Research Journal of Chemical and Environmental Sciences
Res J. Chem. Environ. Sci. Vol 5 [1] February 2017: 79-85
Online ISSN 2321-1040
CODEN: RJCEA2 [USA]

Website: www.aelsindia.com/rjces.htm

RJCES

ORIGINAL ARTICLE

Comparative Analysis of Various Data Mining Prediction Algorithms, Demonstrated using Air Pollution Data of Navi Mumbai

Swati Vitkar
Assistant Professor, SIES (Nerul) College of Arts, Science & Commerce
Email : swativitkar20@gmail.com

ABSTRACT
In today's rapidly increasing industrial environment, air pollution is becoming a reason to worry. The air quality is deteriorating day by day due to careless approach, so causing a serious impact on human health. A need has arouse to predict the level of future air pollutants by entering past data. The development of data-mining applications such as classification, clustering, prediction has shown the necessity for machine learning algorithms to be applied to environment data. In this paper the framework is proposed using Data Mining to study the existing pattern of air pollution data and to predict future pattern of it. The data mining tool WEKA (Waikato Environment for Knowledge Analysis) is used to compare various prediction techniques. It is an open source data mining software and it consists of machine learning algorithms. The aim of this paper is to study the performance and accuracy of different prediction methods such as Bagging, Linear Regression, Rep Tree, Random Forest, Additive Regression and non linear regression algorithm SMOreg on air pollution data of Airoli, Navi Mumbai area.
Keywords: *Data Mining, Air Pollution, WEKA, Bagging, SMOreg, REPtree, Linear algorithm, Regression Analysis, Time Series analysis.*

Received 01.01.2017 Accepted 28.01.2017

INTRODUCTION

Health of people depends on their direct contacts with environment. So, improvement of environment (indoor and outdoor) leads to avoidance of environment-related diseases and thus good health of public. In the long run, good health of a people has a direct effect on the economy of the nation. For good governance, this brings out the significance of a clean environment with investments in public health and education.

Under this study attempts were made to assess the quality of air of specific study sites in Navi Mumbai. Data mining is a good option for designing technology innovatively, in order to serve the community for social causes. It optimizes the environmental impact assessment more effectively, and thus has a great impact on the human beings. Data mining techniques can be efficiently applied on environmental data to discover the new knowledge. Classification and clustering algorithms can be used to understand the data behavior patterns. Predictive analysis is used to evaluate the future trend of data.

The surrounding -- ambient -- air, generally termed as atmosphere, has been taken for granted by individual. Billions of time human beings are breathing in and out from their birth till death, the quality of air has a direct impact on their health, and on other life forms on this planet. The air quality keeps on changing on a daily and even on an hourly basis. The development process and increasing industrialization has led to significant damage and pollution of the atmospheric layer. To improve quality of air, we should have a scientific understanding about the environment and take necessary steps to prevent its further deterioration.

EXISTING SCENARIO AND BACKGROUND

Navi Mumbai the satellite city has a population of 2,600,000. The city has developed considerably in the last decade and with this it is noticed that there is an increase in the air pollutants of Navi Mumbai.

LITERATURE REVIEW

In this study the authors have suggested data mining for environmental modeling, such as where data may be spared, not complete, or heterogeneous. The field of data mining is used to find new pattern in

large amounts of data. Data mining can be applied to environment for e.g.: to find out which pollutant is likely to make larger difference. [3]

[4] stated KDD, Knowledge Discovery in Database process is the inner activity on data mining which is concerned with the discovery of patterns in data. In this paper, the writer also gave an overview of KDD applications in environmental sciences and achieved it by using sample of case studies.

To solve a problem of predicting various defects called Regression through Classification (RvC) the authors have applied a machine learning technique. Software systems are developed which automatically describes the number of defects into a number of fault classes, then learns a model that forecast the fault. Lastly, RvC converts the model class output again into a numeric prediction. [2]

OBJECTIVES

- To create awareness about probable deterioration of environment amongst the people.
- To predict the future trend of environmental data.
- To develop scalable, user friendly data prediction model.
- To help the civic bodies to know the future air pollutant value.

