algorithm. Liang CI and Wang Kaun in their paper [3] have carried out extensive study on air quality management system using fuzzy logic principle. The main purpose of their study was to evaluate the cost of hazards that occur on external environment due to air pollution. The study obtained through this paper acts as reference for government for air pollution decision making. Opera and Mihalache [4] have carried out a comparative study between Artificial neural Networks and Adaptive Neuo-fuzzy interference approaches used for developing air pollution forecasting models. Cagliero and Cerquitelli [5] have proposed a new data mining system named as "Generalized Correlation analyzer of pollution data (GECKO)", which would help in discovering correlations and associations rules among various heterogeneous parameters like pollutant levels, traffic and climate conditions. Liu and Xiang [6] have presented a approach to design a mobile monitoring system for air pollution in a city. Their approach suggest mounting sensing box on bicycle and collecting data from vehicles running on road. Later this data can be transmitted to data center through Gps receiver and Bluetooth. Data Center will publish this data on website to make it available for public.

Baralis and Cerquitelli [7] have proposed the use of business intelligence methodologies and open technologies to design a data analysis engine. The data analysis engine monitors air pollution based on selective Key Performance Indicators. Wang and Xaio [8] have made use of FCM-HMM Multi-Model to model the air pollution Atmosphere system. Authors have applied FCM-HMM clustering to mine inherent states and then applying TS fuzzy inference multi model for each state to predict air pollution index.

3. Research Methodology

Dataset: For the current Research work, the dataset provided by central pollution control board has been exploited, which concentrates on the scientifically accepted parameters responsible for air pollution. The parameter which are being monitored for research work are Nitrogen, Sulfur Dioxide, Carbon Monoxide, Ozone, PM2.5, Relative Humidity, Wind Speed and Wind Direction Specifically dataset contains information of 8 cities of India.

The project is purposed to be uploaded as web service which can be used by 3rd party applications. The data will be uploaded on Central server, where the processing part will be carried out. Client Terminal can connect to central server and obtain the results.

Algorithm used: Usually when Datasets having large number of dimensions is converted into learning model, they produce results which are over fitted. In order to avoid the problem of over fitting, method of feature selection is followed. Feature Selection is a technique in which subset of only relevant features are selected .Feature selection as a rule gives better learning execution. As size of data set is reduced, thus feature Selection also result in minimizing computation cost. In order to reduce the error rate, it is usually advised to remove the parts of the model that depict illegitimate impacts in the training model as opposed

to genuine elements. Reduced Error Pruning is such technique which eliminates the section of tree, thus making reduction in the size of tree. Reduced Error Pruning likewise brings about reducing complexity and giving better and accurate results. CF

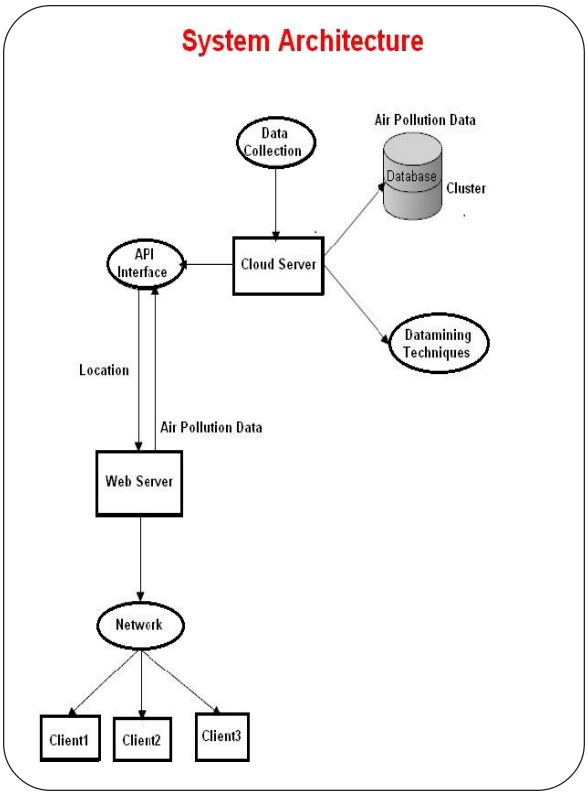

Figure 1. System Architecture of Proposed system

algorithm produces prediction for location according to the reading data of the air pollution. The assumptions are if the readings of some location observed by some sensors are similar, the rating of other location observed by these sensor will also be similar. CF prediction system uses statistical techniques to search the nearest neighbors of the object and then basing on the item rating rated by the nearest neighbors to predict the item rating rated by the object and then produce corresponding prediction list.

