International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 2 (2015) pp. 4619-4626

http://www.ripublication.com

Applying Regression Technique On Environmental Data By WEKA

Vidyullatha Pellakuri

Research Scholar, department of CSE, KL University.

D. Rajeswara Rao

Professor, Department of CSE, KL University, Guntur(dt), AP.

Lakshmi Narayana

Student, Department of ME, NIT, Warangal, AP.

Abstract

Now a day's Data Mining process is implemented in different applications by using various tools and methods. Data mining tools are user-friendly and allow the knowledge driven decisions. For effective and efficient data representation and visualizations these tools are applied. WEKA is such tool which is easily implemented for any data streams such as medical, environmental, spatial, text, web, etc. This paper throws some light on prediction of air pollutants in environmental data for forthcoming year using data mining tool WEKA. The air pollutants data was collected from the power industry in Andhra Pradesh and forecasting the pollutants for forthcoming year. Using WEKA, data is analyzed by correlations and linear regression model within a short time period.

Key-words: - Data Mining, Linear Regression, Preprocessing, WEKA.

1. Introduction

Data mining is a bright and relatively new technology. Data mining [1] is the procedure of identifying rules and trends that exist in the data which can be collectively termed as a data mining model. Data mining models can be applied to specific scenarios, such as forecasting, estimating analyzing risks, assigning probabilities to diagnosis or other out comes. Recommendations are to determine

which products are likely to be sold together. Finding sequences are analyzing customer selections in a shopping cart, predicting next likely events. Grouping can separate the customers or events in to a cluster of related items. Data mining tools are applied to predict behaviours and future trends to create knowledge-driven conclusions. Data mining tool WEKA [2] uses the statistical procedures such as simple linear regression, multiple linear regressions as well as logistic regression models and machine language techniques like clustering analysis, multi layer perceptrons and decision trees. This paper centers on Correlation analysis, linear regression and prediction. Correlation explains the inter-relation between attributes. While comparing the relation between two attributes, it is necessary to find a variation in one attribute that effects a change in the other. Correlation and regression are helpful for statistical analysts to understand relationships between parameters. Regression analysis is a statistical tool that gives the relationship between two or more attributes to find out one variable (dependent) from the other variable (independent). Prediction is almost like classification; the only difference is that it is a continuous one when compared to classification which is qualitative discrete attribute. The aim of prediction is to determine the quantity of the target attribute for unknown objects. This type of analysis is also known as regression and the prediction with time series data is generally named as forecasting.

2. Literature Survey

Gerami Farzad et al[3] Predicted Workplace Accidents by WEKA Software using linear regression model in May 2014. In April- 2014, Vrushali Bhuyar [4] used classification Techniques on Soil Data and Predicted Fertility Rate for Aurangabad District. Velide Phani kumar and Lakshmi Velide [5] discussed and predicted nitrogen, phosphorus and sulphur in soil in less time by linear regression and showed accurate results in Warangal. Chandana Napagoda [6] discussed about Web Site Visit Forecasting Using Data Mining Techniques in 2013. They mainly concentrated on forecasting of web site visits by using prediction methods such as Gaussian, Linear regression, Multilayer Perceptron regression and SMO regression. Nan Gao and Xueming Shu et al[7] gave an idea about Forecasting Model on Emergency Incidents in a city using WEKA tool in the month of October 2013. Elia Georgiana Dragomir [8] suggested an Air Quality Index Prediction using K-Nearest Neighbour Technique in 2010. Rajesh Kumar [9] discussed about decision tree method in predicting the dependent variables like fog and rain for weather forecasting using WEKA. Haizhou DU[10] gave an idea about Wind Power Load Forecasting based on the data mining classification techniques using WEKA Shalini Gambhir et al [11] suggested that Regression model is the best practise method to predict output for Quality of Web Services dataset with WEKA. The present work is focused on the primary pollutants such as sulphur dioxide and Oxides of Nitrogen, data are collected from the power industry of Andhra Pradesh of the previous year and predicted to the next academic year by using the WEKA tool.