SCOPE OF RESEARCH

The research is restricted to the five zones of Navi Mumbai viz Vashi, Nerul, Airoli, Rabale, Mahape. Secondary data of air pollution for five years (2011-15) is collected from MPCB website. Airoli zone is chosen to demonstrate the same, and subsequently it can be implemented on other zones. The details are given below.

Table 1 : Details of air quality monitoring sites under Study

Zone	Location	Type
Vashi	Fire Brigade compound	Residential
Airoli	Airoli fire station	Rural & other Areas
Nerul	Dr.D.Y. Patil College Building	Residential
Rabale	T.B.I.A, Rabale	Industrial
Mahape	Central lab Building, MPCB	Industrial

RESEARCH METHODOLOGY

This research is a combination of exploratory and experimental research. It is interdisciplinary, case study research. It is quantitative in nature. In the first part the technologies that can be used to achieve the objective. As it is implemented for smaller area and then it can be experimented for larger area, it is experimental in nature.

Data Collection : Secondary air pollution data is collected for Navi Mumbai area from MPCB website for the years 2011 to 2016.

Table 2 : Yearly Air Pollution Data for five zones in studied area

Vashi			
YEAR	**SO2**	**NOx**	**RSPM**
2011	18.223	45.117	102.192
2012	23.310	58.305	98.726
2013	34.450	46.216	123.083
Rabale			
YEAR	**SO2**	**NOx**	**RSPM**
2011	19.020	47.166	111.339
2012	18.057	45.688	88.851
2013	18.485	45.232	71.128
2014	18.332	40.575	144.499
2015	19.718	43.343	125.230
Mahape			
YEAR	**SO2**	**NOx**	**RSPM**
2011	17.757	44.743	114.827
2012	17.206	43.913	117.498
2013	17.975	45.576	166.051
2014	18.057	39.355	153.948
2015	19.179	44.370	107.471

Airoli			
YEAR	SO2	NOx	RSPM
2011	20.478	70.563	154.043
2012	18.430	58.514	141.509
2013	22.176	42.173	85.024
2014	17.794	37.115	32.708
2015	22.588	32.273	40.550
Nerul			
YEAR	SO2	NOx	RSPM
2011	15.236	42.212	134.198
2012	15.457	40.266	103.234
2013	16.353	41.995	104.342
2014	16.965	38.176	142.444
2015	16.833	40.421	128.383

Modelling Research :

The framework is proposed using Data Mining to study the existing pattern of air pollution data and to predict future pattern of it. A study about the technology identified is conducted and then implementation model is suggested. Data Mining tool WEKA 3.8.1 is chosen for the analysis. Different analyses are done on air pollution data of Airoli zone for five years (2011-2015) and regression algorithms are used to understand the data behavior patterns. Predictive analysis is used to evaluate the future trend of data using Bagging, Linear Regression, and Random Forest, non linear regression SMOReg, Additive Regression and REPtree. The predicted results for the year 2016 are compared with the actual values. Also various errors are calculated to check the accuracy of these algorithms.

DATA PREPROCESSING

Steps for finding Trend and Prediction analysis on air pollution data of Airoli area using Weka3.8.1

Air pollution data on daily basis is collected from MPCB website, then that data is converted from daily to monthly by taking average. Then further it is preprocessed by taking average of monthly data to convert it yearly. After that it is converted to .CSV (Comma Seperated File) format.

The preprocessed values for SO_2, NOx and RSPM for all five zones (Airoli, Vashi, Nreul, Rabale and Mahape) for the years 2011 to 2015 are displayed below.

REGRESSION ANALYSIS USING WEKA FOR PREDICTING DATA :

A process which is analytical and meant for discovering huge amount of data and searching the unknown, hidden as well as consistent patterns. It also helps in finding the associations between the variables. Further this data is validated and applied on new data set. Data mining is mainly useful for predictive analysis and prediction.

