International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Efficient Filter Based Weighted Prior and Posterior Clustering Framework for Air Pollution Prediction

Atmakuri Krishna Chaitanya [1], K. V. Prasad [2], Dr Ravi Aavula[3], Ch.Lakshmi Kumari [4], Dr.Ch.V.Phani Krishna[5] , Dr. Nallamala Sri Hari[6]

Submitted: 27/01/2024 **Revised**: 05/03/2024 **Accepted**: 13/03/2024

Abstract— The problem of aggregating similar air quality indicators due to sample variability arises when the number of datasets detailing air pollution continues to grow. In order to estimate comparable severity levels out of scant and ambiguous data, traditional techniques of air quality clustering rely on static markers. Locating both local and global data samples for the purpose of severity categorization is an additional problem involved with the single cluster metric-based clustering approach that is currently being used. For the purpose of data clustering, this study presents a hybrid outlier identification model. The goal of this model is to determine which air quality characteristic values are the most severe. We recommend use a clustering technique that considers both local and global measures in order to classify the data that has been filtered. In the end, it all boils down to utilizing the categorization learning model in order to make a prediction regarding how poor the air quality would be. An outlier-based data clustering and classification framework has been shown to perform better than the conventional methods for predicting air quality, according to the results of an experiment conducted on a time series air pollution dataset.

Keywords— Clustering, Air Quality Index, Classification, Outliers

I. INTRODUCTION

In the area of industrial intelligence, anomaly detection plays a vital role and has a great contribution. One of the most useful features of anomaly detection is outlier detection, and the purpose of outlier detection is the detect abnormal or unusual observation. Outlier detection is an important technique nowadays in the industrial intelligence field. In recent decades, several findings implemented outlier detection for different sectors, like stock prediction, intrusion detection, diseases identify and tracing, Autism spectrum disorder (ASD), etc. Outlier detection's main motive is to identify abnormal phenomena over data using specific values which vary from common values[1][2]. Outlier detection is the branch of deep learning, and it is known as the unsupervised approach of deep learning. The reason is that outlier detection is unable to identify between the denser clusters in low-density areas and is mainly used in anomalies[3]. Recently, research scholars implemented various paradigms and frameworks to detect outliers for multiple fields. Through the observation over data, several outlier methods exist that enable to detection of abnormality.

The well-known outlier framework is the density-based LOF model; this model is very efficient and recommended by many implementors.

The type of problems that happened due to the skewness of data is mostly solved by LOF[4], and scholars use this type of framework. The algorithm neighbor-based algorithm, distance, and many more are examples of LOF. The problem is that such techniques can't work efficiently because such methods assume that input data required has the same distribution law. Furthermore, such algorithms follow general rules globally for outlier detection[5]. Another classification algorithm called one classification was introduced by developers and it can be applied to various applications. Such methods are used in multiple problems and their solutions in the field of abnormalities detection[6]. Outlier detection methods are classified into two types: univariate and multivariate methods. The initial form of the outlier approach was univariate, while the most recent form utilized by academics is multivariate. For the statistical data type, outlier detection systems are combined with another parametric category, but for the free models, the non-parametric form is used[7]. Researchers face a huge challenge to implement outlier detection due to its unavailable labeled dataset. The major problem is that

1Institute of Aeronautical Engineering, Dundigal, Hyderabad, India;
ORCID ID: 0000-0002-7081-0477
2Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India
ORCID ID: 0000-0002-5305-4155
3Guru Nanak Institutions Technical Campus
ORCID ID: 0000-0001-8640-1152
4Mahatma Gandhi Institute of Technology, Hyderabad
ORCID ID: 0009-0003-9129-2466
5Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad
ORCID ID: 0000-0001-8452-3877
6Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt.,
ORCID ID: 0000-0003-2462-948X
Corresponding Author Email: prasad_kz@yahoo.co.in