Algorithm

Step 1: All locations are weighted with respect to similarity with the query location. Similarity between query locations is measured as the Pearson correlation between their ratings vectors.

Step 2: Select *n* location that have the highest similarity.

Step 3: Compute a prediction, $P_{a,u}$ from a weighted combination. Similarity between two locations is computed using the Pearson correlation coefficient

$$P_{a,u} = \frac{\sum_{i=1}^{rm} (r_{a,i} - r_a) \times (r_{u,i} - r_u)}{\sqrt{\sum_{u=1}^{rn} (r_{a,i} - r_a)^3 \times \sum_{i=1}^{rn} (r_{u,i} - r_a)^3}}$$

$$P_{a,i} = r_a + \frac{\sum_{u=1}^{n} (r_{u,i} - r_u) \times P_{a,u}}{\sum_{u=1}^{n} P_{a,u}}$$

To start with dataset given by Pollution Control Board of India. All locations with their pollution data are checked for similarity coefficient between them .The places which show similar trends in pollution level are grouped together .In next iteration, similarity coefficient is again calculated among places in group. Output model, thus obtained is predicted with test dataset and error rate is calculated.

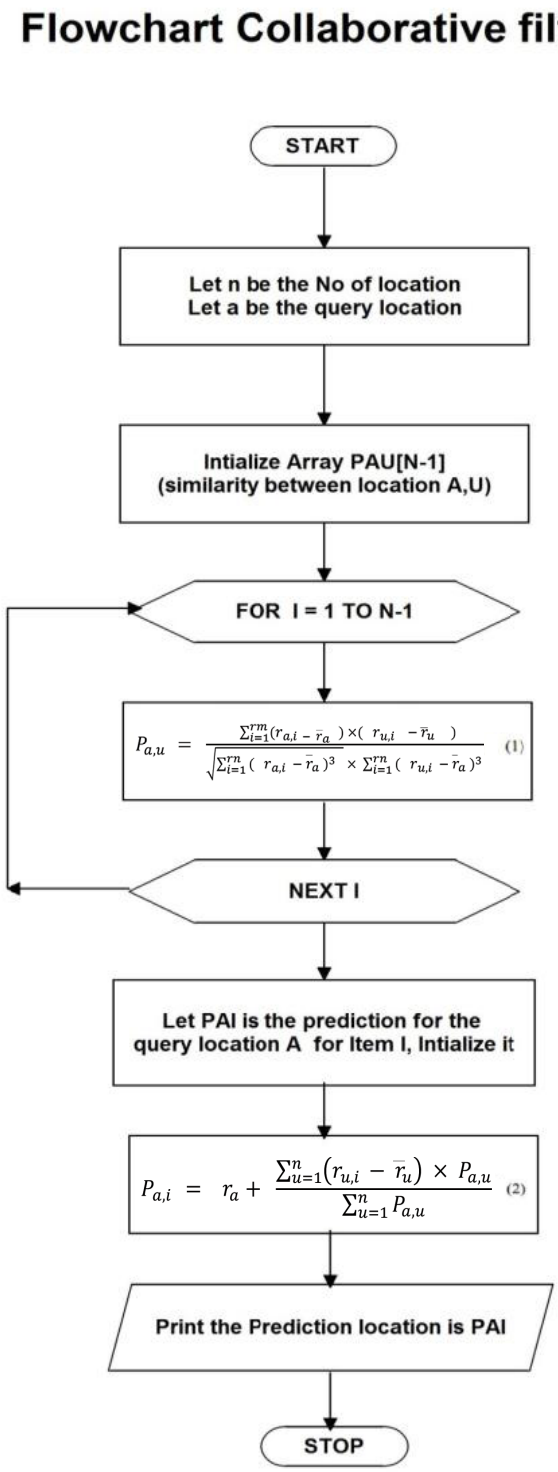

Figure 2. Collaborative filtering flowchart

Artificial Neural Network Prediction Model

Another set of approach that has been used in developing predictive model for regression problem in this project is neural networks. The idea behind using deep learning for predicting air pollution is to develop a pattern for nonlinear relation among factors responsible for air pollution. Artificial Neural Networks works in the same manner a human brain works. In neural networks comprises of multi neurons. Concept of neurons are inspired by neurons present in human brain. A single neuron simulates the functioning of biological neuron by taking input data, processing it and then passing it to other nodes. Weight are associated with each node in hidden layer which are also known as bias. It helps in constraining input processed by each