INTERNATIONAL JOURNAL OF COMPUTERS AND COMMUNICATIONS
DOI: 10.46300/91013.2021.15.3 Volume 15, 2021

Computational analysis of incremental clustering approaches for Large Data

Arun Pratap Singh Kushwah[a], Shailesh Jaloree[b], Ramjeevan Singh Thakur[c],
[a]Reseach Scholar, Dept. of Computer Science,, BU,Bhopal, Madhya Pradesh, India
[b]Dept. of Appl. Maths & Computer Applications, SATI, Vidisha, Madhya pradesh, India
[c]Dept. of Computer Applications, MANIT, Bhopal, Madhya pradesh, India

Abstract: Clustering is an approach of data mining, which helps us to find the underlying hidden structure in the dataset. K-means is a clustering method which usages distance functions to find the similarities or dissimilarities between the instances. DBSCAN is a clustering algorithm, which discovers the arbitrary shapes & sizes of clusters from huge volume of using spatial density method. These two approaches of clustering are the classical methods for efficient clustering but underperform when the data is updated frequently in the databases so, the incremental or gradual clustering approaches are always preferred in this environment. In this paper, an incremental approach for clustering is introduced using K-means and DBSCAN to handle the new datasets dynamically updated in the database in an interval.

Keywords: Clustering, Incremental Clustering, K-means, DBSCAN, Large Data

I. INTRODUCTION

Extensive use of data collecting devices in different areas forces data storage systems to update dynamically in a certain time interval. This creates large amount of data with many analysis issues, needs researchers attention for creating enhanced data mining techniques. Clustering [1][6] is one of the data mining techniques [32] which can be applied on the new dataset to find and optimize the structure within the data. Incremental Clustering [4][5] is relatively new area of research to handle the new data updated frequently in the databases.

Many clustering approaches have been proposed for limiting the search space and building or updating arbitrary shaped clusters in large incremented datasets, out of which Incremental DBSCAN [4][5] and Incremental K-means[7][8] are two very useful and popular clustering techniques suitable for large dynamic datasets such as IOT dataset, sensor dataset, etc. The performance of the incremental K-means and incremental DBSCAN are different with each other. Both algorithms are efficient compare to their existing algorithms with reference to time, cost and other aspects. This paper introduces the incremental variant of these two clustering methods for large datasets. In this paper data is divided in to multiple subsets so that we can computationally check feasibility with all possible subsets and hence data objects are divided into non overlapping clusters so that each and every object is fixed in exactly in one subset.

II. RELATED WORK

Many attempts have been done for clustering of dynamic datasets to limit the search space, find the initial cluster centroid and to update the arbitrary shaped clusters.

A DBSCAN based incremental density-based clustering method was proposed by [4] using three steps namely sorting, region query for each marked data point and merging on the clusters to identifies the correct number of clusters in a dataset. An incremental DBSCAN based technique was Introduced by [5] for incrementally building and updating clusters in the dataset by incrementally partitioning the dataset to reduce the search space of the neighbourhood to one partition in place whole dataset scan. The problem of converge at a local minimum in respect of K-means has been addressed by [7] by employing an incremental K-means method with jumping technique on artificial and real datasets which enables cluster centres to move in such an efficient way that causes reducing the overall cluster distortion. So that it decreases the dependency on initialization of clusters centres. A Comparative analysis of the performance of both K-means and incremental K-means is performed by [8] to investigate that K-means is suitable with static data whereas its counterpart is found best alias with dataset of incremental nature. A study of recent development in incremental clustering techniques were carried out by [9] emphasized representative algorithms, such as Single-Pass Clustering, Suffix Tree Clustering and ICIB after compared their clustering quality and features. A detailed behavioural study of K-means and Incremental K-means has been carried out by [10] to describe the pitfalls of standard K-means algorithm after analysing the distance with threshold limit and group the object into existing cluster or form a new cluster with that object by presuming synthetic dataset. [11] Proposed a feature based incremental method for clustering the incremental data points by employing K-means clustering algorithm on numeric dataset where Set of Data points have been processed followed by merging of closest pair of clusters based on mean value. The problem of initial starting condition effect has been addressed by [12] by defining degree of neighbourhood object in respect of coupling and cohesion using neighbourhood-based rough set model. A method based on K-means have been explored by [13] by emphasizing no need of predicting of input cluster K in advance.[14] investigated a method based on K-means employing Euclidian distance and Cosine distance functions. An incremental method based on K-means investigated by [15] to determine threshold value as a centroid and Similar or dissimilar Groups of cluster will get created on the basis of Euclidian distance. [16] Investigated a method in periodic incremental environment based on grid and density that deals with dynamic update by decreasing the number of region query. The problems in K-means specially inconsistent behaviour under the same initial conditions and its passive attitude towards large datasets were deeply elaborated by [17] and implemented MapReduce to quickly analyse large datasets and determine good initial seeding in less time. In 2015 a comparative study of four popular clustering algorithms K-Means, Bisecting K-Means, Fuzzy C Means and Genetic K-Means was carried out by [18] to investigate