those outlier occurrences are too few, despite the fact that that scientists developed several unstructured approaches for outlier detection. The following are the requirements for finding anomalies: (i) examine attribute values sets; (ii) identify instances that deviate from the configurations. As a result, providing a consistent statistical measurement technique for variation that pertains to all data and conditions is difficult[8]. Moreover, due to the unstructured nature of the data, there appears to be a contradiction between statistically unusual occurrences and real-world instances of significance to consumers. All throughout centuries, a wide range of research papers have received considerable attention. On the other hand, as technology advances at a rapid rate in various sectors (e.g., equipment, electronic devices, Internet services, medical equipment, the structure of the economy, etc.), data quantities and complexity rise. The existing research studies reviewed mostly unsupervised anomalies identification. Conclusively, the unsupervised outlier detection model is a greatly analyzed approach for the detection of outliers. To find the abnormal patterns that were not examined previously, as the labeled data for outlier detection is often a challenging task to trace the abnormal patterns. Recently, the researchers have suggested ways to consider user feedback while developing unsupervised outlier detection models. Thus, many researchers concluded that there is a need to develop purely unsupervised algorithms. Moreover, recent researches have also introduced semi-supervised models. There is a massive amount of data that can be derived from outliers in the form of application domains. Machine learning and statistical learning theory (SLT) will be applied to a labelled training data set. Techniques that build a specific prognosticative model require a training data set. An information instance's labels indicate whether or not it is a traditional or outlier. Outlier detection methods are often supervised or unsupervised because these labels are readily available or used. A training data set with labelled examples of outliers is assumed to be available for use in supervised outlier detection techniques. A prognosticative model is built for each normal and outlier category in such techniques. The two models are compared to see if any previously unseen information falls into one of the two categories. An unsupervised outlier detection technique does not assume that labelled training data will be available. As a result, a common pattern is considered normal, while a rare occurrence is considered an anomaly. They describe several clustering methods that appear in a wide range of fields. In contrast, the majority of traditional clustering algorithms use either numerical or categorical data as the basis for their clustering. Using clustering algorithms, data is grouped into clusters based on their shared characteristics and a high degree of similarity. A wide range of fields, including pattern recognition, computing and medicine, rely on clustering. In order for clustering to be successful, the type of information and application space must be considered. As a result of its extreme deviation, an outlier is considered to have been generated by a separate mechanism from the other observations. Clustering can be used to detect outliers; outliers are defined as observations that don't fit into the general clustering pattern. A variety of approaches to finding k-Means outliers and alternative clustering algorithms are discussed in this section to deal with noise and produce successful results. It has been proposed by the authors , that the stream can be divided into chunks, and each chunk can be clustered using k-median into a variable number of clusters. This method is only applicable to numerical datasets. Outlier Removal Clustering (ORC) was proposed to simultaneously identify outliers and information clusters. With the help of both clustering and outlier detection, this strategy improves its ability to estimate the distribution's centroids. Outliers can be identified using the shortest distance technique, which was proposed by [9]. In this rule, outliers are identified by measuring how far they are from the rest of the information set's data. Algorithm-supported clustering approaches for spotting outliers have been proposed by [10]. The PAM clustering algorithm is used as the first step in this algorithmic rule. Outlier clusters are those that are extremely small in comparison to the rest of the data.

K-Means (KM) is one of the most commonly used clustering algorithms because of its high computational efficiency[11]. The initialization process, which generates the initial clustering, is well known to lead to a local optimum in KM, and the result depends on this process. Different KM runs on the same input data can yield different outcomes. " Algorithms based on natural evolution are being developed. In general, a new population is generated based on the principle of "survival of the fittest" after an initial population has been established. This population's fitness is typically measured in terms of distance, which is the most common method of determining how good this population. Crossover is then carried out over the new population, in which substrings from selected pairs are swapped out for each other. Selection is based on the fitness of both pairs, with a preference given to the fittest pairings[12]. When random points from each cluster are assigned to another cluster, then mutation may occur. For each generation, this process repeats until the fitness of that generation reaches a predetermined threshold or a predetermined number of generations. Genetic algorithms for clustering have been proposed by a number of researchers. Natural evolution can be replicated in an effort to find better solutions over time. The initialization process is not a factor in these genetic algorithms, which always find the global optimum. The problem is that these algorithms are usually computationally intensive. In 1998, Bradley et al. proposed a method for improving the initial points of clustering algorithms, particularly the K-Means clustering algorithm. In order to refine an initial clustering algorithm starting point, they presented a quick and competent algorithm[13]. Clustering problems involving thousands of high-dimensional points can be easily solved using these techniques. [14] developed a method

