International Journal of Advanced Science and Research
ISSN: 2455-4227
Impact Factor: RJIF 5.12
www.allsciencejournal.com
Volume 3; Special Issue 1; March 2018; Page No. 43-45

A study of fuzzy clustering concept for measuring air pollution index

[*1] Savita Vivek Mohurle, [2] Dr. Richa Purohit, [3] Manisha Patil
[1] Lecturer, MIT ACSC, Alandi, Pune, Maharashtra, India
[2, 3] Assistant Professor, MIT ACSC, Alandi, Pune, Maharashtra, India

Abstract
Increase in population leads to increase in human activity which in turn leads to increase in pollution. Not only human activities but also natural processes can both generate air pollution. According to the 2014 World Health Organization, air pollution in 2012 caused the death of around 7 million people worldwide, an estimate roughly, echoed by one from International Energy Agency. The composition of normal dry air should be acknowledge to be is 78.09% NO2, 20.95% of O2, 0.93% of Ar,0.04% of CO2, 0.4% H2O and other gases. But polluted air consists of composition of particulates like NH3, CO2, SO2, NO2, VOC, CH4, PM2, PM10, toxic metals traces like lead, mercury, etc. producing toxic or poisonous gases. These gases are harmful for living begins. Harmful compositions are considered as data points. The pollutant or data points gathered from the atmospheric air data are somewhat non-linear or fuzzy. If number of unwanted composition or gases appears in atmosphere, then those are the outliers in the normal dry air. To remove outlier, data points needs to be clustered so as to group desired data. This paper gives a brief study of a fuzzy clustering concept for measuring pollution index P_i.

Keywords: pollution index, fuzzy clustering, atmospheric air, data point, particulates

1. Introduction
Clustering is a kind of unsupervised classification aimed at grouping a set of data. Clustering is a frequently used method in many fields, including bioinformatics, humanities, social science, information science, and engineering [13]. It helps us to recognize the dissimilar data points from the cluster. The data points in the atmospheric air data may belong to many different clusters. Fuzzy clustering, also referred to as soft clustering is a form of clustering in which each data point can belong to more than one cluster. It is an alternative method to conventional or hard clustering algorithms, which makes partitions of data containing similar subjects. The tendency of adopting machine learning, big data science, and cloud computation in various industries depends on unsupervised learning on data structures to tell the story about consumers' behavior, fraud detection, and market segmentation [1].

2. Literature Review
Jennifer R.Wolch, and Joshua P.Newell in 2014 in their research reviewed the Anglo-American literature on urban green space, especially parks, and compares efforts to green US and Chinese cities. According to authors urban green space strategies may be paradoxical: while the creation of new green space to address environmental justice problems can make neighborhoods healthier and more esthetically attractive, it also can increase housing costs and property values. It can also lead to gentrification and a displacement of the very residents the green space strategies were designed to benefit. Authors say urban planners, designers, and ecologists, therefore, need to focus on urban green space strategies that are ‘just green enough and that explicitly protect social as well as ecological sustainability [2].

Again in 2014, Lina Cao, Xijin Tang applied dynamic topic model (DTM) to explore the changing topics of new posts collected from Tianya Zatan Board of Tianya Club. They also proposed an algorithm to compute the strength fluctuation of each topic. With visualized analysis of the respective main topics in several months of 2012, some patterns of the topics fluctuation on the board were summarized [3].

In 2015, Xiao Feng, QiLi Yajie Zhu, Junxiong Hou, Lingyan Jin, Jingjie Wang presented a novel hybrid model combining air mass trajectory analysis and wavelet transformation to improve the artificial neural network (ANN) forecast accuracy of daily average concentrations of PM2.5 two days in advance. The air mass trajectory was used to recognize distinct corridors for transport of “dirty” air and “clean” air to selected stations. With each corridor, a triangular station net was constructed based on air mass trajectories and the distances between neighboring sites. Their approach shows the potential to be applied in other countries’ air quality forecasting systems. The model was developed from 13 different air pollution monitoring stations in Beijing, Tianjin, and Hebei province (Jing-Jin-Ji area) [4].