that the performance of Fuzzy C Means technique is better than other two counterparts and Genetic K-Means outperforms with others as because it searches for the global optima with respect to diversified datasets. G. R. Kingsy, et al., in 2016 [19] proposed a method based on enhanced K-Means clustering to analyse the air pollution data and analyse comparative study of both K-Means and Possibility Fuzzy C-Means(PFCM). The air quality Index and pollutant dataset is monitored by correlation coefficient. [20] have introduced an incremental method on synthetic data based on IDBSCAN to dynamically partitions the dataset and reduce the search space to each partition instead of scanning the whole dataset and detect arbitrary shaped clusters where clusters are defined as dense regions separated by low density regions. A density-based clustering algorithm was developed by [3] for discovering clusters in large spatial databases with noise to find density based clusters of data points in a region. Mai, et al. in 2020 [21] have introduced a unique parallel incremental data clustering approach IncAnyDBC to effectively update clustering results in the environment where data changes frequently. An efficient framework was proposed by [22] to cluster previous summaries with new data for predicting the summaries of previous data directly from data distribution through supervised learning. [23] Developed an ISIF algorithm and proves that the incremental filter based on given approach can effectively detect spam images with high accuracy along with low false positive rate. [24] In 2020 introduced incremental clustering methods perform significantly better in terms of linkage quality when compared with greedy mapping. An incremental approach based on DBSCAN introduced by[25] by underlying suitable fitness functions for both labelled and unlabelled datasets and suggested method to improve the efficiency by parallelization of the optimization process. Baydoun at al. in 2016[37] proposed a method to apply enhanced parallel implementation of the K-Means clustering algorithm.[27] developed an incremental approach that measure the new cluster centres by directly computes the new data from the means of the existing clusters instead of rerunning the K-means algorithm. [28] Proposes an enhanced version of the incremental DBSCAN algorithm for incrementally building and updating arbitrarily shaped clusters in extensive datasets. An incremental method based on DBSCAN was proposed by [29] to cluster the trajectories of moving objects of varying sizes and shapes. [30] Proposes an incremental DBSCAN, which is fused with a suitable noise removal and outlier detection technique inspired by the box plot method. [31] Proposed a method based on DBSCAN using incremental clustering called AMF(Adaptive Median Filtering)-IDBSCAN which builds incrementally the clusters of different shapes and sizes in large datasets and eliminates the presence of noise and outliers.[34] introduced a hybrid framework for air quality prediction based on K-Means clustering and deep neural network.

III. PROPOSED INCREMENTAL CLUSTERING APPROACH

In this paper incremental variant of K-means and DBSCAN clustering algorithms are introduced. Density-based clustering techniques designated to find clusters based on density of data points in a region. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points. Major features of Density Based Spatial Clustering Methods are as follows [3]:

Pros:

(a) It can handle noise efficiently.
(b) It clusters of different shapes and sizes.
(c) It is faster.
(d) There is no need to define number of clusters in advance.

Cons:

a) Varying densities

b) It has some difficulties in distinguishing separated clusters if they are located too close to each other, even though they have different densities.

Many researchers have attempted to overcome certain deficiencies in the original DBSCAN like identifying patterns within datasets of varied densities and its high computational complexity. Hence, a number of augmented forms of DBSCAN algorithm are available.