3. Materials and Methods

The available data mining tools are weka, Rapid miner, Tanagra and Orange. This report gives the details of data mining tool WEKA (Waikato Environment for Knowledge Analysis) [12] which is a collection of visualization tools and machine learning algorithms for data mining tasks like data analysis and predictive analysis used in many different application areas, in particular for educational uses and research. It is produced at the University of Waikato, New Zealand and it is fully enforced in the Java programming language, free availability and simplicity of use due to its graphical user interfaces. After installing the data mining tool WEKA 3.7.4, the graphical user interface consists of four applications such as Explorer, Experimenter, knowledge flow and simple CLI [13]. The explorer application consists of pre-processing panel, classify panel, cluster panel, Association panel, selection attributes panel and visualization panel. Preprocess panel facilitates for importing the data in the form of ARFF (attribute relational flat file) and preprocessing the data by normalization. The Classify panel enables the user to apply Gaussian process, regression process, decision rules and decision tress to the resulting data set and to estimate the accuracy of the predictive model & to visualize the data in margin curves, threshold curves and cost benefit analysis and so on. The Cluster panel gives access to the clustering algorithms such as simple k-means, Cobweb, Density based cluster, hierarchal cluster etc. The Associate panel provides Apriori algorithm for frequent item set mining. The Select attributes panel provides algorithms for placing the most predictive attributes in a dataset. The Visualize panel shows a scatter plot matrix where individual scatter plots can be selected and magnified and studied further using various selection operators.

4. Data Analysis

Data is collected and written the data into WEKA in Attribute-Relation File Format (ARFF) [14] as shown in the following syntax fig [1].

Header section:
@relation <relation-name>
@attribute <attribute-name> < data-type>
The data types are numerical or nominal
Data Section:
@data
Instances are represented in a single line and separated by commas. The missing values are represented by a single question mark.

Fig: 1 ARFF Syntax

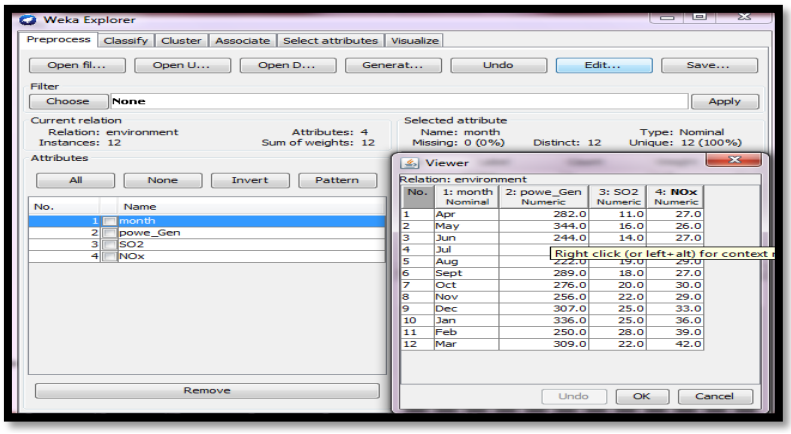

Fig: 2 loading a dataset into WEKA

The ARFF file is imported in to WEKA [15] by choosing the "open file" option. Once the data is loaded, WEKA recognizes attributes that are shown in the Fig[2]. WEKA computes some basic statistics on each attribute. The following statistics for each attribute value Min, max, mean, standard deviation (Std-Dev) is shown for continuous attributes in the Fig[3]. A data mining tool WEKA is used to analyze the environmental data for correlation and linear regression between variables of power generation and air pollutants. The correlation is a positive between the variables. Then, develop a linear regression for both pollutant variables Sulphur dioxide and Nitrous oxide. From the classifier panel of WEKA, choose the option "functions" and select linear regression model & make the cross-validations fold at 10, then click on start button, then linear regression model is shown in Fig [4] & [5]. For the linear regression model of sulphur dioxide is 0.0659* power + 1.0373 where nitrous oxide is 0.1046*power+1.3112 are shown. From this, the emission of sulphur dioxide and nitrous oxide can be calculated at any point of time for known value of power generation.

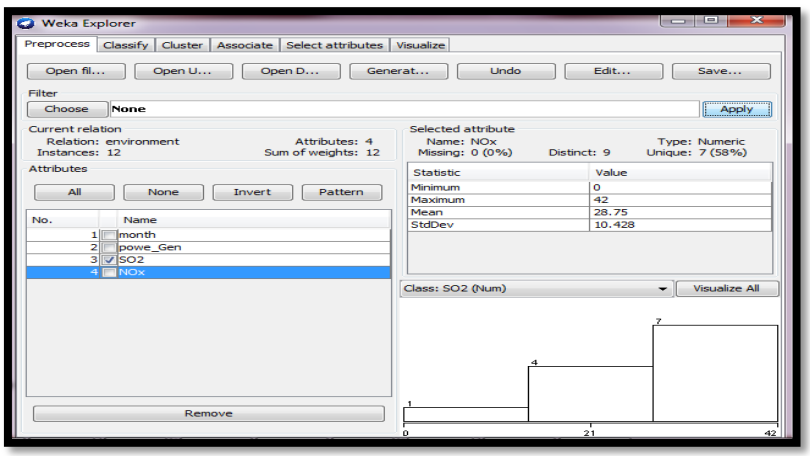

Fig:3 pre-processing the environmental data

Likewise, the values of sulphur dioxide and nitrous oxide for the year forthcoming year are calculated and shown in fig [6].