node. They comprises of Input Layer, Hidden Layer and one output layer. Input layer comprises of input factors or also known as independent variables. The number of nodes in input layer is equal to the number of independent variables. The hidden layer comprises of nodes which process as neurons, as the process the input information and give output.

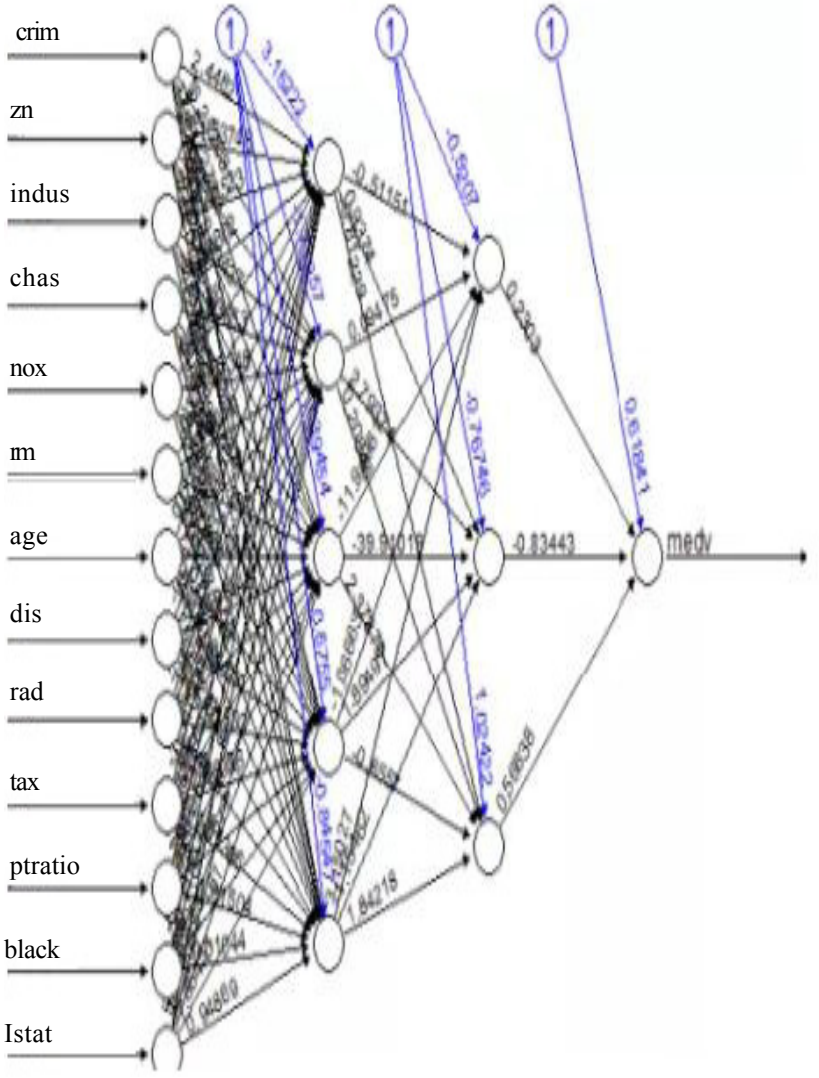

Figure 3. Weights on each node

The number of hidden layers Depends on the configuration of neural network. The artificial neural networks make use of weights to adjust the error rate in predicting output variable. The model in artificial neural network works in same way as functioning of human brain. The Back propagation method is the most common approach in which artificial neural networks perform their operation. Back Propagation method is usually performed in two stages. In first Stage, it carries out data in forward direction in network from left to right It takes information for developing predictive model from input layer and processing of relation between

independent variables is carried out by hidden layers. The predicted result is carried out to output layer. Error rate is calculated by analyzing predicted output with actual output. To minimize the error rate ,the feedback is back propagated in the network and weights adjustment operation is performed in next iteration. Weights are the deciding factors which give weightage to each input variable and thus decide what value should be given to each input variable while developing predictive model. The process of carrying information from input layer and processing the results by hidden layer and then transferring it to output layer is iterated again and again until the error rate is minimized and predicted output result is close or similar to actual output result. This type of learning is usually performed under supervision and hence such training model is known as supervised learning.

