Webology (ISSN: 1735-188X)
Volume 19, Number 5, 2022

An Efficient Density Clustering Based Ensemble Classification Learning Model For Large Real-Time Spatial Aqi Database

A.Nageswara Rao[1], Dr. Bendi Venkata Ramana[2]

[1]Research Scholar, Dept of CSE, JNTUK Kakinada, AP,India.

[2]Prof & Dean, Dept of IT, AITAM, Tekkali AP, India.

Abstract
Air quality analysis plays a vital role in the multi-region based severity detection. Since, most of the conventional density based clustering approaches use static homogeneous type of air quality data for severity prediction. However, most of the conventional models are not applicable to dynamic sub-region based cluster analysis for severity prediction. In this work,a novel weighted density inter and intra cluster based ensemble learning approach is developed for air quality prediction process. Experimental results show that the proposed multi-level weighted density based clustering approach has better efficiency for sub-clustering and severity detection process than the conventional approaches.

Keywords: Air quality analysis, multi-level clustering, heterogeneous data samples.

1.Introduction

Spatial pattern mining for air quality detection refers to the process of identifying patterns and relationships in air pollution data collected from different locations in space. The goal is to uncover insights into the distribution and trends of air pollution, which can be used to predict air quality and inform decision-making.The process of spatial pattern mining involves the use of statistical and data mining techniques to analyze large amounts of air quality data. This can include calculating spatial correlation measures to understand the relationship between different attributes of air pollution, such as temperature, wind speed, and pollutant concentrations. The data can also be analyzed using association rule mining and pattern clustering techniques to identify co-patterns of air pollution in different locations.Air pollution is a major environmental issue that has a significant impact on human health and the environment. In order to address this issue, it is important to understand the distribution of air pollution and predict its future trends. Spatial pattern mining can play a crucial role in this regard by providing valuable information about the distribution of air pollution and helping to predict air quality.

The process of spatial pattern mining involves the use of statistical and data mining techniques to analyze large amounts of air quality data. The resulting patterns can provide important insights into the distribution and trends of air pollution. For example, a pattern that shows a consistent relationship between wind speed and pollutant concentrations can help to predict air quality in different locations. This information can be used by decision-makers to take measures to improve air quality, such as reducing emissions from sources of pollution or changing the location of industrial activities.Reducing emissions from sources of pollution is one of the most effective ways to improve air quality. For example, if a pattern shows a high concentration of pollutants in an area that is heavily industrialized, decision-makers can implement measures to reduce emissions from industrial sources, such as stricter regulations, investment in cleaner technologies, or the relocation of industrial activities to less populated areas.Another way to improve air quality is to change the location of industrial activities. If a pattern shows that pollutants are being carried by the wind to areas with high populations, decision-makers can take measures to prevent this, such as relocating the source of pollution or implementing measures to block the wind from carrying pollutants.

The resulting patterns can provide valuable information about the distribution of air pollution and help to predict air quality. For example, a pattern that shows a consistent relationship between wind speed and pollutant concentrations can help to predict air quality in different locations. This information can be used by decision-makers to implement measures to improve air quality, such as reducing emissions from sources of pollution or changing the location of industrial activities.Finding patterns in air pollution in space is challenging. The aim in searching for spatial air pollution co-patterns is to uncover connections between different locations. A spatial air pollution co-pattern refers to a group of features that consistently appear together in space. These co-patterns have various applications, such as predicting weather, air pollution levels, and its changes. Spatial statistics and spatial data mining are effective ways to identify air pollution co-patterns. Spatial statistics calculates the spatial correlation measure to understand the relationship between different attributes and their impact on each other[1]. Spatial data mining utilizes implicit knowledge from spatial databases through association rule mining and pattern clustering. Association rule mining seeks to find answers to difficult mining problems and links cases with their features. Clustering, on the other hand, organizes items based on the similarity of their features into layer-based or mixed clustering. Finding the relationship between spatial objects and events is a crucial area of research in spatial machine learning, and pattern-mining techniques have been used for this purpose. However, most models lack the consideration of limited memory and processing speed and often produce repeating patterns, making it challenging to make decisions based on spatial events[2-3].

Processing and extracting critical patterns from vast spatial data has become a challenge for organizations as spatial technology and storage advance. The difficulty in finding meaningful patterns from these data is increasing. To address this issue, various models have been investigated to find frequent spatial relationships, including join-less models, probabilistic prevalent models, and join-based models in uncertain spatial datasets. However, some of these models can result in a large number of patterns that are difficult to understand and apply[4].

Spatial classification is a useful technique for creating classification models that support various spatial characteristics. It involves analyzing different spatial objects, such as regions and areas, to develop classification models. The goal of spatial classification is to identify classification rules from both spatial and non-spatial attributes.

