Air Pollution Analysis using Clustering Algorithms

[1]D.Sathya,[2] J. Anu, [3]M. Divyadharshini
[1]*Assistant Professor II, Department of Computer Science and Engineering, Kumaraguru College of Technology Coimbatore*
[2,3]*UG Final Year, Computer Science and Engineering, Kumaraguru College of Technology Coimbatore*

Abstract-*Air pollution has enormous influence on the amount of constituents in the atmosphere that leads to effects like global warming and acid rains. The main cause of pollution is hazardous gases from traffic system with a large number of private vehicles. In order to help the people to find the best healthy area in the city that are suitable for living, analysis of Air Pollution data is very important. Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. In the paper, the K-means clustering algorithm and Expectation Maximization (EM) clustering algorithm are applied on air pollution datasets generated from The City Pulse project. The paper helps the viewer in determining the actual level of pollution in different locations by position. The paper also gives the comparison of K-means and EM Clustering algorithm using the air pollution dataset. The clustering algorithms discussed in the paper are implemented in WEKA platform.*

Keywords– *Likelihood, Clustering, Air Pollution***.**

I. INTRODUCTION

Air pollution data is not only huge in volume, containing about 500 observations and eight attributes. City Pulse[1] aims to provide real time solutions to interlink data from IoT and associated social networks and to achieve real- time information for the maintainable and smart city applications. The impact of local air pollution on the environment and human health has been studied. With different climate situation, (effected by the wind, temperature, pressure, humidity, etc.), the pollutants pose different air qualities. When human beings expose to the contaminated air when driving in heavy traffic, nearby the freeways or at the 'downwind' areas, those people may suffer breathing problems and asthma attacks, which will lead to risk of heart attacks among people with heart disease.

The existing works of Air Pollution Analysis have implemented different data mining techniques. The classification technique has been used to forecast the Air Quality in the city. The association technique has been used to determine the consequences of Air Pollution. In the paper, clustering technique is used to determine the healthy and unhealthy area in the city.

The harmful effects of air pollution substances on human health are listed below, Ozone (O3), particulate matter (PM10 and PM2.5) and sulphur dioxide (SO2).

A. Ozone (O3)

Scientific research shows that low-level ozone not only affects people either with impaired respiratory systems or healthy adults and children as well. Exposure to ozone (O3) for even short duration, at low levels, certainly reduces lung function and promotes respiratory inflammation in normal. It can be determined by symptoms such as nausea, coughing, chest pain and pulmonary congestion. Results from researches on animal shows that frequent exposure to high concentration of ozone (O3) for longer period of time can lead to lung damage.

B. Carbon Monoxide (CO)

It will get dissolved in the bloodstream and reduces oxygen supply to the body's organs and tissues. The health treatment from CO is most important for those who suffer from cardiovascular disease. Also, healthy people are affected, but only when exposed to higher concentrations. Exposure to high CO concentration is related with visual impairment and reduced work capacity, and poor learning ability.

C. Sulphur Dioxide (SO2)

The major health threats associated with exposure to higher levels of SO2 include effects on respiratory diseases, breathing, aggravation of existing heart related disease, and alterations in pulmonary defences. Major subclasses of the people that are most sensitive to SO2 include asthmatics and people

who are infected by cardiovascular disease or lung disease (like emphysema or bronchitis).

D. Nitrogen Dioxide (NO2)

Nitrogen oxides are important component in the formation of ozone and may affect both earthy and watery ecosystems. Nitrogen dioxide can aggravate the lungs and lower resistance to respiratory illnesses like influenza. The long term exposure to levels that is much higher than those originally found in the ambient air may lead to raised rate of acute respiratory diseases in infants.

E. Particulate Matter (PM-10 and PM-2.5)

In [8], Main harmful effects on people's health from the emissions of particulate matter are: Effects on breathing and respiratory systems, Lung disorders, cancer and death in early age. The elderly, children, and people, who are suffering from severe lung disease, influenza, or asthma, expected to be specifically sensitive to the harmful effects of particulate matter.

II. RELATED WORKS

In [2], proposed a system with k- means clustering and 5 clusters to cluster 53 year of cluster data from 1951 to 2003 of Air Pressure, Air humidity and dusty days per month. For this purpose thy have used Clementine software. Dusty days are classified into 5 classes and decision rule has been exported between air pressure, air humidity and dusty days of January, February and March of each year and other month of year dusty days.

In [3], proposed a system that performed a comparative study between several air quality forecasting methods and tools and explained the comparison work performed between several statistical methods and classification algorithms, on the basis of their performance for specific air quality time series in Athens, Greece.