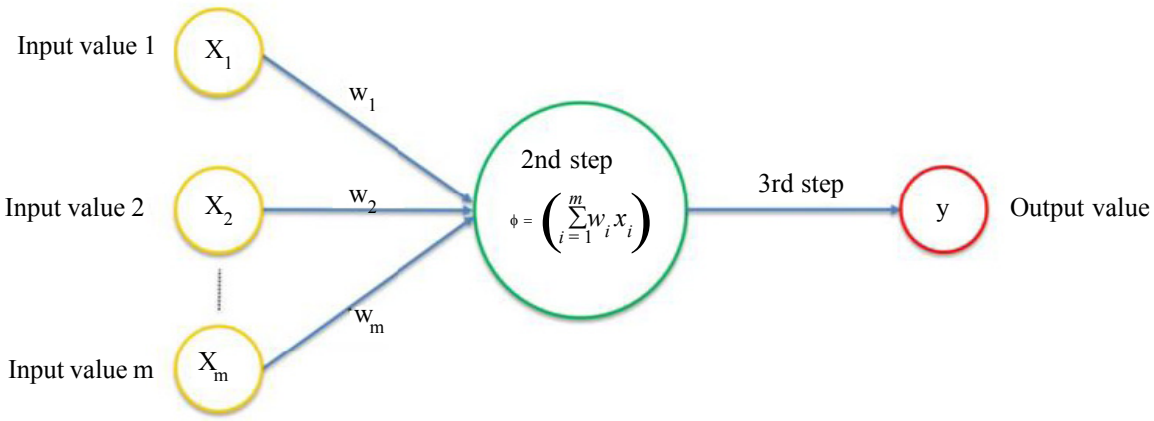

Figure 4. Basic Structure of neural Network

As variables in neural network are usually related to each other in nonlinear manner, therefore in order to inject the non-linearity in network, it is required to make use of activation function. There are mainly predefined activation function which can be used to inject non linearity in our prediction model. Some of the activation function which can be used as threshold activation function, sigmoid activation function, Rectifier activation function, Hyperbolic Tangent function which is also known as TanH function,softmax activation function, etc. In this project, it is dealing with non-linearity by using sigmoid activation function. The graph for sigmoid activation function is usually given as follows.

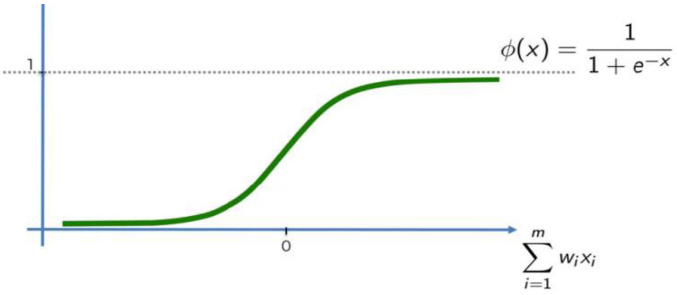

Figure 5. Graph for Sigmoid Activation function

Sigmoid function is defined mathematically as:

$$f(S_{ab}^{j}) = 1 / (1 + e^{(-Sabj)})$$

Where S_{ab}^{j} is the intermediate state of 'b' neuron. The output of b^{th} neuron in hidden layer can be denoted by following equation as

$$O^{k}_{hn} = f(S^{k}_{hn}) = \frac{1}{(1 + exp\,(-\Sigma_m \omega_{mn}\, 0^{k}_{im} - b_{hn}))}$$