One approach to spatial classification is the ID3 algorithm, which takes into account both the non-spatial features of the classified instance and the non-spatial features of its neighboring instances. However, this method is not effective in dealing with non-spatial features and cannot analyze hierarchical relationships between spatial and non-spatial attributes.

A modified two-phase approach to spatial classification was later introduced, which involves obtaining the spatial predicate value using a low-cost calculation and then conducting correlation analysis. The final stage involves refining computations to produce a novel decision tree. Despite its small size, this tree is the most accurate one.To address the limitations of traditional spatial classification, a new method called Rough Set-based spatial classification has been suggested[5-6].

2.Related works

To achieve spatial instance classification, the rough set approach's fundamental principles are necessary. It effectively distinguishes between spatial and non-spatial relationships. Every data point contained inside a specific class with a fixed scope must only contain data points that meet the minimum threshold. Only those data points can be included in the clustering process that meet the minimum threshold value. Construction of clusters with any shape is most effectively accomplished using a density-based approach. It also has the ability to separate noise from information that is useful. The following are some of the widely used density-based techniques: DBSCAN scheme, OPTICS scheme, GDBSCAN scheme, DBRS scheme, DENCLUE scheme, etc. They used the external library GeoSpark to implement the framework they had developed above in the Apache Spark environment. This method's main goal is to effectively process large amounts of spatial data. Through the Leaflet JavaScript library, the co-patterns of generated air pollution are visualised on a map. For each and every transaction derivation, a variety of different techniques must be used, including spatial partitioning, reference features, and DBSCAN clustering. The primary goal of the FP-growth scheme is to produce intriguing and reliable air pollution co-patterns. With the aid of the framework mentioned above, performance evaluation is also done to improve the processing of spatial big data. Future work will be required to combine the aforementioned framework with a web-based interface. When Apache Spark is used as the backend, many enhanced web applications can be created. It will significantly enhance all of the benefits of the technique suggested above. As there are many different types of space data involved, they were able to pinpoint the real problem with co-location pattern detection. Many redundant patterns are also created during this co-location pattern generation process. These redundant patterns increase computation time and take up unnecessary space. More intriguing information can be retrieved for long-sized patterns. Therefore, mining longer patterns is preferable to mining smaller patterns[7]. Here, a sophisticated method for identifying top-k-size maximal co-location patterns is presented. Additionally, a brand-new datastructure called the MCP tree is presented. To create top-k-size maximal co-location patterns, it is not necessary to produce all candidate co-

locations. To save time and space, it can be done by only looking at partial candidates[8]. A specific transformation sequence is used to transform each programme. Every single pattern candidate has a predefined transformation sequence specific to it. They put their theory into practice using an image integration application. Comparing the results of the evaluation process to other traditional mining approaches that had previously been developed, it produces better performance and efficiency. An ordered neighborhood method for mining co-location patterns is used for static air pollution dataset. The maximal clique enumeration approach is the foundation of the above-mentioned technique. To successfully mine spatial boolean data, none of the conventional mining techniques can be used. This method's experimental results generate every single maximal clique in a synthetic dataset. Additionally, the proposed method performs better than a join-based approach. Here, they have examined various techniques for applying traditional frequent itemset mining methods to spatial datasets. An innovative method for accurately identifying spatial co-location patterns was proposed in [9]. They presented a hybrid technique in this research study that incorporates all the advantages of conventional, earlier methods. This method uses a collection of instances to identify every single spatial co-location pattern. To begin, a neighbourhood detection process is used to locate each unique co-location pattern with a size of 2. The execution of the aforementioned procedure is aided by a spatial access technique known as KD-tree. Tree generation and search queries are just two of KD-unique tree's features. Here, an effective pruning technique known as the participation index is used. This participation index's main duty is to eliminate all non-prevalent patterns. Other co-location patterns with higher sizes will be discovered after size 2 co-location patterns have been successfully identified. To produce these co-location patterns, a conventional join method is combined with another sorting method. The KD-tree structure can be compared to various other index structures in the future. The method mentioned above can be improved in order to raise the system's overall effectiveness. To find regional co-location patterns, they created a sophisticated method [10]. This method effectively extracts a set of continuous variables from spatial datasets. Detecting regions is the main goal of the approach discussed above. These areas are designed to co-locate multiple continuous variables. They created a brand-new framework that can function in a continuous domain. Regional mining co-location is regarded as a significant clustering issue. The external fitness function is significantly maximised in this situation. The z-scores of the pertinent continuous variables are taken into account to assess how interesting co-location patterns are. They used a variety of experimental evaluation techniques that are adequate for identifying both known and unidentified regional co-location patterns. The idea of randomised hill climbing is included in the CLEVER approach, a more successful prototype-based method for finding regions. Additionally, it looks for larger neighbourhood sizes and cluster sizes that vary. In exploratory data analysis, the data owner can "view" a data set in terms of condensed statistical parameters and graphical display to "get a feel" for any patterns or trends present in the data set through interactive and visual techniques. In essence, descriptive modelling creates more elevated "views" of a data set. In addition to having models that describe the relationship between variables, it also includes determining the overall probability distributions of the data (dependency modeling). The data must also be divided into groups using segmentation or cluster analysis. Segmentation is slightly different from cluster analysis. While segmentation looks for homogeneous groups