A. Incremental Dbscan

DBSCAN does not require the number of clusters as a parameter. Rather it infers the number of clusters based on the data, and it can discover clusters of arbitrary shape. The ε-neighborhood is fundamental to DBSCAN to approximate local density, so the algorithm has two parameters:

Step. Ɛ: The radius of our neighborhoods nearby a data point *x*.

Step. *min[x]*: The minimum number of data points in a neighborhood to define a cluster.

DBSCAN clusters the data points using these two parameters, into three categories:

Core records: An entity *x* is a *core record* if *f*(x, *ε*) [ε-neighborhood of *x*] contains *min[x]*; |*f*(*x*, *ε*)| >= *min[x]*.

Edge Records: An entity **y* is an *edge record* if *g*(*y*, *ε*) contains less than *min[x]* data points, but *y* is *reachable* from some *core records x*.

Noise: A data record *r* is a *noise* if it is considered as other class.

The clusters are created when the points are less than edge record but greater than core records and if greater than edge record, the points are considered as outliers.

In incremental step we have calculated the distance of new point from the core point of the existing clusters and if distance is greater than edge points of all the clusters the point is marked as noise else the point is assigned to the existing cluster.

B. Incremental K-Means

An iterative algorithm applied to split the dataset into K pre-defined distinct non-overlapping subgroups called clusters, where each data point belongs to only one group. It makes the intra cluster data points as similar as possible while also keeping the clusters together as far as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. [17].

Incremental K-means algorithm working:

1. Specify number of clusters K.
2. Initialize the centroids by scuffling the dataset followed by random selection of K data points for the centroids without replacement.
3. Keep iterating till changes observed in the centroids assignment of data points to the clusters is not changing.
4. Compute the sum of the squared distance between data points and all centroids.

5. Allocate each data point to the closest-cluster (centroid).
6. Compute the centroids for the clusters the average of all data points that belong to each cluster.
7. Take new data points compare with created cluster centres
8. Define radius (threshold) of each cluster
9. If distance is minimum and within radius assign to the cluster
10. Else create new cluster
11. Repeat 7 to 10 till all data points assigned
12. Stop

K-means has several limitations. The actual K-means algorithm takes lot of time when it is applied on a large database. That is why the incremental clustering concept appears to provide quick and efficient clustering technique on large dataset [22].

IV. EXPERIMENTAL RESULTS

The dataset taken to experiment is the synthetic dataset created using RandomRBF a weka 3.8 function with 2 classes and random set of centers with each cluster. Attributes taken 10 with 1000 and 2000 records.

When DBSCAN applied in 1000 samples, the initial starting points are automatically considered since Farthest First method is internally applied. After algorithm completes its execution, it gives the final centroids (core points) for Cluster 0: 0.749, 0.911 and Cluster 1: 0.964 -0.861. In next step for finding clusters for 2000 new samples, the previous centers (core points) are considered as initial starting point and edge points as threshold (radius). by which it grasps the incremental clustering. Figure 1 presents the clusters created using DBSCAN with 1000 data samples. Blue dots showing cluster 0 and red showing cluster 1.

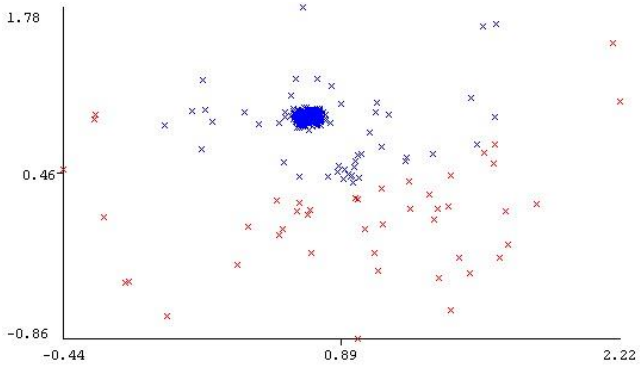

Figure 1 results for DBSCAN with 1000 samples

Classes to Clusters evaluation of DBSCAN with 1000 samples given bellow gives the results with 1.9% incorrectly clustered instances:

```
   0    1  <-- assigned to cluster
 908    0 | a0
  19   73 | a1
Cluster 0 <-- a0
Cluster 1 <-- a1
```

Incorrectly clustered instances: 19.0 1.9 %

In Figure 2 clustered instances shown for DBSCAN with 2000 data samples, which depicts nearly same result as shown in Figure 1 with improved density.