In [4], the system presented association rule mining for finding association patterns on the various air pollutants. For this, Apriori algorithm of association rule technique in data mining is used. As association rule mining produces several sequence rules of contaminants, the system design has enhanced the reproducibility, selectivity and reliability of air pollution sensor output.

In [5], a system applied the K-means clustering algorithm on air pollution data level of pollution depending on available datasets generated from The City Pulse project and also it provided the level of pollutants in various locations of city by using the longitude and latitude coordinates. Polluted locations has been determined which helped in getting smart environment.

In [6], in order to comply with requirements of oil and gas industry, an air quality monitoring system was proposed based on ZigBee wireless sensing technology. It uses ZigBee wireless network to send results to the monitoring centre so that, if some abnormal situations happens, a quick warning will be generated to remind staff to take effective measures to prevent major accidents and protect human lives in industry.

III. PROPOSED WORK

In proposed system, the air pollution data set of 500 values are used which consists of attributes like ozone, sulphur dioxide, nitrogen oxide, particulate matter. Several clustering algorithms [7] are used to analyze the input data. The security and privacy of data is important and the data should be transmitted within short time period. Here the two algorithms are implemented to compare the shortest computation time. Based on the result, best algorithm is chosen. Two Algorithms are

- K-MEANS
- EM

A. *K-MEANS*

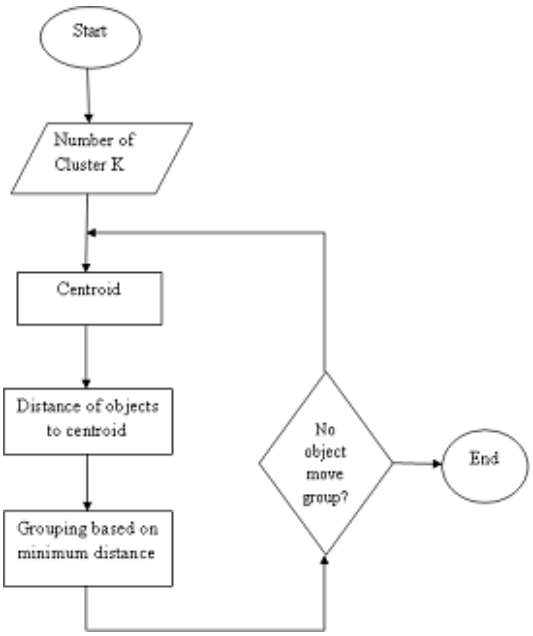

Fig.1 K-means algorithm

The k-means algorithm [9] gets the data set D and parameter k as the input, and then divides a data set D of n objects into k groups. This partition depends upon the similarity measure so that the result obtained will have similarity within the cluster as high but the similarity between the clusters is low. Cluster similarity is measured using the mean value of the objects in a cluster, which can be showed as the cluster's mean. The k-means algorithm works as follows. First, it randomly selects k of the objects, each of which initially defined as a centre or cluster mean. For each of the remaining objects, an object is moved to the cluster to which the similarity is more, based on the similarity measure which is the distance between the item and the average of the cluster. It then for each cluster calculates the new mean. This process repeats until there is no change in the mean values in the clusters.

Algorithm: K-means
Input: *E = {e1, e2,..., en}* (set of objects to be clustered)
K (number of clusters)
Output: *C = {c1, c2,..., ck}* (set of cluster centroids)
L={l(e) | e = 1,2,...,n} (set of cluster labels of E)
Methods:
1. Randomly select k points from the data set D as the initial cluster means (centroids);
2. Assign each object to the group to which is the most closest, based on the means values of the objects in the cluster;
3. Reestimate the mean value of the objects for each cluster;
4. Repeat the steps 2 and 3 until no change in the means values for the groups.

B. EM ALGORITHM

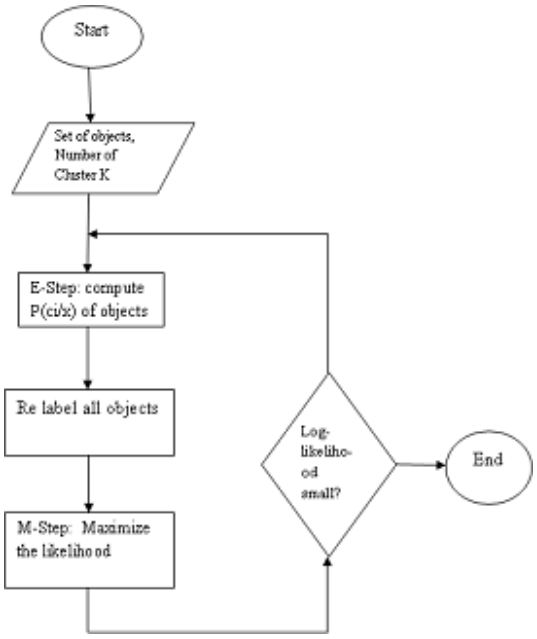

Fig.2 EM Algorithm

The EM algorithm [10] is an iterative algorithm to compute the Maximum Likelihood estimate in the presence of missing or hidden data. Each iteration of EM has two steps: E step and M step. In the expectation, or E-step, the missing data are estimated provided the observed data and current estimate of the model parameters. In the maximization or M-step, the likelihood function estimated on the E step is maximized with the assumption that the missing data are known.