connected to the variable to be modelled, clustering algorithms look for "natural groups." Regression and classification are both used in predictive modelling. Given that extended objects like points, lines, and polygons—which are typically stored as BLOBS, or Binary Large OBjects—are included, the data inputs for spatial data mining are also more complex than the data inputs for traditional data mining. The spatial attribute and non-spatial attribute are two distinct types of attributes present in the data inputs for spatial data mining. Non-spatial attributes, which are used to characterise non-spatial features of objects like name, population, or the unemployment rate for a city, are essentially database-type attributes. They resemble the attributes that are used as data inputs in traditional data mining. Any discipline that analyses multivariate data frequently uses cluster analysis. In 2010, the term "data clustering" appeared in 11,700,000 result entries, according to a Google Scholar search. The significance of clustering in data analysis and Data Mining in general is discussed in this extensive literature. One of the most fundamental forms of intelligence is the capacity to organise meaningful groups of objects. Humans, even young babies, are remarkably adept at performing this task. One learns to differentiate between, say, apples and bananas or plants and animals in early childhood. But making the computer capable of automatically grouping is a challenging and frequently ill-defined issue. The numerous scientific disciplines and applications that have used clustering techniques are extensively listed. Image segmentation is a crucial issue in computer vision, and it is simply formulated as a clustering issue. Unsupervised classification, also referred to as cluster analysis, categorises unknown groups without the use of a training set or prior knowledge of the groups. Analyzing discriminant functions classifies recognised groups. Due to the use of some domain knowledge, the discriminant function analysis process is fairly straightforward. Voting Methodology based Using a direct approach or re-labeling method are other names for ensembles. The correspondence problem between the labels of known and derived clusters does not need to be explicitly solved in the other type of algorithms. A simple voting result can be used to reassign objects in clusters to determine the final consensus partition after the voting approaches have first solved the correspondence problem. Label correspondence is the specific factor that makes unsupervised combination challenging and difficult. Here, the main goal is to permute the cluster labels in order to achieve the best possible agreement between the labels of two or more partitions. The ensemble's partitions must all be relabeled using a consistent, user-defined reference partition. Usually, the ensemble clustering or a fresh clustering of the data set is used to determine which partition to use as the reference. Then, a meaningful voting process assumes that each partition has the same number of clusters as what must be present in the target partition[11-14].

Path-based clustering with automatic outlier detection was also presented in [15]. It captures the empirical finding that, with a reasonable extension, group structures in embedding spaces might provide the necessary variations. The main issue is that local connectivity and homogeneity characterise it. Even in situations where the parametric form of such a transformation is unknown, path-based clustering is still useful. This is based on the fact that path based clustering has already achieved success in solving two key problems of perceptual organisation, edge grouping and texture segmentation.

A different consensus function has been created by [16] using information theoretic principles and generalised mutual information (MI). It was established that, in the specially

transformed space of labels, the underlying objective function is equivalent to the total intra-cluster variance of the partition. Therefore, in such a space, the k-means algorithm can find corresponding consensus solutions quickly. As the components of the combination, they proposed two different weak clustering algorithms. The first one is a clustering of random 1-dimensional projections of multidimensional data, and it can be generalised to clustering in any random subspace of the original data space. By clustering and dividing the data using a number of random hyper planes, a second weak clustering algorithm is produced. For instance, data is divided into two groups if only one hyper plane is used. This generalisation allows for clustering in any arbitrary data space subspace. Although this algorithm has a low computational complexity, it requires a few restarts to prevent convergence to local minima. Another probabilistic model of consensus using finite mixture multinomial distributions in the cluster labels space was presented by [17]. The Expectation Maximization Algorithm was used to find a combined partition as a resolution to the corresponding maximum likelihood issue (EM). With regard to the parameters of the corresponding finite mixture distribution, the likelihood function of an ensemble is optimised. The target consensus partition's target clusters are represented by each component of this distribution. They completely avoid dealing with the challenging label correspondence problem with their strategy. For clustering of large data sets, this algorithm's excellent scalability and understandable underlying model are crucial. However, if sample size is fixed, accuracy degradation will unavoidably happen as the number of partitions increases. Their method worked with missing data, such as labels for missing clusters or labels for specific ensemble patterns. For instance, when the ensemble is generated using the bootstrap method, there may be some circumstances in which the label is absent. Their method works with arbitrary partitions that have different numbers of clusters—some of which may not be equal to the desired number in the consensus partition[18].