As this data is very huge and it is time series numeric data, limited algorithms are available for predicting this data. Regression analysis can be carried out in various ways. Some of them are given below.

Linear regression:

It works by estimating coefficients for a line or hyperplane that best fits the training data. This regression algorithm is very easy, fast to train and can have great performance if the output variable for the data is a linear combination of the inputs. It is good idea to evaluate linear regression on any problem before moving onto more complex algorithms.

Reptree :

Quinlan proposed a tree model called Reduced Error Pruning (REP). Each node is replaced with its class which is more popular (starting at the leaves). If the accuracy of prediction is not affected then the change is kept. It is very simple and fast which is the main strength of Reduced Error Pruning. For better results cross validation of tenfold can be applied.

Bagging

A famous ensemble learning method is Ensemble selection – a proved accurate and elective policy. One drawback of ensemble selection is that it is unsteady and sometimes obverts the hill climb set is the only limitation it has. The bagging strategy is proposed here to improve further ensemble selection. Bagging is more accurate and robust classifier than the original. (Bauer, E., Kohavi, R, 1998).

Non-linear regression :

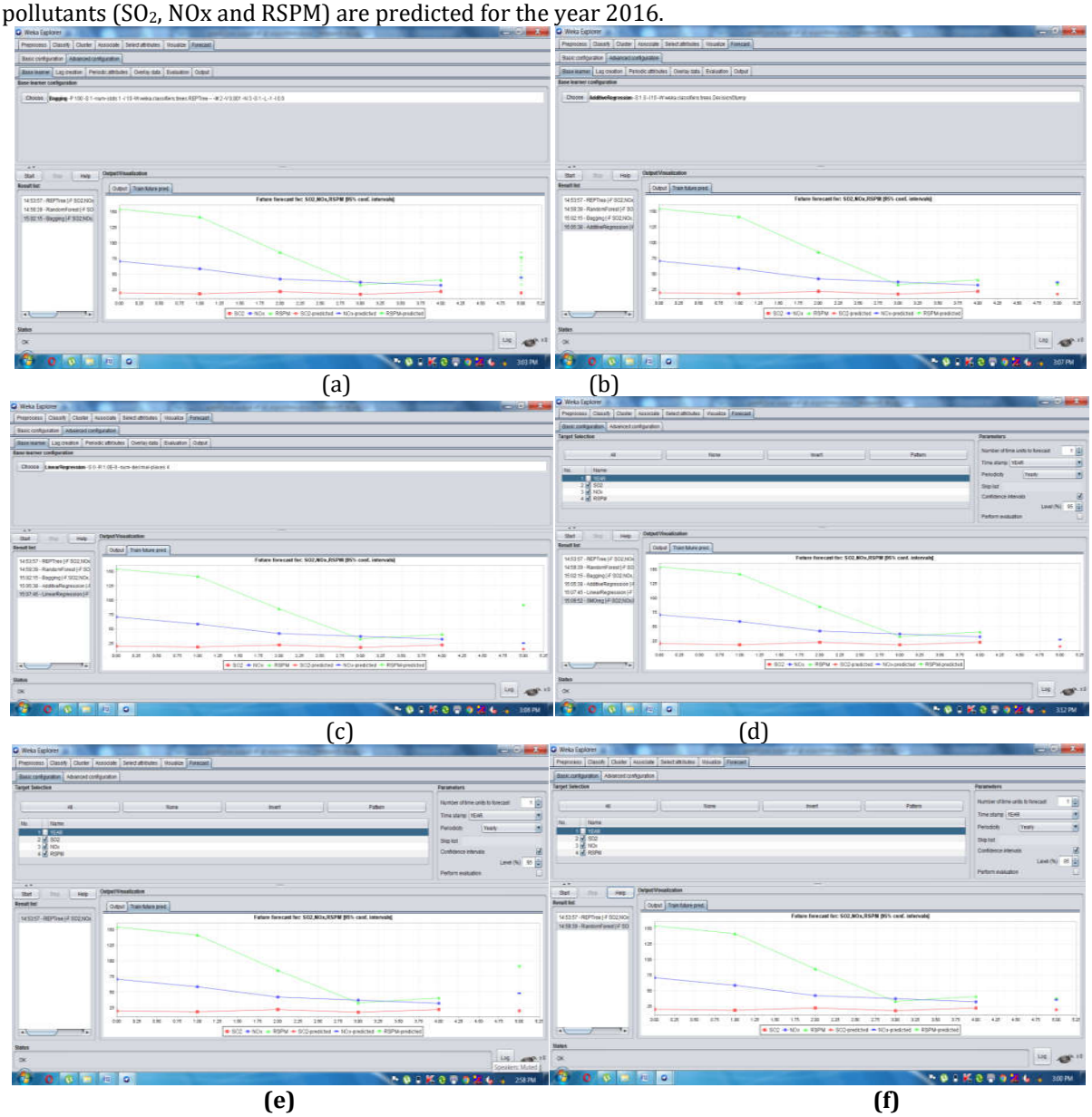
Generally nonlinear regression is best suited when curve passing through data. It defines Y as a function of X when the data is get fitted into the equation for one or more parameters. It seeks the parameter values which produce the curve and appears near to the data. In statistics, nonlinear regression is a form of regression analysis, in which data are modelled by a function which is a nonlinear in nature and depends on (one or more) independent variables. Generally time series analysis engrosses predicting numeric outcomes. In this method the appropriate statistical tools are used to explain time dependent data point series. This method is being used to generate forecast of future values based on the past (known) values. Generally time series applications are useful in: weather forecasting, capacity planning, sales forecasting etc.