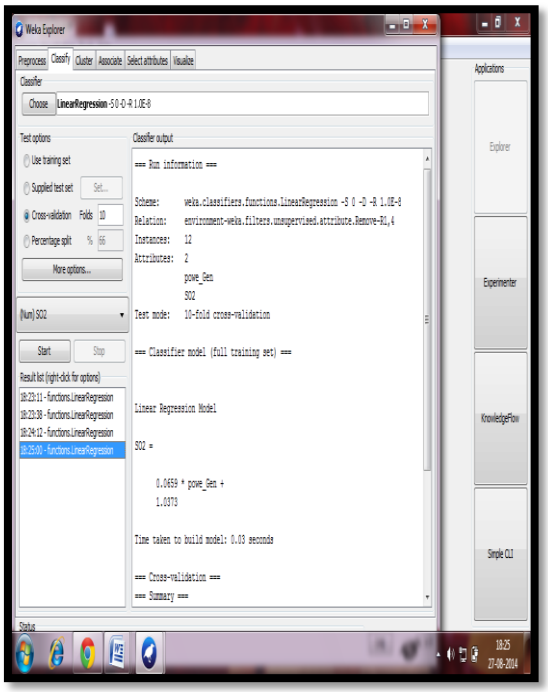

Fig:4 Linear regression model using WEKA for Pollutant Sulphur dioxide.

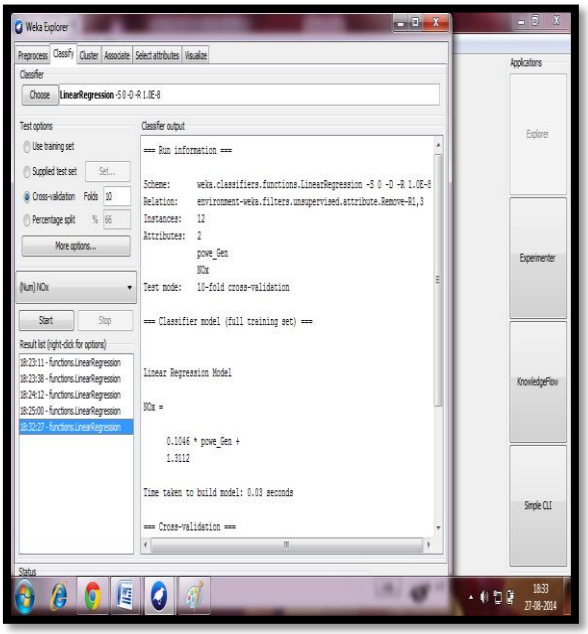

Fig:5 Linear regression for Pollutant NOx

5. Results and Discussion

For the data mining tool WEKA, it is possible to predict the air pollutants using linear regression model. The results are shown in the fig [6]. The predicted values are depending upon the target values which are fixed by Central Electricity Authority (CEA). However there is a slight variation between the target and actual value due to bad quality of coal, inefficient combustion and more auxiliary motor loads in power industries. There are different methods for checking the accuracy of predicted value from linear regression like correlation coefficient, square of residuals, correction factor and standard estimate of error. By using the Correction factor, the predicted values are approximately equal to actual values. Let us check the predicted values to the actual values for any one of the month, suppose for the month july-14 the predicted values of sulphur dioxide and nitrous oxide are 24.76 & 38.97 µg/m3 which is having power generation target of 360 MU which are fixed by Central Electricity Authority (CEA) but, the actual measured values during the same month of sulphur dioxide and nitrous oxide are 23 µg/m^3, 35 µg/m^3. For solving this problem, the correction factor is used. Using the correction factor the predicted values are approximately equal to the actual values. The Correction factor = actual value / target value. so, that the pollutants such as sulphur dioxide and nitrous oxide values are multiplied with correction factor, we will get the approximate results which are almost matching with forecasting values. It is clearly observed that, by using data mining tool WEKA, the prediction is very easy and it can be generated the values within a short time period.