3. Proposed Model

In recent years, an increasing number of sensor devices have generated a large amount of temporal data which can be treated as time series data. Extracting numerical features from time series data would have a huge influence for human decision, such as revealing human interpretable characteristics of the human activity data, data forecasting for social behaviors as well as clustering and classification. According to the definition of AQI, ambient air pollutants are concentrations of particulate matter (PM2.5 and PM10), SO2, CO, NO2, O3. These time series are nonstationary and seasonality with high level of noise and outlier. Nonstationary means that the statistic properties change with time. Many studies have analyzed air quality time series in the stationary framework, however, this assumption is invalid. Air quality time series should be pre-processed to deal with the negative impact caused by noise and outlier.

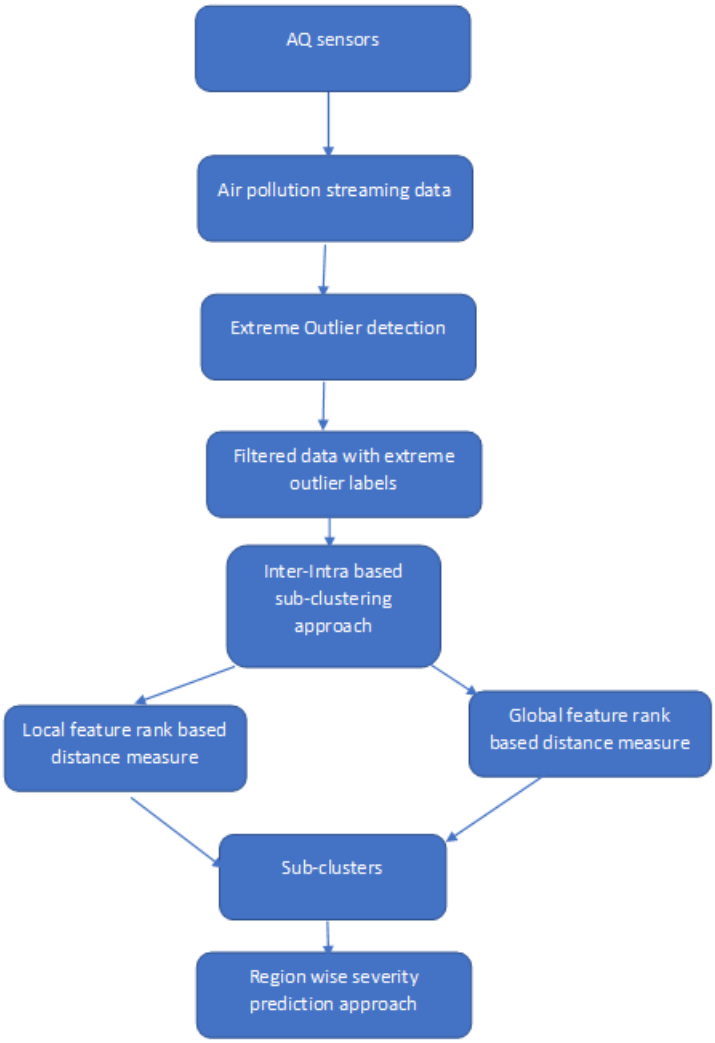

Figure 1: Proposed framework

Traditional time series clustering and statistic feature extraction are unable to overcome the patterns of noise and outlier. In order to solve these problem, a novel algorithm need to be proposed. Among all techniques applied to analyzing time series data, clustering is the most widely used one without costly human supervision or time-consuming annotation of data. Clusters are formed by grouping objects which have maximum similarity in an unlabeled dataset. According to nonstationary time series, feature-based representations of time series are universal. These features are statistical values for stationary time series with low dimensions. For a specific dataset, the selection of feature vector and suitable distance measure is a very challenging task, and these two measures can extremely affect the results of time series clustering. In the proposed work, a hybrid multi-level air quality severity prediction framework is implemented on the large air pollution data. Here, the real-time air pollution data is taken from the realtime sensors deployed in the various places of india. Theses sensors feed the various air pollution data records in periodic manner. These data samples are taken from the india national air quality index website https://app.cpcbccr.com/AQI_India/.