for making the K-Means algorithm more adaptable to different types of data. The K-Means algorithm, in general, performs well when it comes to clustering data sets. In order to cluster real-world data using K-Means, the algorithm must be applied to only numerical values.

II. RELATED WORK

To introduce an efficient outlier detection method, the research study conducted by Liu et al., (2020) suggested a high volume dataset outlier detection method named KOF (KDE-based Outlier Factor). The presented model was based on KDE (Kernel Density Estimation) approach. In this study, the KOF model was first introduced to find out the local outlier score for the data set. Then the UKOF method (the Upper-bound and pruning-based model) was proposed to handle the data updates. Moreover, the lazy update version of UKOF was also introduced for large-scale updates, which was named LUKOF. This study used ten realistic datasets to analyze the proposed model. The study concluded that the proposed approach produced better results in comparison to other already introduced methods. Specifically, the proposed method performed 3600 times better in speed than other models. Generally, there is always an issue with the financial time series dataset that always shows high volatility and outliers. Thus, to detect the structural breaks in financial time series, especially interval-valued based, the study was conducted by Cappelli et al., (2021), in which the A Theoretical regression trees approach was used. The employed approach has the edge over other approaches in that it automatically detects the occurring structural breaks for unidentified dates. This study used a time series dataset comprised of daily prices of the American International Group spanning from January 03, 2005, to December 18, 2018. The empirical results of the study demonstrated that the presented approach was effective in detecting the structural breaks and outliers in IVTS datasets. This approach had the edge over other outlier detection models as this framework was faster, automatic, and easy to implement. While dealing with large-scale, high-volume data streams, detecting outliers is one of the critical and challenging tasks, especially in the field of data mining and machine learning. There are numerous applications of an outlier detection model in the real world as it can be applied in detecting faults and frauds in industrial settings. It can also be used for recognizing patterns and can also be used in the field of finance and economics. Thus, various studies were conducted to develop an efficient and robust outlier detection algorithm that performs better than the state-of-the-art approaches.

In industrial automation, detecting outliers is a critical and challenging problem. To overcome the limitations of already state-of-the-art methods, a study was conducted in which an effective method was presented, which was named as CELOF algorithm. This proposed approach was presented to overcome the main two limitations of previously presented models. The first limitation was that a large memory was consumed to store the data. The other limitation was the poor detection of outliers in high-dimensional data. Moreover, the study results indicated that the CELOF algorithm was on average 15% more accurate in performance than other earlier methods. Additionally, the results concluded that the CELOF framework consumes 1% lesser time than the original LOF. Thus, it was suggested by this research that this model could be used without getting prior information and data distribution (Chen et al., 2021).

For multivariate functional data, Lejeune et al. (2020) presented another outlier detection method, which primarily focused on multivariate functional depth to detect the outliers. This research study introduced a detection method that could be performed on curve shape-based multivariate functional data. To assess the performance of the method, this study employed synthetic and realistic datasets. Another research was conducted to provide a robust outlier detection method that efficiently spots the outliers. This study employed the KNN (k-nearest neighbor) rule to develop each test pattern's validation set by using the UCI repository dataset. This approach accurately computes the competencies with neighbor trends in this area than the traditional KNN algorithm. A probabilistic model was employed to measure the classifier competencies, which used posterior probabilities of the one-class classifier. The empirical results of the study indicated that the proposed approach outperforms various static ensemble methods and single frameworks (Wang & Mao, 2020). Furthermore, another study attempted to introduce such a method that provides robust results while using the raw sea clutter and simulated data. Still, this model is difficult to implement in general as this model consumes more time in computations (Yu et al., 2021).To address the existing issues of previously presented models, this study was conducted in which a novel approach was introduced which uses simple reduction for outlier detection to classification. The empirical findings suggested that the proposed method is more robust than the other existing approaches regarding computations problems and explanatory potential (Abe & Zadrozny, n.d.).

