ISSN: 0974-2115
www.jchps.com Journal of Chemical and Pharmaceutical Sciences

Identification of lead contaminant in river water quality data

S Brintha Rajakumari*, C.Nalini
Dept. of CSE, Bharath University, Chennai.
***Corresponding author: E-Mail: brintha.ramesh@gmail.com**

ABSTRACT

Data Mining is an art and science of extracting hidden information from the large datasets. Water pollution assessment is a critical area of study because it directly affects the human beings. The simple k Mean, EM and cobweb clustering algorithms are used to analyze the lead contaminant in the river water quality dataset. The simple KMean and hierarchical clustering methods takes less time to build the model than the other methods. This paper presents the comparison of different clustering algorithm used for the river water quality dataset and predicts the cluster, which contains the highest value of lead in river water.

Keywords: clustering, Data mining, River Water Quality, Weka tool.

1. INTRODUCTION

Due to the huge amount of data being created from normal business operations, there is a demand to take information in modern data repositories to make effective decisions. Data mining is a data driven approach and applies statistical and machine intelligence tools to extract useful, exact and meaningful patterns from the large volume of data sets. Data mining tools are used every day to solve the real problem in business, engineering and science.

Water quality monitoring helps in drawing the water quality trends and prioritizing pollution control efforts and calculate approximately the environment and the level of pollution control required and effectiveness of pollution control measures.

The paper prepared the heavy metal pollution index of the river in India and applied multivariate statistical analysis, factor analysis and cluster analysis to predict the source of heavy metals. The cluster analysis is used for spatial characterization to evaluate the most important parameters that affect the water health of the river.

The analysis of complex data sets with enhanced evaluate the water quality and a variety of environmental issues, including study the spatial and temporal patterns of water quality, chemical group associated with hydrological conditions, amount of pollution sources, the principal component analysis (PCA) and cluster analysis (CA) have been widely used.

2. METHODS

Analysis of lead in river water: Weka tool contains lots of algorithms to do the research. We have used WEKA tool to analyze the river water quality data. First, we have upload the lead dataset which contains water quality site name, river name, data of water taken from the river and the numerical value of the lead metal in the river. Totally there are 67 observations in the lead dataset that are all exceeded limit of lead content in the river water. In the dataset 12 data were taken from the period of Feb 2012, 26 data from June 2012, 12 data from Nov 2011, 10 for the period of Oct 2012 and 4 for the period of Mar 2013. The initial lead data sets are loaded into the wake tool is in figure1 and the arff file is in figure 2.

Clustering methods: The clustering algorithm divides the huge data into a group of similar objects. We have used four different algorithms in this paper to analyze the dataset. They are

- Simple KMean Clustering Algorithm
- EM Clustering Algorithm
- Cobweb Clustering Algorithm
- Hierarchical Clustering Algorithm

3. RESULT AND DISCUSSION

Weka tool is used to cluster the river water quality data is in figure 3. In the diagram 4, X axis represent the period of river water taken to monitor the status of the lead metal and y represent the value of the lead.

Time taken to build a model for training set is 0.03 seconds in EM clustering algorithm and the incorrectly classified instance is 42.0 and the percentage is 62.6866 %. The visualization of lead data using an EM clustering algorithm is in figure 4. Time taken to build a model for full training data is 0.03 seconds in Cobweb clustering algorithm and the incorrectly clustered instances is 61.0 and percentage 91.0448 %. The visualization of the cobweb clustering algorithm is in the figure 5.

The KMean clustering algorithm takes Time to build a model for full training data is 0.02 seconds and incorrectly clustered instances is 47.0 and the percentage is 70.1493 %. Similarly Time taken to build model the full training data is 0.02 seconds and incorrectly clustered instances is 38.0 and the percentage is 56.7164 %. The figure 6 and 6 contains the highest lead river contamination of the river Ramganga in June 2012 and the lead amount is 48.92 is found in the cluster 1. We have compared the time taken to build the model using training dataset. From

that analysis the Simple KMean and Hierarchical Clustering method takes less time than the Cobweb and EM clustering methods.