it will produce prediction model by neural network approach and simple linear model and produce a comparison between them. Before starting with processing of constructing predictive model, this will pass data through the initial stage of preprocessing in order to remove any kind of noise and maintain uniformity in data. In this stage it mainly check for any kind of anomalies in dataset like missing data values, or scaling of data in proper range. In case of missing data values, it can apply strategies like filling missing data with mean of all data, or substituting the most frequent value .In order to scale data, it follow the process of feature selection. In this process, scale range of data values from -1 to 1. This process is important to eliminate the bias in calculating Euclidean distance due to wide difference in data values next step involves partition of data in two parts. First is training set and another is Test Data set. It is highly recommended to take training dataset and test dataset in ratio of 80:20 or 75:25. Training Dataset will be used to train predictive model and make this neural network model learn and form patterns. Test dataset will be used for checking accuracy. Linear regression model is produced by using gml( ) function . In order to measure prediction accuracy, it is calculating measure of Mean Square. When it is done with linear regression model, then switch to make the prediction model for same data set using the approach of Neural Networks. In neural networks approach also, the very first stage to start with first stage of data preprocessing. In neural network modeling, the data can be preprocessing with the process of normalization in order to process any kind of anomalies present in dataset. In order to come up with neural network prediction model for our dataset, it is required to come up specifications of how many number of hidden layers and neuron in each layer required. Usually more hidden layers mean more complex neural network and consequently it requires more computation power to train such model, so this methodology restrict our self to use two hidden layers. While selecting the number of neurons, the basic rule of thumb has been considered that number of neurons should fall in range between size of input and size of output layer. So, it would prefer to roughly number of neurons to be somewhat around two-third of size of input layer. The data that has been collected for prediction of air pollution has been obtained from online repository of Central Board for Pollution control in India. Tha data is collected from three states Bangalore, Delhi and Chandigarh. The influential factors which are mainly responsible for air pollution are mainly categorized in two categories as Climatic factors and pollutant concentration. Climatic factors include factors like atmospheric pressure, solar radiation, speed of wind, direction of wind, humidity, rainfall and time. Whereas pollutant concentration includes factors like concentration of PM_{10}, NO2, NO3, CO.

The predictive model will analyze relationship between climatic factors and previous concentration of air Pollutant and tries to predict the concentration in next one hour. This approach has named as PKPC "Prediction using Known Pollutant Concentration". In this method, already known concertation of pollutant component will be used along with climatic factors to form relationship model and predict the concentration in next one hour .However, prediction process can be extended to more number of hours by feed backing the results of first time prediction into next iterations. However in such case accuracy of results will be lower than accuracy obtained by using PKPC approach. The configuration of neural network largely depends on the number of hidden layers that are being used in network. Number of hidden layers must be selected in coherence with the number of nodes in input layers otherwise, it can lead to the problem of overfitting if more number of hidden layers are selected or it could affect the accuracy of result, if number of hidden layers selected is less than required. Since Dataset can comprise of entities in different range, so process of normalization comes into picture. It needs to normalize data in order to minimize the biasness in Weight which could come due to disproportionate values of data entries. Normalization process usually involves mapping each value into the range of [0, 1] by using normalization formula as follows:

$$I_{norm} = (I_i - I_{min}) / (I_{MAX} - I_{min})$$

Where I_{norm} is the result of normalized process applied on Item *Ii*. *Imin* is the least value of item in dataset, where as I_{MAX} is the highest value of item in dataset.

In the present work, this paper used 6 ANN models for predicting the value of each pollutant concentration. Each ANN models will be configured with different number neurons in hidden layers and prediction results are validated by analyzing the value of statics measure such as RMS (root mean square value), MRE (mean Relative error) and MAE (mean absolute error). The above mentioned Static variables are calculated for each configuration of ANN model with different number of hidden layers and process is iterated for other pollutants as well.

In this paper, neural net package is used for developing predictive models and testing our training data. Input layer comprises of attributes / parameters that have been captured by dataset to produce a predictive model for air pollution. Input layer is usually constrained to one and the number of nodes in input layer is equal to number of features in dataset. Hidden layers are usually used to capture patterns. The output layer is mainly responsible for processing results and giving output. If the processing results if