III. PROPOSED MODEL

The purpose of this study is to provide a strategy for anticipating the severity of air quality issues by applying a hybrid multi-level framework to the huge amounts of data pertaining to air pollution. The data presented here on air pollution in India comes from sensors that are located all throughout the country and are updated in real time. There are numerous records that detail air pollution, and these sensors contribute data to those records at regular intervals. It is possible to access the official website of the air quality index in India, which can be accessed at https://app.cpcbccr.com/AQI_India/. This information was obtained from that website.

As can be seen in Figure 1, the dataset containing the air quality in the input is placed through a number of different stages of filtration and prediction algorithms. We begin by filtering the data samples and locating the

outliers by employing the severe outlier detection methodology that was given during the first stage. A comparison is made between the training data and two extremely extreme limits, one at the very top and one at the very bottom. This comparison is used by the outlier identification method to filter the training data for the most likely outliers. After that, we use the multi-variate clustering algorithm to incorporate the data that has been filtered. The clustering approach that has been provided identifies class-wise grouping on the dataset that has been cleaned up by utilizing a combination of posterior and prior predictions that are weighted by probability combinations. Lastly, in order to provide an accurate prediction regarding the severity of the test samples, a clustered dataset concerning the classification problem is provided.

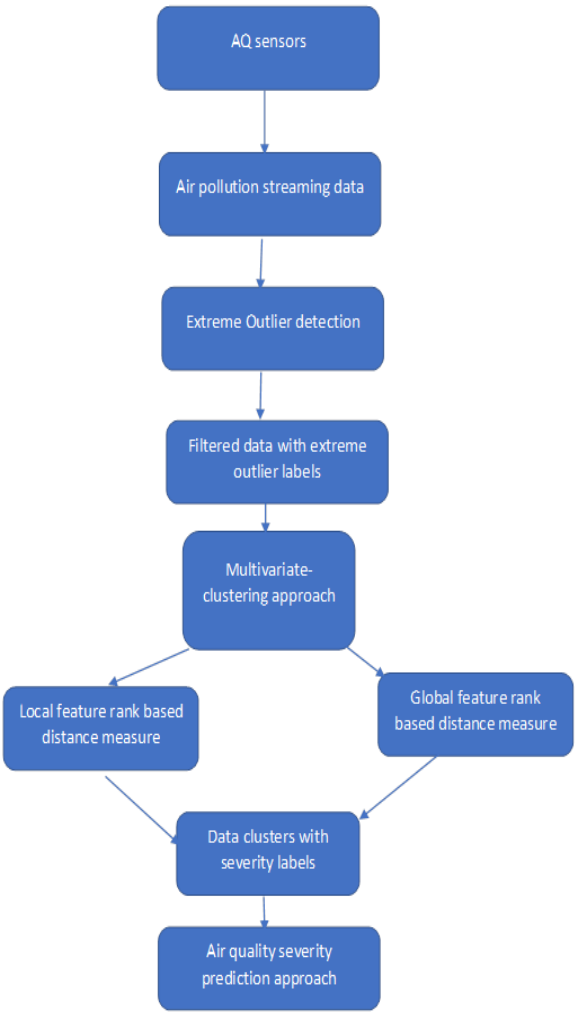

Fig 1. Proposed local and global probability based clustering framework

Algorithm 1: Proposed Extreme range outlier detection

1. Read the training data D.
2. Compute the following hybrid IQR method as

$F[] = AttIndices();$

$P1 = V(|F|/4));$

$P2 = (V(|F|/2) + V(|F|/2+1))/2;$

$P3 = (V(|F|-|F|/4-1) + V(|F|-|F|/4))/2;$

$U\_E[] = P3 + \eta.log(\Gamma(P3-P1))$

$L\_E[] = P1 - \eta.log(\Gamma(P3-P1))$

$U\_Outlier = P3 + \eta.max(chisquareprob(\lambda 3+\lambda 1,9)).log(\Gamma(P3-P1)))$