In 2016, Jiamin Li; Harold W. Lewis, introduced an innovative threefold intelligent hybrid system of combined machine learning algorithms HISYCOL (henceforth). First, it deals with the correlation of the conditions under which high pollutants concentrations emerge. On the other hand, it proposes and presents an ensemble system using combination of machine learning algorithms capable of forecasting the values of air pollutants. Moreover, their approach improves the accuracy of existing forecasting models by using unsupervised machine learning to cluster the data vectors and trace hidden knowledge. Finally, it employs a Mamdani fuzzy

inference system for each air pollutant in order to forecast even more effectively its concentrations [9].

3. Method

Fuzzy clustering contrasts with hard clustering by its nonlinear nature and discipline of flexibility in grouping massive data. It provides more accurate and close-to-nature solutions for partitions and herein implies more possibility of solutions for decision-making [1]. Air sample can be collected from air quality measuring instruments or sensors for all kind of gases. Low-cost gas sensors get more and more interest in the field of air pollution monitoring [11, 10], in complement with conventional methods such as optical/spectroscopic analyzer. Compared to the reference methods defined in the Air Quality Directive [12, 10], a low cost gas sensor would considerably reduce both installation and maintenance costs and allow larger spatial coverage especially in remote areas. Clusters of air data collected by the instruments like low cost gas a sensor, the data point belongs to more than one cluster and associated with the data point are a membership grade which indicates the degree to which the data point belongs to the different clusters. This mainly happen in fuzzy clustering. The concept of the fuzzy clustering approach is as follows. The approach demonstrates the measuring of pollution index P_i. Fig 1 below show the graphical representation of fuzzy C means clustering approach.

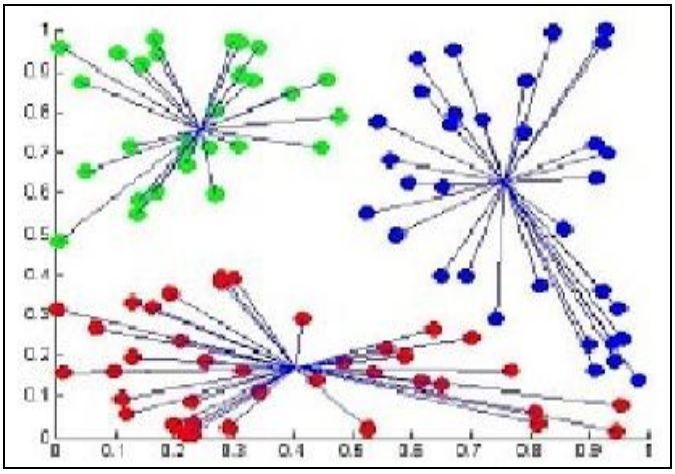

Fig 1: Cluster formations in Fuzzy Clustering

Fuzzy clustering works by assigning membership to each data point corresponding to each cluster center. The membership assigned on the basis of distance between the center of cluster and data point. If the data point is near to center of cluster, then it more a member of that particular cluster that is the membership towards that cluster increases. Addition or summation of all the membership of each data points should be equal to 1. After eachiteration cluster centers and membership is updated according to the given formula:

$$\mu_{ij} = 1/\sum^{c} (d_{ij} = d_{ik})^{(2/m-1)}$$

i=1n is the number of data points. vj represents the j^{th} cluster center.

m is the fuzziness index m € [1, ∞]. c represents the number of cluster center.

μij represents the membership of i^{th} data to j^{th} cluster center dij represents the Euclidean distance between i^{th} data and j^{th} cluster center.

The main objective is to minimize the distance between data point and the center, so as to find the similar data and separate dissimilar data, since in fuzzy clustering on data point can appear in different clusters.

Following are the steps in Fuzzy C-Means Clustering [14]:

1. Randomly select *'c'* cluster centers.
2. Calculate the fuzzy membership $'\mu_{ij}'$ using:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

3. Compute the fuzzy centers $'v_j'$ using:

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, \ldots . c$$

4. Repeat step 2 and step 3 until the minimum *'J'* value is achieved or $||U^{(k+1)} - U^{(k)}|| < \beta$.

Where,

k is the iteration step.
β is the termination criterion between [0, 1].
$U = (\mu_{ij})_{n*c}$ is the fuzzy membership matrix.
J is the objective function.

$X = \{x_1, x_2, x_3..., x_n\}$ be the set of data points
$V = \{v_1, v_2, v_3..., v_c\}$ be the set of centers.