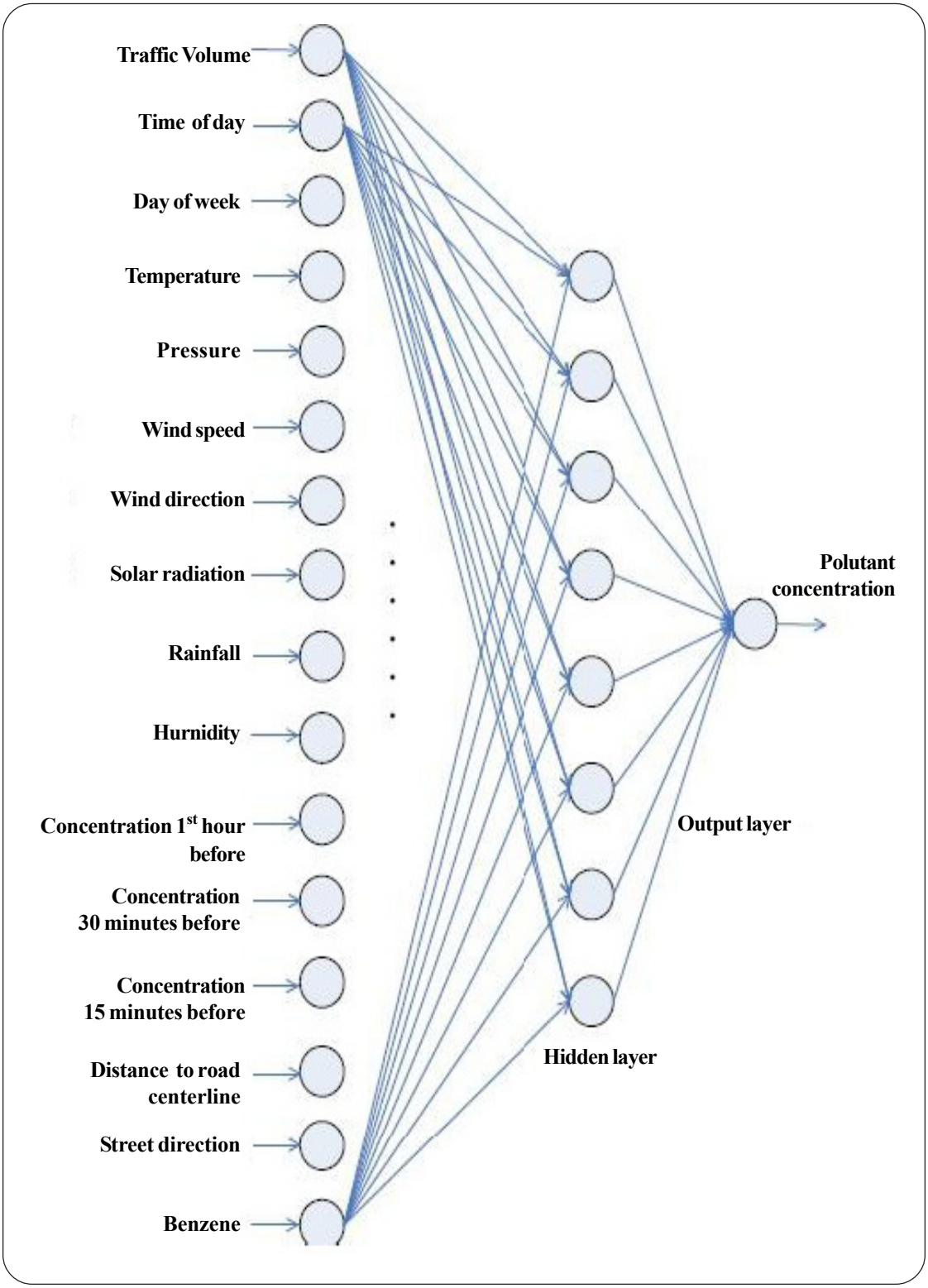

Figure 6. Architecture of suggested neural network with single hidden layer

for classification problem, then number of output units will be similar to number of features in input layer. Else if problem is of regression, then output layer will process data and give one particular value.

4. Results and Analysis

Number of neurons in hidden layer	4	6	7	8	10	12
Root Mean Square Value (RMS)	302.6	283.6	232.9	206.6	252.6	286.8
Mean Absolute Error (MAE)	244.5	238.9	183.0	152.6	193.4	217.9
Mean Relative Error (MRE)	12.75	12.57	10.32	8.59	11.23	11.74

Table 1. Prediction Results of CO based ANN for different number of neurons in Hidden Layer

Number of neurons in hidden layer	4	6	7	8	10	12
Root Mean Square Value (RMS)	288.6	289.6	243.9	211.6	272.6	296.8
Mean Absolute Error (MAE)	257.5	232.9	177.8	147.6	197.4	227.9
Mean Relative Error (MRE)	11.58	13.37	11.36	8.93	10.26	12.32