$L\_Outlier = P3 + \eta.max(chisquareprob(\lambda 3+\lambda 1,9)).log(\Gamma(P3-P1)))$

3. for i=1 to |A|
   do
4. if((V(F[i])<=U_E[i]&&V(F[i]>U_Outlier[i]||(V(F[i])>=L_E[i]&&V(F[i]<L_Outlier[i])
5. Assign instance I as Anomaly .
6. d(|F|+1)=1;
7. else
8. d(|F|+1)=0;
9. done
10. Perform proposed classification model on the filtered dataset for data prediction.

The first method presents the proposed model for identifying extreme outliers at both low and high levels, which are essential for forecasting severity. This model is implemented in the first algorithm. This approach allows for the identification of outliers by applying the computations of extreme values that were provided in step 2 to each attribute respectively. In the second stage, the outlier regions, the extreme lower limit, and the extreme upper limit are established by making use of the input training data. Steps three and four are reflective of the necessary requirements that must be met in order for the anomaly detection to filter out the out-of-the-ordinary ranges. In the sixth through eighth phases, the outlier labels for each feature are represented as a class label.

Algorithm 2. Prior and Posterior Weighted Density based clustering approach

Step 1: Compute min and max value of each attribute.

Step 2: // filter data initialization and cluster initialization

Initialize best cluster objects using k-means as initial clusters CK.

Step 3: Initial k centers

Cn_k, initial cluster standard devaitions $C_{\sigma(k)}$, cluster sizes Cs_k

Step 4: To each cluster in K(number of clusters)
do
To each attribute At in attribute space AS[].
if(At is not numerical)
then
ModelNom[]=AddParams(k, At,
WeightedFreqCount(At));
end if
else
ModelNum[]=AddParams($V(Cn_i)$, $\min(C_{n(k)})$
end loop // attributes
initPriorProb(k)= Cs/ Sum(Cs) // Compute and assign prior probability using cluster size.
Step 5: Compute log likelihood estimation of each training data by using following procedure
To each instance in training data TS.
do
Compute the exponential joint density estimation values to each instance using the prior probabilities as initial weights as $w[i] = w[i] + \exp(2 * \sqrt{Cs[i]})$ --1
Compute the squared joint probability estimation to the instance as instance updated weight in the model

construction phase.

for(m = 0; m < | w |; m++)
do

$$S[m]=\log(\exp(\sqrt{w[i]-MaxInd(w[i])})$$

R[m]=R[m]+S[m] ---(2)

done
Normalize(S[m],R[m])
Computing the proposed log likelihood estimation is given as

$$PLLE = rMaxInd(w[i]) * MaxInd(w[i] + \log(S[m])) \quad ---(3)$$

Step 6: Finally, the estimated clusters are maximized using the following updated weights and prior probabilities in the model optimization phase.

for(p = 0; p < | TS |; p++)
do
for(q=0;q<k;q++)
do

$$PostProb[q]=w[i] * e^{S[m]} * \min(\sigma_A) \quad ---(4)$$

done
done
Step 7: Repeat steps 5-6 till N number of iterations.

In the second algorithm, we see how to use joint probability and log likelihood estimates to construct a weighted prior and posterior estimation based clustering strategy. The first step is to use the k-means technique to identify k clusters. For the purpose of estimating probabilities, these k-clusters are considered to be randomly distributed clusters. Here, the weights utilized in the probability estimation method are initially set using eq (1). It is possible to modify the clustering process's weights using Eq (2). During optimization, the posterior estimate is calculated using Eqs. (3) and (4) in order to maximize cluster quality.