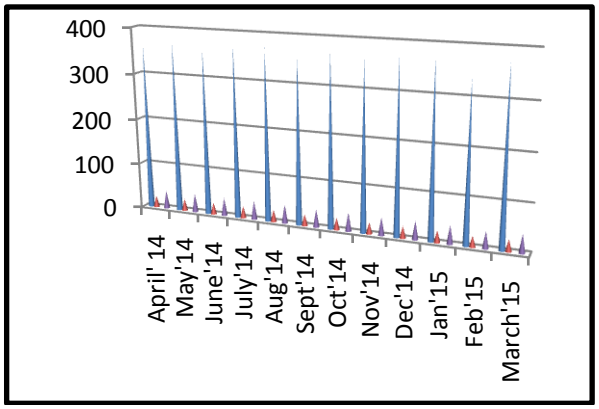

Fig: 6 Predicted values for environmental data.

6. Conclusions

This research gives a valuable knowledge on air pollutants pattern and its factors. This previously unknown knowledge is mined from the dataset provided us by a power industry in Andhra Pradesh State. Coming to an end, the regression technique using data mining tool WEKA yields a better results within a short time period. In future, we would like to analyze a latest dataset with more attributes comprising of last five years, to inspect the results by using soft computing techniques and to recognize that

how the air pollutants are disseminating within a period of five years. This research provides a valuable knowledge to plan and enhance the pollution control program efficiently in Andhra Pradesh.

7. Acknowledgement

My sincere thanks to Dr. D. Rajeswara Rao, Research guide of KL University, for motivating and helping me in every task. My heartful thanks to the power industry of Andhra Pradesh State for providing me the facility for data collection and sampling. I am thankful to Mr. Lakshmi Narayana for providing valuable suggestions in my research work.

8. References

[1] J. Han and M. Kamber, (2000) "Data Mining:Concepts and Techniques," Morgan Kaufmann.

[2] Ian H.Witten and Elbe Frank, (2005) "Datamining Practical Machine Learning Tools andTechniques," Second Edition, Morgan Kaufmann, San Fransisco.

[3] Gerami Farzad, Bartashak Masoumeh,Kourosh Rocky, Razieh Honarmand"Prediction of Workplace Accidents with Knowledge Discovery Approach Using Weka Software", Nova Explore Publications, *Nova Journal of Engineering and Applied Sciences*Vol 2(5), May 2014:1-8

[4] Vrushali Bhuyar "Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District", *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)* Volume 3, Issue 2, March – April 2014.

[5] Velide Phani kumar and Lakshmi Velide "Data mining plays a key role in soil data analysis of Warangal region", *International Journal of Scientific and Research Publications*, Volume 4, Issue 3, March 2014

[6] Chandana Napagoda."Web Site Visit Forecasting Using Data Mining Techniques", *international journal of scientific & technology research*, volume 2, issue 12, december 2013.

[7] Nan Gao, Xueming Shu, Jiting Xu, Biao Wen, Peng Chen, and Peng Wu "The Study of Quantitative Forecasting Model on City Emergency Incidents", *International Journal of Information and Education Technology*, Vol. 3, No. 5, October .

[8] Elia Georgiana Dragomir "Air Quality Index Prediction using K-Nearest Neighbor Technique*", buletinul universităţii petrol – gaze din ploieşti*, Vol. LXII No. 1/2010

[9] Rajesh Kumar "Decision Tree for the Weather Forecasting", *International Journal of Computer Applications* (0975 – 8887)Volume 76– No.2, August 2013.

[10] Haizhou DU "Intelligent Optimization Research of Wind Power Load Forecasting", *Journal of Pattern Recognition & Image Processing* 4:4 (2013) 507-513.

[11] Shalini Gambhir et al "Regression model for Quality of Web Services dataset with WEKA*", International Journal of Electronics and Computer Science Engineering* (IJECSE, Volume 2, Number 3)

[12] *Geoffrey Holmes, Andrew Donkin, and Ian H. Witten "WEKA: A Machine Learning Workbench".*

[13] E.Frank, Machine Learning With WEKA, University of Waikato, New Zealand.

[14] B. Mobasher, Data Preparation and Mining with WEKA

[15] Weka Machine Learning Project, http://www.cs.waikato.ac.nz/~ml/index.html.