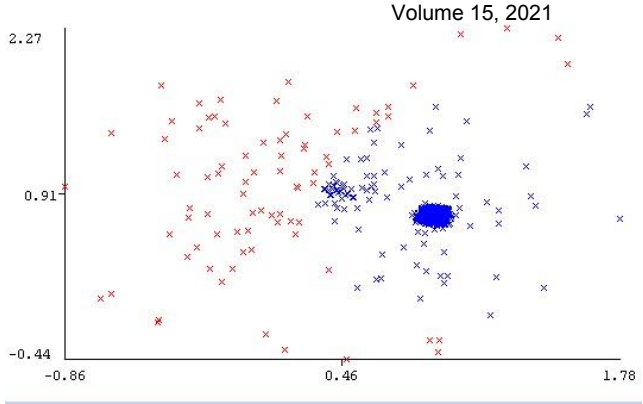

Figure 2 Results of DBSCAN with 2000 samples

Classes to Clusters evaluation of DBSCAN with 2000 data samples shown below, the incorrectly clustered points are 4.65%, which is nominal means the incremental approach working in this case.

```
   0    1  <-- assigned to cluster
1826    0 | a0
  93   81 | a1
Cluster 0 <-- a0
Cluster 1 <-- a1
```

Incorrectly clustered instances: 93.0 4.65 %

When k-Means is applied in 1000 samples, the initial starting points are Cluster 0: 0.726, 0.878 for first attribute and Cluster 1: 0.735, 0.901 for second attribute. After algorithm completes its execution, we have acquired Final cluster centroids for Cluster 0: (a0: 1.050, 0.729) and for Cluster 1: (a1: 0.171, 0.903) in incremental step the samples are updated to 2000 and the initial starting points are updated as previous clusters (a0: 1.050, 0.729) and (a1: 0.171, 0.903) for both attributes for next 2000 samples. Figure 3 shows the clustered points for K-means with 1000 data samples.

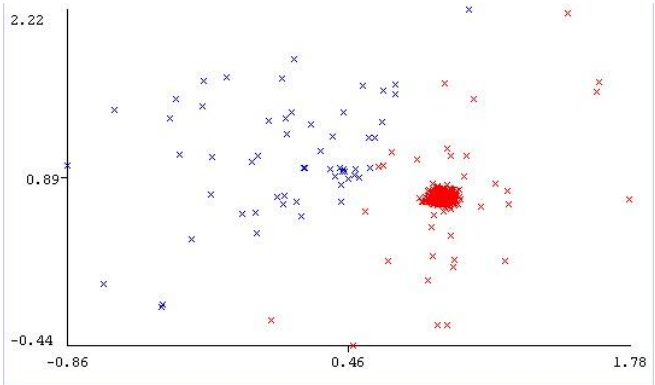

Figure 3 Result of K-means with 1000 samples

Classes to Clusters evaluation of K-means with 1000 data samples shown below:

```
   0        1       <-- assigned to cluster
 908        0 | a0
  57       35 | a1
Cluster 0 <-- a0
Cluster 1 <-- a1
```

Incorrectly clustered instances: 35.0 (3.5%) means only 3.5% points are clustered incorrectly, which is nominal. Figure 4 shows the clustered points for Incremental K-means with 2000 data samples

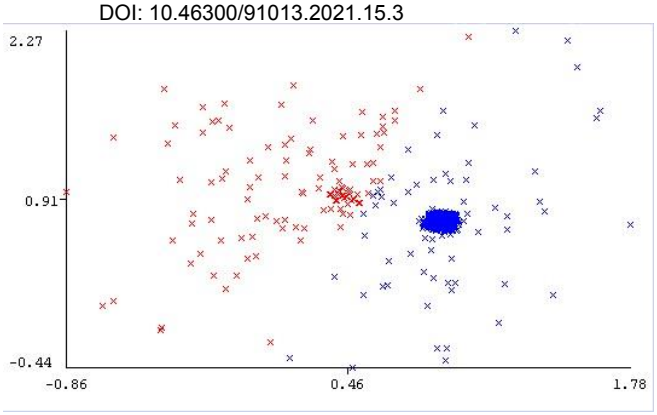

Figure 4 Result of Incremental K-means with 2000 samples

Classes to Clusters evaluation of Incremental K-means with 2000 data samples given below:

```
    0    1  <-- assigned to cluster
1826   0 | a0
   62  112 | a1
Cluster 0 <-- a0
Cluster 1 <-- a1
```

Incorrectly clustered instances: 62.0 (3.1%) which means the incremental approach is working correctly with only 3.1% error.