Algorithm: EM clustering
Input: *E = {e1, e2,..., en}* (set of objects to be clustered)
k (number of clusters)
x(number of objects)
Output: *C = {c1, c2,..., ck}* (set of cluster centroids)
L= {l(e) | e = 1,2,...,n} (set of cluster labels of E)
Methods:
1. Initialization: Randomly select initial parameters of k distributions.
2. Iteration:
E-step:
(i) Compute the P(Ci|x) for all objects x by using the current parameters of the distributions.
(ii) Re-label all objects according to the computed probabilities.

M-step:

(i)Re-estimate the parameters of the distribution to maximize the likelihood of the objects assuming their current labelling.

3. Stopping Criteria:

At convergence-when the change in log-likelihood after each iteration becomes small.

4. Repeat the step 2 until the stopping condition occurs.

IV. EXPERIMENTAL RESULTS

In [1][5], the Air Quality Dataset have the attributes like Ozone, Sulphur dioxide, Nitrogen dioxide, Carbon dioxide, Particulate Matter, latitude and longitude. The dataset contains numerical values and 500 observations.

The dataset are taken from the CityPulse Project [1]. We have used K-Means and Expectation Maximization Clustering algorithm in Weka platform as it provides the Visualization of the Clusters obtained after the execution of the algorithm.

Fig.3 Comparison on K-Means and EM

Fig.3 shows the potential of k-means to provide equal sized clusters leads to bad results, while EM provide better results by maximizing the likelihood on artificially generated dataset. EM algorithm is good at dealing with outliers compared to k-means algorithm. The results of EM have lesser variation than that of the results of k-means clustering [9]. EM algorithm showed higher accuracy (over 87%) of the results and high speed.

V. CONCLUSION

The data mining techniques have many applications in Air Pollution analysis which makes decision making very effective. We have implemented Clustering algorithm for Air pollution data to locate healthy and unhealthy areas in the city. The proposed system helps in government projects like Smart City. In the future work, more efficient Clustering algorithm can be implemented to make decisions regarding Air pollution data and Comparative study can be done between those algorithms.

References

[1]"CityPulse project" and the URL is http://iot.ee.surrey.ac.uk:8080/datasets.html.

[2] Ebrahim Sahafizadeh, Esmail Ahmadi, "Prediction of Air Pollution of Boushehr City Using Data Mining", 2009 Second International Conference on Environmental and Computer Science, pp. 33-36, Dec. 2009.

[3] Ioannis N. Athanasiadis and Kostas D. Karatzas and Pericles A. Mitkas, "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, August 2006.

[4] Umesh M. Lanjewar1, J. J. Shah, "Air Pollution Monitoring & Tracking System Using Mobile Sensors and Analysis of Data Using Data Mining", International Journal of Advanced Computer Research,Volume-2 Number-4 Issue-6 December.2012.19.

[5] Doreswamy, Osama A.Ghoneim, B R Manjaunath, "Air Pollution Clustering Using K-Means Algorithm in Smart City", International Journal of Innovative Research in Computer and Communication Engineering,Vol. 3, Special Issue 7, October 2015.

[6] Wenhu Wang, Yifeng Yuan, Zhihao Ling, "The Research and Implement of Air Quality Monitoring System Based on ZigBee", 2011 7th International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1-4, Sept. 2011.

[7] Wei Tian1, Yuhui Zheng1, Runzhi Yang2, Sai Ji1 and Jin Wang1" A Survey on Clustering based Meteorological Data Mining" International Journal of Grid Distribution Computing Vol.7, No.6 (2014), pp.229-240.

[8] Fani A. Tzima, Kostas D. Karatzas, Pericles A. Mitkas, Stavros Karathanasis, ―Using data-mining techniques for PM10 forecasting in the metrop olitan area of Thessaloniki, Greece‖, Proceedings of International Joint Conference on Neural Networks, pp. 2752-2757, Aug. 2007.

[9] Study and implementation of K-Means on https://en.wikipedia.org/wiki/K-means_clustering.

[10]Study and implementation of EM on https://en.wikipedia.org/wiki/EM-means_clustering.