Random Forest

(RF) is versatile classification algorithm suited for the analysis of these large data sets. RF is popular because this model of classification have high-prediction accuracy and gives information on the importance of variables for classification. [6]

RESULTS AND DISCUSSIONS

The .csv file is imported from Weka Explorer- Preprocess module. Being time series data, different regression algorithms such as linear regression, non linear regression algorithm SMOreg, bagging, REPtree, Random Forest and Additive Regression are used for forecasting. And values for the same air pollutants (SO_2, NOx and RSPM) are predicted for the year 2016.



(a) (b)

(c) (d)

(e) (f)

Figure 1: Graph for predicted values of SO_2, Nox and RSPM using all six algorithms

Table 3 : Predicted air Pollutants Values for Airoli Zone for the year 2016

Predicted values for 2016	SO2	NOX	RSPM
REPTree	20.2932	48.1276	90.7668
RandomForest	19.3463	35.6075	36.7708
Bagging	20.0064	44.7356	75.9123
Additive Regression	17.794	37.1245	32.708
Linear Regression Model	15.1332	25.7589	90.7668
SMOReg	14.996	27.4213	-0.2931

Table 4 : Actual Air Pollutants values values of Airoli Zone for the year 2016

2016	SO2	NOx	RSPM
Airoli	30.495	47.040	37.273

Then compared its results with the actual data and also the accuracy of the algorithms is decided by calculating Mean absolute error, Root mean squared error, Mean absolute percentage error and Mean squared error shown below.

Mean Absolute Error			
Algorithm	**SO2**	**NOX**	**RSPM**
REPTree	2.2256	10.9406	38.0061
Random Forest	1.0117	1.9869	1.4539
Bagging	2.3212	7.5486	29.2261
Additive Regression	0.0003	0.0063	0
Linear Regression	0.0462	0.0095	38.0061
SMOreg	0.0058	0.0537	0.1449

Table 5 : Mean Absolute Error for all four algorithms

Root mean squared error			
Algorithm	**SO2**	**NOX**	**RSPM**
REPTree	2.2403	11.6634	44.443
Random Forest	1.2042	2.3733	1.5748
Bagging	2.3286	8.5627	32.6605
Additive Regression	0.0004	0.0077	0
Linear Regression	0.0462	0.0123	44.443
SMOreg	0.0059	0.0555	0.1778