The overall performance analysis of K-means and DBSCAN is given in the Table 1; we can see that after updating database with new values the incremental K-means and Incremental DBSCAN gives comparative performance. The lower Sum of Squared Error (SSE) 2.97 for Incremental K-means with less time 0.05 and the lower SSE 0.82 for Incremental DBSCAN with less time 0.05 shows that the proposed incremental approach is working and giving comparative results.

Table 1 Performance analysis of clustering methods

Method	Sample/ Update	SSE	Time in Sec.	Iterations
K-Means	1000	5.25	0.08	6
	2000	2.97	0.05	4
DBSCAN	1000	1.24	0.06	3
	2000	0.82	0.05	2

V. CONCLUSION

The frequent changes in the database raise an issue of pattern recognition dynamically. This paper tries to give a solution to clustering for dynamically updating data. The proposed incremental clustering approach worked well on new dataset having less error. More incremental methods for clustering can be created that can handle larger variety of data. These methods can also be applied in stream data for real time pattern recognition.

REFERENCES

[1] Javier Bejar Alonso, "Strategies and Algorithms for Clustering Large Datasets: A Review," *Report* 2013.

[2] M. Kiruthika and S. Sukumaran, "A Survey on partitioning and hierarchical based data mining clustering techniques", *International Journal of Applied Engineering Research*, vol.13. No.24, pp.16787-16791, 2018.

[3] M. Ester, et al., "A density-based algorithm for discovering clusters in large spatial databases with noise" in Proceedings of the *International Conference on Knowledge Discovery and Data Mining (KDD'96)*, United States: AII Press, pp.226-231 1996.

[4] S.U. Rehman and M.N.A. Khan," An Incremental Density-Based Clustering Technique for Large Datasets". *Computational Intelligence in Security for Information Systems*, pp.3–11, 2010.

[5] A. M. Bakr, et al., "Efficient incremental density-based algorithm for clustering large datasets,"*Alexandria Engineering Journal*, vol.54,no.4, pp.1147–1154,2015,

[6] Saroj and Tripti Chaudhary, "Study on Various Clustering Techniques," *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 6 no.3 , pp.3031-3033, 2015.

[7] D. T.Pham, et al., "An Incremental K-means algorithm. Proceedings of the Institution of Mechanical Engineers," in Part C: Journal of Mechanical Engineering Science, vol.218, no.7, pp.783–795, 2004.

[8] Sanjay Chakraborty and N.K. Nagwani, "Performance Evaluation of Incremental K-means Clustering Algorithm," *IFRSA International Journal of Data Warehousing & Mining* , vol.1, no.1, pp.54-61, 2011.

[9] Yongli Liu, et al., "Research on Incremental Clustering," 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet), pp.2803-2806, 2012.

[10] Nidhi Gupta and R.L Ujjwal, "An Efficient Incremental Clustering Algorithm," *World Of Computer Science and Information Technolgy Journal (WCSIT)*, vol. 3, no. 5, pp.97-99,2013.

[11] A.M.Sowjanya and M.Shashi, "Cluster Feature-Based Incremental Clustering Approach(CFICA) For Numerical Data," *International Journal of Computer Science and Network Security(IJCSNS)*, vol.10, no.9, 2010.

[12] Fuyuan Cao, et al., "An initialization method for the K-Means algorithm using neighborhood model," *Computers & Mathematics with Applications*, vol. 58, no. 3, pp. 474-483, 2009.

[13] Anupama Chadha and Suresh Kumar. "An Improved K-Means Clustering Algorithm: A Step Forward for Removal of Dependency on K,"*International Conference on Reliability Optimization and Information Technology (ICROIT)*, Faridabad, India, pp. 136-140, 2014.