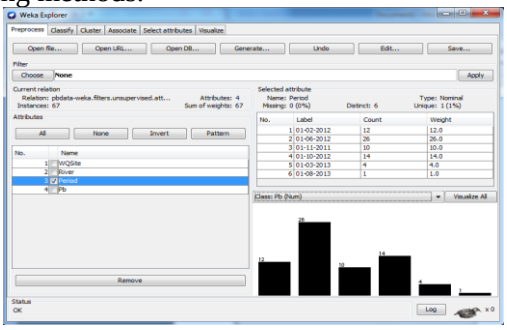

Figure.1.Lead Data in Weka Tool

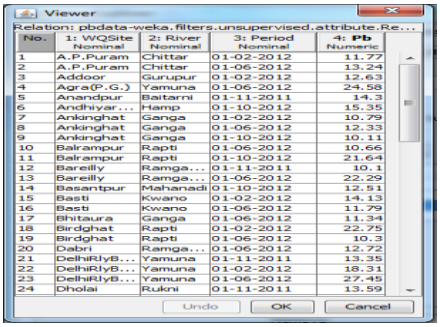

Figure.2.ARFF file of Lead Data

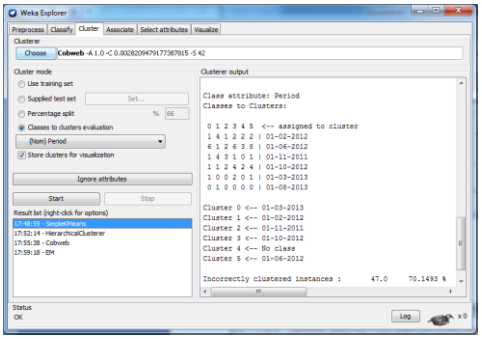

Figure.3.Clustering lead data

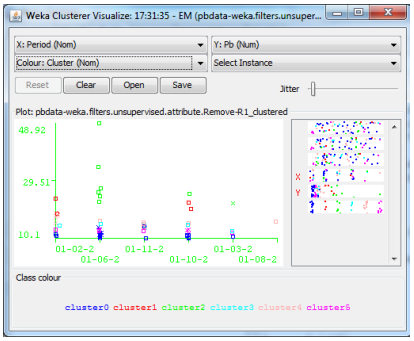

Figure.4.Visualization of lead data using EM

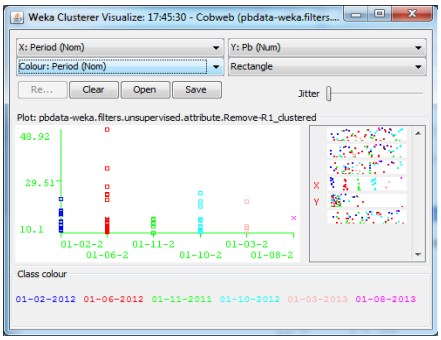

Figure.5.Visualization of lead data in Cobweb

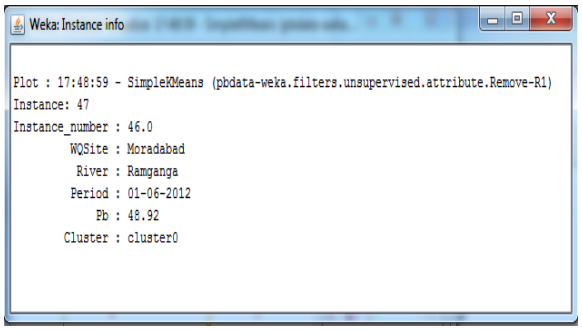

Figure.6.The highest value of lead using SimpleKeans

Table.1.Time and Incorrect classified Instance

Algorithm	Time in Seconds	Incorrect Classified Instance
Cobweb	0.03	91.04%
EM	0.03	62.68%
HierarchicalClusterer	0.02	56.71%
SimpleKMeans	0.02	70.14%

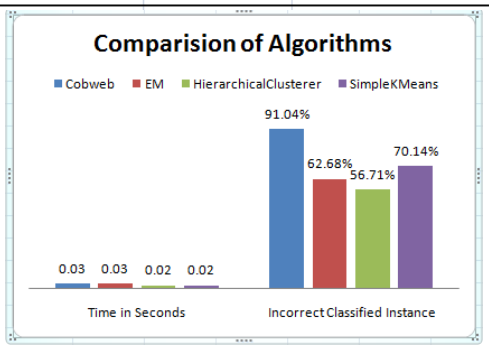

Figure.8.Comparison of algorithms

4. CONCLUSION

Any country river is the main source of water. Degradation of river water quality affects the health status of human beings. Lead metal causes reduce hemoglobin and kidney damage. We have applied data mining technique to predict the value of lead in Indian River water quality dataset. We have used cobweb, EM, simpleKMean and Hierarchical clustering algorithm to group the data and compared the time taken to cluster the lead data and also displayed the highest lead in river water quality dataset.

REFERENCES

Bhardwaj RM, Water Quality Monitoring In India-Achievements And Constraints, Iwg-Env, International Work Session on Water Statistics, Vienna, 2005, June 20-22.

Kumar Manoj, Pratap Kumar Padhy and Shibani Chaudhury, Study of Heavy Metal Contamination of the River Water through Index Analysis Approach and Environmetrics, Bull. Environ. Pharmacol. Life Sci.; 1(10), 2002, 07 – 15.

Sukhdev Kundu, Categorization of Pollution Load in surface Water System using Multivariate Techniques,Advances In Bioresearch, 3(1), 2012, 64 - 68.

Juahir H, Zain SM, Yusoff MK, Hanidza TIT, Samah MAA, Toriman ME and Mokhtar, Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. Env. Monit. Ass. 173(1-4), 2010, 625-641.

Kazi TG, Arain MB, Jamali MK, Jalbani N, Afridi HI, Sarfraz RA, Assessment of water quality of polluted lake using multivariate statistical techniques: A case study. Ecotox Environ Safe, 72, 2009, 301-309.

Omo-Irabor OO, Olobaniyi SB, Oduyemi K, Akunna J, Surface and ground water quality assessment using mutivariate analytical methods: a case study of the Western Niger Delta, Nigeria. Phys Chem Earth, 33, 2008, 663-673.

Ouyang Y, Evaluation of river water quality monitoring stations by principal component analysis. Water Res, 39, 2005, 2621-2635.

Status of Trace and Toxic Metals in Indian Rivers, River Directorate River Data Directorate Planning Organisation Planning & Development Organisation, May 2014.