Table 6 : Root Mean Squared Error for all four algorithms

Mean squared error			
Algorithm	**SO2**	**NOX**	**RSPM**
REPTree	5.019	136.0343	1975.177
Random Forest	1.4502	5.6327	2.48
Bagging	5.4222	73.3193	1066.7099
Additive Regression	0	0.0001	0
Linear Regression	0.0021	0.0002	1975.177
SMOreg	0	0.0031	0.0316

Table 7 : Mean Squared Error for all four algorithms

Mean absolute percentage error			
Algorithm	**SO2**	**NOX**	**RSPM**
REPTree	10.8983	30.9725	102.7
Random Forest	4.5565	5.5398	3.5741
Bagging	11.2154	21.7417	76.6714
Additive Regression	0.0013	0.016	0
Linear Regression	0.2242	0.0254	102.7
SMOreg	0.0278	0.1491	0.3945

Table 8 : Mean Absolute Percentage Error for all six algorithms

Finally all four errors that is Mean Absolute Percentage Error, Root mean squared error, Mean Squared error are compared and it is observed that Additive Regression model is best suited for prediction because all the error values for the same is minimal as compared to other error values.

CONCLUSION

Air pollution play hazardous role in the health of the humans and plants. As there are many different sources of pollution, finding the effects of air pollution on health are very complex and their individual effects differ from one to the other. The ambient air quality is assessed from various parts of Navi Mumbai and industrial area. The online data has been collected from Maharashtra Pollution Control Board (MPCB), ambient air quality data for the past five years from 2011 to 2015. The data are pre processed and Data can be further processed by data mining tool and proper decision support can be given to the policy makers.

This paper proves that data mining techniques are valuable tools that could be used for environmental monitoring and natural resource science field, and are thus of interest to NGO's, Municipal corporations etc. This work is a better starting point for implementation of data mining for real world examples. The use of open-source data mining tools is the main advantage of this study. WEKA is a very useful alternative over other existing tools available. The current values of the air pollutants are provided to different algorithms such as Bagging, Linear Regression, REP tree, Random forest, Additive Regression and non linear regression algorithm SMOreg. The output of these algorithms were tabulated and compared with the actual predicted values. The comparison was made by calculating Mean Absolute percentage Error, Mean Absolute Error, Root mean squared, Mean Squared Error are calculated for all these algorithms. It was observed that the output of non linear regression algorithm Additive Regression has minimum errors and is closer to the actual values. Hence Additive Regression is selected for further predictions.

In future this research can be extended to predict the air pollution outside of Navi Mumbai and in other states. This study proves that data mining technology is the optimal solution to design and develop a scalable and secure model to predict the future values and help the policy makers to take the remedial measure to minimize it.

FUTURE ENHANCEMENT

Data mining technique when combined with GIS contribute towards the new technology of spatial data mining. Thus survey and analysis of pollution pattern using GIS can become further useful for achieving sustainable management of water resources and air pollutants.

APPENDIX

Table : National Ambient Air Quality Standards

Pollutants	Time weighted average	Concentration in ambient air		
		Sensitive area	Industrial area	Residential area
Carbon monoxide	8 hrs	1.0mg\m^3	5.0mg\m^3	2. 0mg\m^3
	1 hr	2.0mg\m^3	10.0mg\m^3	4. 0mg\m^3
Oxides of nitrogen	Annual	15µg\m^3	80 µg/m^3	60 µg\m^3
	24 hours	30 µg\m^3	120 µg\m^3	80 µg\m^3
Sulphur dioxide	Annual	15 µg/m^3	80 µg/m^3	60 µg/m^3
	24 hours	30 µg/m^3	120 µg/m^3	80 µg/m^3
Respirable particulate matter(<10 um)	Annual	50 µg/m^3	120 µg/m^3	60 µg/m^3
	24 hours	70 µg/m^3	150 µg/m^3	100 µg/m^3
Suspended particulate matter	Annual	70 µg/m^3	360 µg/m^3	140 µg/m^3
	24 hours	100 µg/m^3	500 µg/m^3	200 µg/m^3
Lead	Annual	0.50 µg/m^3	1.0 µg/m^3	0.75 µg/m^3
	24 hours	0.75 µg/m^3	1.5 µg/m^3	1.0 µg/m^3