Table 2. Prediction Results of O_3 based ANN for different number of neurons in Hidden Layer

Number of neurons in hidden layer	4	6	7	8	10	12
Root Mean Square Value (RMS)	271.6	254.6	267.9	219.6	282.6	296.8
Mean Absolute Error (MAE)	232.8	231.7	195.0	155.6	173.4	227.9
Mean Relative Error (MRE)	12.75	12.57	10.32	8.59	11.23	11.74

Table 3. Prediction Results of PM_{10} based ANN for different number of neurons in Hidden Layer

Number of neurons in hidden layer	4	6	7	8	10	12
Root Mean Square Value (RMS)	284.8	285.4	231.7	208.6	255.8	296.4
Mean Absolute Error (MAE)	244.5	238.9	183.0	152.6	193.4	217.9
Mean Relative Error (MRE)	9.67	11.27	10.69	9.63	10.23	12.87

Table 4. Prediction Results of NO_2 based ANN for different number of neurons in Hidden Layer

Graph shown in figure thus denotes the accuracy of the system. However accuracy and efficiency of system largely depends on the Similarity Coefficient being used while calculating similarity between two cities. For example, Cosine Similarity Coefficient does not consider the weight of values, thus may resulting in poor accuracy. However, Pearson Coefficient provides better accuracy as it considers actual data values while computing similarity coefficient. In order to Improve Pearson Coefficient, three heuristics can be considered Impingement, juxtaposition and Idolization Juxtaposition factor, not only calculates the absolute difference between two ratings, but also considers whether these ratings are in agreement or not, giving penalty to ratings in

disagreement. The Impingement factor represents how strongly two items match with each other. The last factor Idolization denotes how common two item have. If both items average data values has a large difference with the average of total data values, the two values can provide more information about the similarity of the two items.