In the Figure 1, different levels of filtering and prediction approaches are designed on the input air quality dataset. A weighted density based clustering model is implemented on the

filtered data for data classification process. Finally, an ensemble learning model is applied on the clustered data for air quality class prediction as shown in Figure1.

Hybrid Weighted Density based intra and inter sub-clustering approach

Step 1: Read dataset

Step 2: To each data point p[i] in D

Step 3: Perform weightage density to each data object using the Gaussian transformation function as

$$wd_i = \sum_{j \in L, i \neq j} \exp(-(d_{ij} / d_c)^2)$$

$$d_{i,j} = \sqrt{\sum_{t=1}^{m} \left(x_i^t - x_j^t\right)^2}$$

Step 4: To each object, find the highest density using the following measures

$$hd_j = \max\{wd_i\}$$

Step 5: Computing the k nearest neighbors by using the k-randomized centres as k initial clusters as.

The set of k nearest neighbors of center CPi is defined

$$NP_i^k = \{P_j / \min(d_{ij}), i! = j\}$$

Step 6: Compute the inter cluster similarity and intra cluster similarity to each k-neighbor initial clusters using the following formula.

$$IntraClu(p_c, p_i) = \frac{1}{n_i - 1} hd_c \cdot \sum_{m=1} d(p_c, p_m)$$

$$InterClu(p_c, p_i) = \min_{1 <= m <= k} \left(\frac{1}{n_m} hd_c \cdot \sum_{r=1} d(p_c, p_r)\right)$$

Step 7: Iterate until all points are assigned to k=clusters or no more changes in clusters.

Ensemble learning model

Let the input dataset is represented as $ID = \{In_1, In_2, In_3, \ldots In_n\}$ be the given dataset

Set of proposed classifiers and base classifiers are represented as ensemble classifiers as

$EC = \{HKNN, HDT, SVM\}$,

$MC = \{\}$; / / Model classifier

CO={};Classifier output

for j = I to l

do

if (j > 1)

s(j) = Set of wrongly classified records of jth model MC(i) on S(h)

$CO = CO \cup CC(j); //CC(j): \text{Correctly classified instances}$

end if

done

HKNN

Input : Clustered training data D, test samples T, k value and classes C.
Output: class prediction
Procedure: To each instance p in D

do

Compute lognormaldistance$(D(t,p)/t \in D, p \in D, t != p) = \log(\sum (||ti|| - ||pj||)^2)$;

done

Sort k neighbors according to their distances.

sort(k, D(t,p))

Compute probability distributions using the neighbor distances.

To each instance t(i) in k-neighbors(sort(k, D(t,p)))

do

$DistProb[] = \frac{1}{\sqrt{2.\pi}} \int e^{-D(t(i),p)^2} dD(t(i),p)/|N|; N = \text{Total attributes}$

done

To each test sample t in k-neighbors(sort(k, D(t,p)))

do

Compute class membership probabilities of each test sample t

assign class to t sample using classifier.

done

HDT: Optimized decision tree using random forest

To each attribute in A[]

do

Finding rank of the attribute as

s=D.log(D);

$p1 = -s/((\sqrt{\sum D[i]})^3 * \sqrt{chisqr(D) * \sum D[i]}$

$p2 = -(s*condentropy(D))/(chisqr(D) * \sum D[i])^3$

Rank(A[i])=Max{p1,p2}

done

4.Experimental results:

Experimental results are performed on the multi-region air quality dataset taken from the realtime databases. These data samples are taken from the india national air quality index website https://app.cpcbccr.com/AQI_India/. The sample data of the india air pollution with different locations are given in Table 1.