Algorithm 3. Proposed Classification model

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

$$\text{Compute lnd}(D(p1,p2)) = \log(\sum(||pi|| - ||pj||)^2);$$

done
2. sort(k, D(t,p))
// Sort k-neighbors according to their distances.
3. for each instance t (i) in k-neighbors (sort(k, D(t,p)))
//compute probability estimation for the sorted neighbors

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

done
4. for each test sample t in k-neighbors (sort(k, D(t,p)))
Compute class membership probabilities of each test sample t
assign class to t sample using classifier.
done

Description: Improved KNN model efficiently classifies each instance. Algorithm gets clustered training data D, test samples T, k value and classes C as input. In Step1 for each instance p in dataset D ,logarithmic distance is calculated. Step 2 arranges all k-neighbours of each instance p according to their distances. In Step 3 for each k in neighbours computes the Distribution probability. Step 4 computes class membership probabilities of each test sample t and assign class label using the classifier.

IV. Experimental Results

Data on air pollution is used in real time in a Java context, and the results of the experiment are used in that setting. The information shown here on the level of air pollution in India comes from real-time sensors that have been installed all around the country. These sensors provide frequent feeds to the systems that record data on air pollution. The official website for the Indian air quality index, which can be seen at https://app.cpcbccr.com/AQI_India/, the source of this information is the Indian government. A selection of representative statistics on air pollution in India is presented in Table 1, which includes data from a variety of locales.

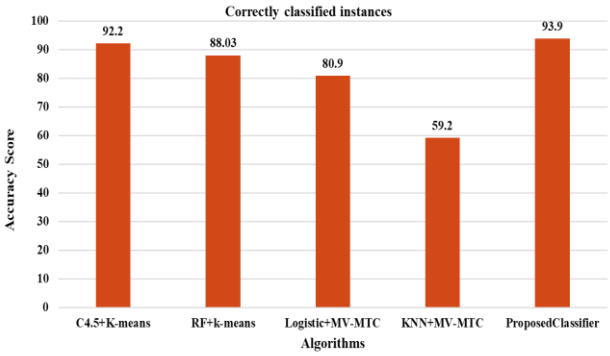

Fig 2: Comparative analysis of proposed cluster based classification model to traditional approaches on air pollution data.(classified instances)

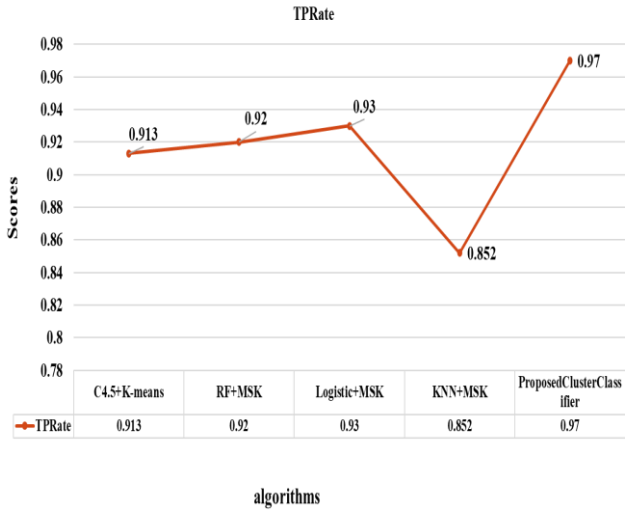

Fig 3. Comparative analysis of proposed model to traditional approaches on air pollution data.(true positive rate)

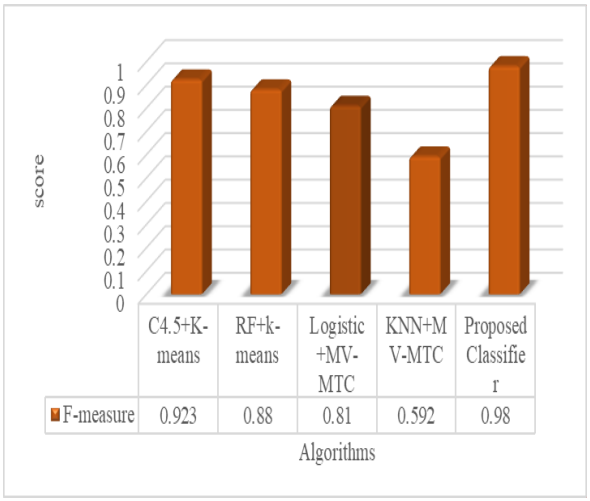

Fig 4. Comparative study of Average cluster based classification approaches and its measures