[14] S. K. Sunori, et al., "K-Means Clustering of Ambient Air Quality Data of Uttarakhand, India during Lockdown Period of Covid-19 Pandemic*," 6th International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, pp. 1254-1259, 2021.

[15] Md. Hossain, et al., "A dynamic K-means clustering for data mining," *Indonesian Journal of Electrical Engineering and Computer Science(IEECS)*, vol.13, no.2, pp.521-526, 2019.

[16] C. Zhuo, et al., "A Fast Incremental Clustering Algorithm Based on Grid and Density," *Third International Conference on Natural Computation (ICNC 2007)*, Haikou, China, 2007, pp. 207-211, 2007.

[17] R.M. Esteves, et al., "Competitive K-Means, a New Accurate and Distributed K-Means Algorithm for Large Datasets," *5th IEEE International Conference on Cloud Computing Technology and Science*, Bristol, UK, pp. 17-24, 2013.

[18] Shreya Banerjee, et al., "Empirical evaluation of K-Means, Bisecting K-Means, Fuzzy C-Means and Genetic K-Means clustering algorithms," *IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*, Dhaka, Bangladesh, pp. 168-172, 2015.

[19] G. R. Kingsy, et al., "Air pollution analysis using enhanced K-Means clustering algorithm for real time sensor data," *IEEE Region 10 Conference (TENCON)*, Singapore, pp. 1945-1949, 2016.

[20] Y. El-sonbaty, et al., "An efficient density based clustering algorithm for large databases," *16th IEEE International Conference on Tools with Artificial Intelligence*,pp.673-677,2004.

[21] S.T. Mai, et al., "Incremental Density-based Clustering on Multicore Processors", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp.1-1,2019.

[22] X.Zhao, et al., "Incremental face clustering with optimal summary learning via graph convolutional network," *Tsinghua Science and Technology*, vol. 26, no. 4, pp. 536-547,2021.

[23] Ying He, et al., "Incremental clustering-based spam image filtering using representative images,"*International Conference on System science, Engineering design and Manufacturing informatization*, Guiyang, China,pp. 323-327,2011.

[24] D. Vatsalan, et al.,"Incremental clustering techniques for multi-party Privacy-Preserving Record Linkage," *Data & Knowledge Engineering*, vol.128, pp.101809, 2020.

[25] E. Azhir, et al., "An efficient automated incremental density-based algorithm for clustering and classification," *Future Generation Computer Systems*, vol.114, pp.665-678, 2021.

[26] M. Baydoun, et al.,"Enhanced parallel implementation of the K-Means clustering algorithm," *3rd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*, Beirut, pp.7-11, 2016.

[27] D. Vanisri, "A Novel Fuzzy Clustering Algorithm Based on K-Means Algorithm," in *International Review on Computers and Software (IRECOS)*, vol.9 ,no.10, pp.1731, 2014.

[28] A.M. Bakr, et al., "Efficient incremental density-based algorithm for clustering large datasets," in *Alexandria Engineering Journal*, vol.54, issue.4, pp.1147–1154, 2015.

[29] R. Ranjith, et al., "Anomaly detection using DBSCAN clustering technique for traffic video surveillance," *7th International Conference on Advanced Computing (ICoAC)*,pp.1-6, 2015.

[30] S. Wan and Y-P Wang ," The Comparison of Density-Based Clustering Approach among Different Machine Learning Models on Paddy Rice Image Classification of Multispectral and Hyper spectral Image Data," *Agriculture*, vol.10, no.10, pp.465, 2020.

[31] A. Chefrour, and L. Souici-Meslati , "AMF-IDBSCAN: Incremental Density Based Clustering Algorithm Using Adaptive Median Filtering Technique," *Informatica.* vol.43, no.4, pp. 495–506, 2019.

[32] P. N. Tan, et al., Cluster Analysis: Basic Concepts and Algorithms. Introduction to Data Mining, pp.487-568, 2006.

[33] Amit Yadav and Gambhir Singh, "Incremental k-means clustering algorithms: a review," *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, vol.5, no 4,pp.56-59, 2015.

[34] D. Ao, et al., "Hybrid model of Air Quality Prediction Using K-Means Clustering and Deep Neural Network," Chinese Control Conference (CCC), Guangzhou, China, pp. 8416-8421,2019.