No.	1: id	2: country	3: state	4: station	5: last_update	6: pollutant_id	7: pollutant_min	8: pollutant_max	9: pollutant_avg	10: city
	String	Nominal	Nominal	Nominal	Nominal	Nominal	String	String	Nominal	Nominal
1	1	India	Andhra_Pradesh	Secretariat, Amaravati - APPCB	23-02-2020 12:00	PM2.5	20	98	46	Amaravati
2	7	India	Andhra_Pradesh	Secretariat, Amaravati - APPCB	23-02-2020 12:00	OZONE	20	63	43	Amaravati
3	99	India	Delhi	Aya 0gar, Delhi - IMD	23-02-2020 12:00	NO2	20	46	29	Delhi
4	104	India	Delhi	Bawa0, Delhi - DPCC	23-02-2020 12:00	NO2	20	94	44	Delhi
5	225	India	Delhi	Okhla Phase-2, Delhi - DPCC	23-02-2020 12:00	NO2	20	89	46	Delhi
6	228	India	Delhi	Okhla Phase-2, Delhi - DPCC	23-02-2020 12:00	CO	20	149	87	Delhi
7	1	India	Andhra_Pradesh	Secretariat, Amaravati - APPCB	23-02-2020 12:00	PM2.5	20	98	46	Amaravati
8	7	India	Andhra_Pradesh	Secretariat, Amaravati - APPCB	23-02-2020 12:00	OZONE	20	63	43	Amaravati
9	99	India	Delhi	Aya 0gar, Delhi - IMD	23-02-2020 12:00	NO2	20	46	29	Delhi
10	104	India	Delhi	Bawa0, Delhi - DPCC	23-02-2020 12:00	NO2	20	94	44	Delhi
11	225	India	Delhi	Okhla Phase-2, Delhi - DPCC	23-02-2020 12:00	NO2	20	89	46	Delhi
12	228	India	Delhi	Okhla Phase-2, Delhi - DPCC	23-02-2020 12:00	CO	20	149	87	Delhi
13	439	India	Harya0	Sector-2 IMT, Manesar - HSPCB	23-02-2020 12:00	CO	20	101	58	Manesar
14	470	India	Harya0	F-Block, Sirsa - HSPCB	23-02-2020 12:00	PM2.5	20	44	33	Sirsa
15	673	India	Madhya Pradesh	Bhopal Chauraha, Dewas - MPPCB	23-02-2020 12:00	CO	20	60	46	Dewas
16	968	India	Rajasthan	Adarsh 0gar, Jaipur - RSPCB	23-02-2020 12:00	NO2	20	78	36	Jaipur
17	971	India	Rajasthan	Adarsh 0gar, Jaipur - RSPCB	23-02-2020 12:00	CO	20	59	37	Jaipur
18	1084	India	Uttar_Pradesh	Yamu0puram, Bulandshahr - UPPC...	23-02-2020 12:00	CO	20	125	99	Bulandshahr
19	1128	India	Uttar_Pradesh	A0nd Vihar, Hapur - UPPCB	23-02-2020 12:00	PM2.5	20	89	37	Hapur
20	1164	India	Uttar_Pradesh	Pallavpuram Phase 2, Meerut - UP...	23-02-2020 12:00	PM2.5	20	309	86	Meerut
21	1244	India	West_Bengal	Ballygunge, Kolkata - WBPCB	23-02-2020 12:00	NO2	20	112	57	Kolkata
22	1	India	Andhra_Pradesh	Secretariat, Amaravati - APPCB	23-02-2020 01:00	PM2.5	20	98	46	Amaravati
23	99	India	Delhi	Aya 0gar, Delhi - IMD	23-02-2020 01:00	NO2	20	46	30	Delhi
24	104	India	Delhi	Bawa0, Delhi - DPCC	23-02-2020 01:00	NO2	20	94	44	Delhi
25	225	India	Delhi	Okhla Phase-2, Delhi - DPCC	23-02-2020 01:00	NO2	20	89	47	Delhi
26	228	India	Delhi	Okhla Phase-2, Delhi - DPCC	23-02-2020 01:00	CO	20	148	96	Delhi
27	437	India	Harya0	Sector-2 IMT, Manesar - HSPCB	23-02-2020 01:00	CO	20	101	64	Manesar
28	648	India	Kerala	Plammoodu, Thiruva0nthapuram - ...	23-02-2020 01:00	PM2.5	20	180	74	Thiruva0nth...
29	671	India	Madhya Pradesh	Bhopal Chauraha, Dewas - MPPCB	23-02-2020 01:00	CO	20	60	47	Dewas
30	968	India	Rajasthan	Adarsh 0gar, Jaipur - RSPCB	23-02-2020 01:00	NO2	20	78	37	Jaipur

Table 1: Sample multi-region air quality dataset

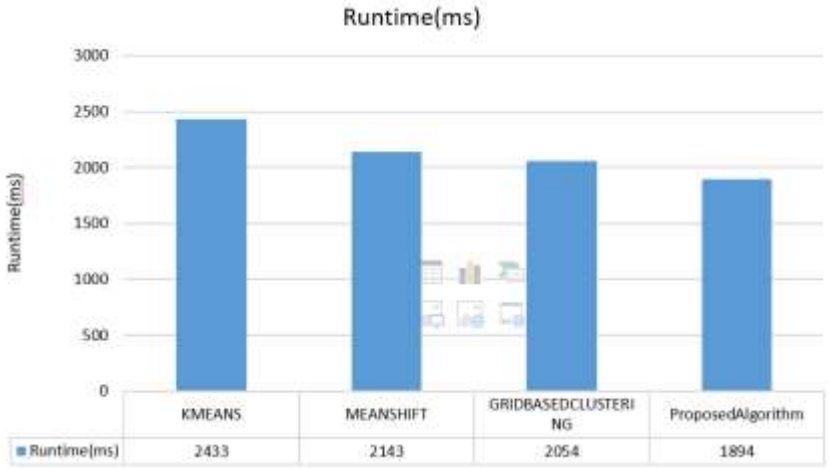

Figure 2: Comparative analysis of proposed model to the conventional models for runtime computation.