V. CONCLUSION

The purpose of this work is to present a hybrid cluster-based classification strategy in order to improve the process of forecasting the severity of conditions involving air pollution. Finding and forecasting the severity based on the location is extremely important because the majority of traditional algorithms have difficulty doing so when working with training data that has not been filtered properly. The purpose of this research is to offer a hybrid method for identifying extreme outliers, which would make it possible to filter out noise that is prevalent at the extremes. In addition, the filtered data is utilized in a clustering model that is founded on previous and posterior concepts for the purpose of the classification process. Based on the experimental findings, it can be concluded that the current framework is superior to the conventional methods in terms of optimization for the procedure of air pollution detection.

REFERENCES

(Periodical style)

[1] Yan, Y., Cao, L., & Rundensteiner, E. A. (2017, August):Scalable top-n local outlier detection. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1235-1244.

[2] Hu, Z., Bodyanskiy, Y. V., Tyshchenko, O. K., & Boiko, O. O. (2018):A neuro-fuzzy Kohonen network for data stream possibilistic clustering and its online self-learning procedure. Applied soft computing, Vol. 68, pp. 710-718.

[3] Li, W., Mo, W., Zhang, X., Squiers, J. J., Lu, Y., Sellke, E. W., & Thatcher, J. E. (2015):Outlier detection and removal improves accuracy of machine learning approach to multispectral burn diagnostic imaging. Journal of biomedical optics, 20(12), 121305.

[4] Chen, L., Wang, W., & Yang, Y. (2021) : CELOF: Effective and fast memory efficient local outlier detection in high-dimensional data streams. Applied Soft Computing, 102, 107079.

[5] Wang, B., & Mao, Z. (2019):Outlier detection based on Gaussian process with application to industrial processes. Applied Soft Computing, Vol.76, pp. 505-516.

[6] Cruz, R. M., Sabourin, R., & Cavalcanti, G. D. (2017). META-DES. Oracle: Meta-learning and feature selection for dynamic ensemble selection. Information fusion, Vol. 38, pp. 84-103.

[7] Santoyo, S. (2017):A brief overview of outlier detection techniques. Towards data science.

[8] Boukerche, A., Zheng, L., & Alfandi, O. (2020):Outlier detection: Methods, models, and classification. ACM Computing Surveys (CSUR), 53(3), pp. 1-37.

[9] Guansong Pang, Kai Ming Ting, and David Albrecht. 2015:LeSiNN: Detecting anomalies by identifying least similar nearest neighbours. In Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW'15). IEEE, 623–630.

[10] Haizhou Du, Shengjie Zhao, Daqiang Zhang, and Jinsong Wu. 2016:Novel clustering-based approach for local outlier detection. In Proceedings of the 2016 IEEE Conference on Computer Communications Workshops. pp. 802–811.

[11] 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, Dec. 2021, doi: 10.1016/j.neucom.2021.09.079.

[12] J. Song and M. E. J. Stettler:"A novel multi-pollutant space-time learning network for air pollution inference," Science of The Total Environment, p. 152254, Dec. 2021, doi: 10.1016/j.scitotenv.2021.152254.

[13] A.-L. Balogun, A. Tella, L. Baloo, and N. Adebisi:"A review of the inter-correlation of climate change, air pollution and urban sustainability using novel machine learning algorithms and spatial information science," Urban Climate, vol. 40, p. 100989, Dec. 2021, doi: 10.1016/j.uclim.2021.100989.

[14] P. Govender and V. Sivakumar:"Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)," Atmospheric Pollution Research, vol. 11, no. 1, pp. 40–56, Jan. 2020, doi: 10.1016/j.apr.2019.09.009.