In context to atmospheric air data points, air contains pollutant like PM2 or PM10. Since air data is fuzzy, these pollutants may be present in many clusters. The above Fig. 1 shows three clusters. Therefore c=3. Suppose there are hundreds of data points, among those some may be repeated in many clusters. It becomes necessary to calculate the membership of PM2 or PM10 either of the pollutant by calculating d_{ij} (Euclidean distance between minimizing i^{th} and j^{th} data cluster center). If the air is polluted than it may contain PM2 in every cluster, hence it denotes the degree of belonginess of PM2 which range from 0….1. A fuzzy clustering method can issue good results even for complex data by taking the relative structure for all clusters into account for the clustering results [13].Since atmospheric data found is in three dimensional space, the total number of dissimilarity in the clusters specify the pollution index in the air samples collected, that these data points are not the real composition of dry air. Parameter m specifies degree of fuzziness. Parameter m matter a lot is atmospheric air data points, since it is considered as fuzzy or scattered data. The dissimilarity D is calculated using data points between three clusters. The dissimilarity means data point is outlier to a specific cluster and hence pollution index P_i can be calculated.

4. Conclusion

Air pollution has become the most critical problem now a day. It occurs when harmful substances including particulates and

biological molecules are introduced into Earth's atmosphere. It is very necessary to quantify the extend of pollution being created. This paper tries to tries to co-relate the concept of fuzzy clustering approach with atmospheric air pollution data points collected.

5. References

1. Jiamin LI, Harold W. Lewis. Fuzzy Clustering Algorithms — Review of the Applications, 2882016 IEEE International Conference on Smart Cloud SmartCloud, 2016, (s):282-288.
2. Jennifer R. Wolch, Jason Byrne, Joshua Newell P. Urban green space, public health, and environmental justice: The challenge of making cities 'just green enough', Landscape and Urban Planning, Elevier. 2014; 125:234-244.
3. Lina Cao, Xijin Tang. Topics and trends of the on-line public concerns based on Tianyaforum. Journal of Systems Science and Systems Engineering. 2014; 23(2):212-230.
4. Xiao Feng, Qi Li, Yajie Zhu, Junxiong Hou, Lingyan Jin, Jingjie Wang. Artificial neural networks forecasting of $PM_{2.5}$ pollution using air mass trajectory based geographic model and wavelet transformation, Atmospheric Environment, Elsevier. 2015, 107:118-128.
5. Suganthi L, Iniyan S, Anand A, Samuel. Applications of fuzzy logic in renewable energy systems – A review, Renewable and Sustainable Energy Reviews, Elsevier. 2015; 48:585-607.
6. Bob McKercher, Noam Shoval, Eerang Park, The [Limited] Impact of Weather on Tourist Behavior in an Urban Destination, Journal of Travel Research, 2014; 54(4):442-455.
7. Maria Guadalupe Cortina–Januchs, JoelQuintanilla–Dominguez, AntonioVega–Corona, Diego Andina, Development of a model for forecasting of PM_{10}concentrations in Salamanca, Mexico, Atmospheric Pollution Research, Elsevier. 2015; 6(4):626-634.
8. PierpaoloD'Urso, RiccardoMassari. Fuzzy clustering of human activity patterns, Fuzzy Sets and Systems, Elesvier. 2013; 215(16):29-54.
9. Ilias Bougoudis, Konstantinos Demertzis, Lazaros Iliadil. HISYCOL a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in Athens, Neural Computing and Applications. 2016; 27(5):1191-1206.
10. Laurent Spinelle, Michel Gerboles, Maria Gbriella Villani, Manuel Aleixandre, Fausto Bonavitacola. Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO_2. Sensors and Actuators B: Chemical, Elevier. 2017; 238:706-715.
11. Kumar P, Morawska L, Martani C, Biskos G, Neophytou M, Di Sabatino S, *et al.* The rise of low-cost sensing for managing air pollution in cities, Environ. Int. 2015; 75:199-205. 10.1016/j.envint.2014.11.019
12. Directive 2008/50/EC of the European Parliament and the Council of 21 May 2008 on ambient air quality and cleaner air for Europe and Directive (EU) 2015/1480 of 28 August 2015 on reference methods, data validation and location of sampling points for the assessment of ambient air quality.
13. Tosei Hatori, Mika Sato. A fuzzy clustering method using the relative structure of the belongingness of objects to clusters, 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014. Procedia Computer Science. 2014; 35:994-1002.
14. https://sites.google.com/site/dataclusteringalgorithms/fuzzy-c-means-clustering-algorithm/fuzzy2.bmp?attredirects=0.