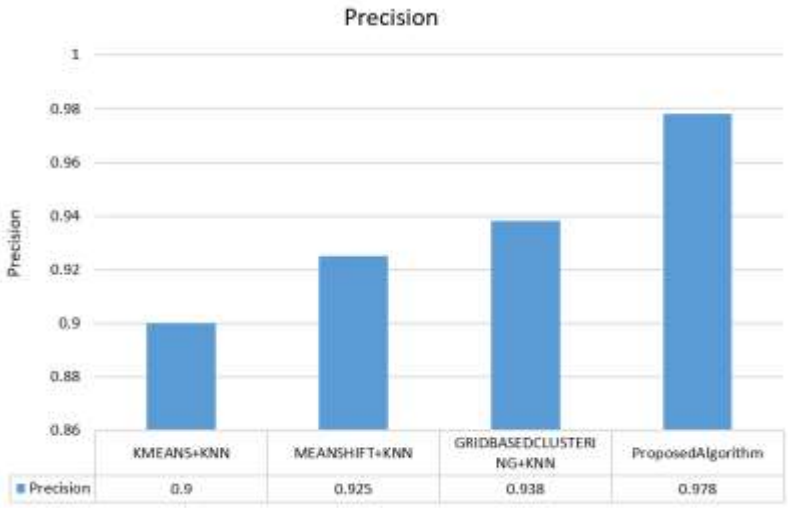

Figure 3: performance analysis of proposed approach to the conventional models for precision rate.

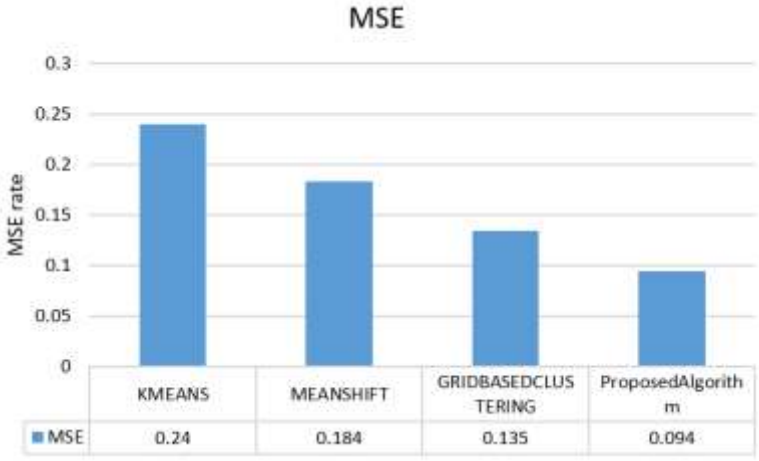

Figure 4: Comparative analysis of proposed multi-variate cluster based classification model to the conventional models for mean square error rate.

5.Conclusion

In this paper, a novel inter and intra based multi-level clustering and classification framework was proposed on the air quality severity detection process. Since, most of the traditional approaches are difficult to predict the region wise severity detection due to the variation in data samples and feature types. In this work, a novel weighted density inter and intra cluster based ensemble learning approach is developed for air quality prediction process. Experimental results show that the proposed multi-level weighted density based clustering approach has better efficiency for sub-clustering and severity detection process than the conventional approaches.

References

[1] H. Ahn, J. Lee, and A. Hong, “Urban form and air pollution: Clustering patterns of urban form factors related to particulate matter in Seoul, Korea,” Sustainable Cities and Society, vol. 81, p. 103859, Jun. 2022, doi: 10.1016/j.scs.2022.103859.

[2] W. Alahamade, I. Lake, C. E. Reeves, and B. De La Iglesia, “A multi-variate time series clustering approach based on intermediate fusion: A case study in air pollution data imputation,” Neurocomputing, vol. 490, pp. 229–245, Jun. 2022, doi: 10.1016/j.neucom.2021.09.079.

[3] M. Asgari, W. Yang, and M. Farnaghi, “Spatiotemporal data partitioning for distributed random forest algorithm: Air quality prediction using imbalanced big spatiotemporal data on spark distributed framework,” Environmental Technology & Innovation, vol. 27, p. 102776, Aug. 2022, doi: 10.1016/j.eti.2022.102776.