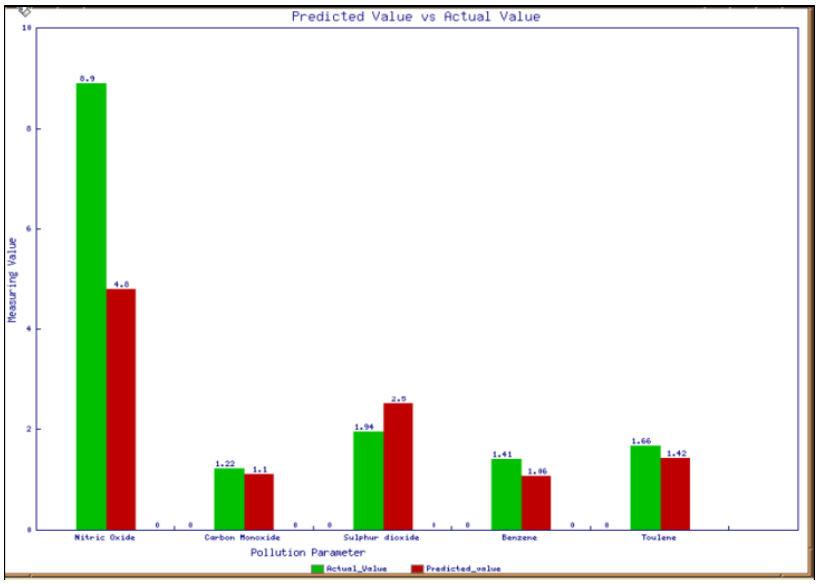

Figure 7. Plotting of points by Collaborative filtering approach

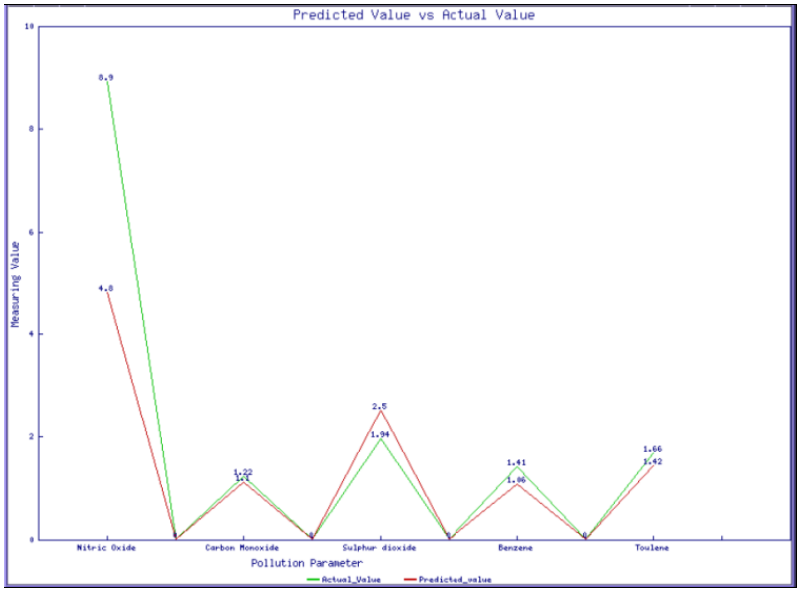

Figure 8. Plotting of points by Collaborative filtering approach

5. Future Scope of Work

The work has been carried out with PKPC (Prediction using Known Pollutant Concentration) approach in which known pollutant concentration has been made use of in predicting next one hour concentration. However the work can be extended by feed backing the predicted output concentration back into model as input variables and thus results can be predicted for N hours. Even though Accuracy in this case will be affected, but however that can improvised by selecting more factors which are responsible for air pollution and developing patterns between them.

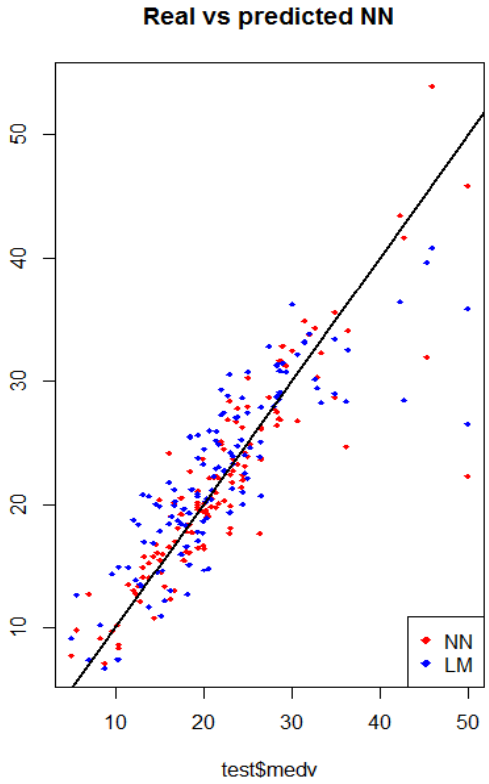

Figure 9. Combined Plot of points for both neural networks and Linear Regression model

References

[1] Boja., Karumari. (2016). Development and evaluation of pollution forecasting model using soft-computing methods for PM10 and SO2 in Ambient Air IEEE 2016.

[2] Martinez, E., Aceves, MA. (2014). Enhancement of a Neuro-Fuzzy Models Using Ant Colony Optimization for the Prediction Level of CO Pollution IEEE 2014.

[3] CI, Dong –Liang., Kaun, Wang. (2009). Knowledge-based air quality management study by Fuzzy Logic principle, IEEE 2009.

[4] opera and Mihalache. (2016). A comparative study of computational intelligence techniques applied to PM2.5 air pollution forecasting IEEE 2016.

[5] Cagliero., Cerquitelli. (2016). Modeling Correlations among Air Pollution-Related Data through Generalized Association Rules IEEE 2016,

[6] Liu., Xiang. (2016). Collaborative Bicycle Sensing for Air Pollution on Roadway, IEEE 2016.

[7] Baralis., Cerquitelli. (2016). Analyzing air pollution on the urban environment, IEEE 2016 .

[8] Wang., Xaio. (2016). Prediction of air pollution based on FCM-HMM Multi-model IEEE 2016.

[9] Frischbier, S., Petrov, I. (2010). Aspects of Data-Intensive Cloud Computing, From Active Data Management to Event-Based Systems and More. p. 57-77.

[10] Open Group. (2011). SOA Reference Architecture, The Open Group, .

[11] Sears, R., Ramakrishnan, R. (2012).bLSM: A general purpose log structured merge tree, *In:* Proc. of SIGMOD 2012.

[12] Bass, C., Clements, P., Kazman, R. (2013).Software Architecture in Practice, Addison Wesley.