Source : U.S. EPA National Ambient Air Quality Standards

ACKNOWLEDGEMENT

The author would like to thank Maharashtra Pollution Control Board for online Data.

REFERENCES

1. Luan, J., Zhao, C.-M., and Hayek, J. (2004). Exploring a new frontier in higher education research: A case study analysis of using data mining techniques to create NSSE institutional typology : Paper presented at the California Association for Institutional Research, Anaheim, California, November 17-19, 2004.
2. Bibi, S., Tsoumakas, G., Stamelos, I. and Vlahavas, I.P., 2006, March. Software Defect Prediction Using Regression via Classification. In *AICCSA* (pp. 330-336).
3. Spate, J., Gibert, K., Sànchez-Marrè, M., Frank, E., Comas, J., Athanasiadis, I., & Letcher, R. (2006). Data Mining as a tool for environmental scientists. In *Proceedings of the iEMSs Third Biennial Meeting:" Summit on Environmental Modelling and Software"*. International Environmental Modeling and Software Society.
4. Dzeroski, S. (2003). Environmental Applications of Data Mining. *Lecture Notes of Knowledge Technologies, University of Trento*.
5. Weka, University of Waikato, New Zealand, http://www.cs.waikato. ac.nz/ml/weka/ Retrieved on 6/12/2016 at 8.10 pm.
6. Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & van Hijum, S. A. (2012). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?. *Briefings in bioinformatics*, bbs034.
7. Collier, K., Carey, B., Sautter, D., & Marjaniemi, C. (1999, January). A methodology for evaluating and selecting data mining software. In *Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on* (pp. 11-pp). IEEE.
8. Nghe, N. T., Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. *FIE'07. 37th Annual* (pp. T2G-7). IEEE.
9. Luan, J. (2002). Data Mining and Knowledge Management in Higher Education-Potential Applications.
10. Gholap, J., Ingole, A., Gohil, J., Gargade, S., & Attar, V. (2012). Soil data analysis using classification techniques and soil attribute prediction. *arXiv preprint arXiv:1206.1557*.
11. Sandhya, N., Anuradha, K., Althaf, S., Basha, H., Premchand, P., & Govardhan, A. (2009). Rank Analysis Through Polyanalyst using Linear Regression. *IJCSNS-International Journal of Computer Science and Network Security*, *9*(9), 290-293.
12. Kosorus, H., Honigl, J., & Kung, J. (2011, August). Using R, WEKA and RapidMiner in time series analysis of sensor data for structural health monitoring. In *Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on* (pp. 306-310). IEEE.
13. Takeuchi, H., Mayuzumi, Y., & Kodama, N. (2011, August). Analysis of time-series correlation between weighted lifestyle data and health data. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE* (pp. 1511-1514). IEEE.
14. Eltoft, T. (2002). Data augmentation using a combination of independent component analysis and non-linear time-series prediction. In Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference *on* (Vol. 1, pp. 448-453). IEEE.
15. Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., & Murthy, K. R. K. (2000). Improvements to the SMO algorithm for SVM regression. *IEEE transactions on neural networks*, *11*(5), 1188-1193.

CITE THIS ARTICLE

Swati Vitkar. Comparative Analysis of Various Data Mining Prediction Algorithms, Demonstrated using Air Pollution Data of Navi Mumbai. Res. J. Chem. Env. Sci. Vol 5 [1] February 2017. 79-85