[4] R. Borge, D. Jung, I. Lejarraga, D. de la Paz, and J. M. Cordero, “Assessment of the Madrid region air quality zoning based on mesoscale modelling and k-means clustering,” Atmospheric Environment, vol. 287, p. 119258, Oct. 2022, doi: 10.1016/j.atmosenv.2022.119258.

[5] A. Caron, N. Redon, P. Coddeville, and B. Hanoune, “Identification of indoor air quality events using a K-means clustering analysis of gas sensors data,” Sensors and Actuators B: Chemical, vol. 297, p. 126709, Oct. 2019, doi: 10.1016/j.snb.2019.126709.

[6] Y. Chen et al., “Air quality data clustering using EPLS method,” Information Fusion, vol. 36, pp. 225–232, Jul. 2017, doi: 10.1016/j.inffus.2016.11.015.

[7] S. De and B. Chakraborty, “An energy-efficient wireless sensor network construction algorithm for air quality condition detection system,” Computers & Electrical Engineering, vol. 91, p. 107064, May 2021, doi: 10.1016/j.compeleceng.2021.107064.

[8] C. Deng, H.-C. Choi, H. Park, and I. Hwang, “Trajectory pattern identification and classification for real-time air traffic applications in Area Navigation terminal airspace,” Transportation Research Part C: Emerging Technologies, vol. 142, p. 103765, Sep. 2022, doi: 10.1016/j.trc.2022.103765.

[9] Y. Geng, W. Ji, Y. Xie, B. Lin, and W. Zhuang, “A sub-sequence clustering method for identifying daily indoor environmental patterns from massive time-series data,” Automation in Construction, vol. 139, p. 104303, Jul. 2022, doi: 10.1016/j.autcon.2022.104303.

[10] B. W. Hobson, H. B. Gunay, A. Ashouri, and G. R. Newsham, “Clustering and motif identification for occupancy-centric control of an air handling unit,” Energy and Buildings, vol. 223, p. 110179, Sep. 2020, doi: 10.1016/j.enbuild.2020.110179.

[11] M. Hulkkonen, A. Lipponen, T. Mielonen, H. Kokkola, and N. L. Prisle, “Changes in urban air pollution after a shift in anthropogenic activity analysed with ensemble learning, competitive learning and unsupervised clustering,” Atmospheric Pollution Research, vol. 13, no. 5, p. 101393, May 2022, doi: 10.1016/j.apr.2022.101393.

[12] Y.-C. Lin, H.-S. Shih, and C.-Y. Lai, "Classification of air quality zones and fine particulate matter sensitive areas by risk assessment approach," Environmental Research, vol. 215, p. 114208, Dec. 2022, doi: 10.1016/j.envres.2022.114208.

[13] I. A. Sakellaris, J. G. Bartzis, J. Neuhäuser, R. Friedrich, A. Gotti, and D. A. Sarigiannis, "A novel approach for air quality trend studies and its application to european urban environments: The ICARUS project," Atmospheric Environment, vol. 273, p. 118973, Mar. 2022, doi: 10.1016/j.atmosenv.2022.118973.

[14] M. Shaheen, S. ur Rehman, and F. Ghaffar, "Correlation and congruence modulo based clustering technique and its application in energy classification," Sustainable Computing: Informatics and Systems, vol. 30, p. 100561, Jun. 2021, doi: 10.1016/j.suscom.2021.100561.

[15] P. Srividya and L. N. Devi, "An optimal cluster & trusted path for routing formation and classification of intrusion using the machine learning classification approach in WSN," Global Transitions Proceedings, vol. 3, no. 1, pp. 317–325, Jun. 2022, doi: 10.1016/j.gltp.2022.03.018.

[16] J. Torres-Serra, A. Rodríguez-Ferran, and E. Romero, "Classification of granular materials via flowability-based clustering with application to bulk feeding," Powder Technology, vol. 378, pp. 288–302, Jan. 2021, doi: 10.1016/j.powtec.2020.09.022.

[17] T. Wang, H. Du, Z. Zhao, A. Russo, J. Zhang, and C. Zhou, "The impact of potential recirculation on the air quality of Bohai Bay in China," Atmospheric Pollution Research, vol. 13, no. 1, p. 101268, Jan. 2022, doi: 10.1016/j.apr.2021.101268.

[18] X. Wang, L. Wang, Y. Liu, S. Hu, X. Liu, and Z. Dong, "A data-driven air quality assessment method based on unsupervised machine learning and median statistical analysis: The case of China," Journal of Cleaner Production, vol. 328, p. 129531, Dec. 2021, doi: 10.1016/j.jclepro.2021.129531.