RESEARCH ON ENERGY SPECTRUM ANOMALY DETECTION METHOD FOR STATE CONTROL RADIATION ENVIRONMENTAL MONITORING BASED ON LOF ALGORITHM

by

Niu YUNLONG [1,2]***, Luo YUNFEIA*** [1]***, Cong FEIYUN*** [1]***, Zheng HUIDI*** [2]***, and Liu XINGGAO*** [1*]

[1] Zhejiang University NGICS Platform, College of Control Science and Engineering, Zhejiang University, Hangzhou, China
[2] Radiation Environment Monitoring Technology Center, Ministry of Ecology and Environment, Hangzhou, China

Scientific paper
https://doi.org/10.2298/NTRP2401047Y

The state-controlled radiation environment automatic monitoring stations are important facilities for monitoring the quality and trends of the radiation environment. However, the analysis and utilization of monitoring data from these stations have limitations. To address this issue, this paper proposes an approach that utilizes the high-dimensional data visualization method based on t-SNE and PCA-BIRCH algorithm's spectrum clustering model to classify historical spectrum data. Subsequently, a local outlier factor-based environmental spectrum abnormal fluctuation detection model is established to enhance discrimination of non-environmental factors causing spectrum anomalies. Results indicate significant differences in spectrum aggregation under different meteorological conditions. The clustering algorithm effectively separates clear and rainy weather spectra based on spectral features, with minimal time consumption and reduced interference from mixed spectrum data. By utilizing local density features, we establish an anomaly detection model, assigning an abnormal score to each spectrum. Comparative analysis demonstrates improved discriminability of the anomaly detection model for unclassified datasets. Our algorithm accurately identifies spectrum fluctuations caused by non-environmental factors amidst complex backgrounds, offering valuable technical support for environmental quality assurance and nuclear emergency decision-making.

Key words: radiation environment, data visualization, PCA-BIRCH algorithm, clustering analysis

INTRODUCTION

Currently, China has built and operated 500 national radiation environment air automatic monitoring stations (hereinafter referred to as *automatic stations*). Moreover, the central government has invested 1.2 billion yuan to build a radiation automatic monitoring network covering all districts and cities in the country [1]. Through radiation environment quality monitoring and supervisory monitoring, it is possible to comprehensively grasp the quality status of the radiation environment, changes in radiation environment trends, and emissions of radioactive pollutants. This approach provides a scientific basis for environmental law enforcement and radiation pollution prevention, meets the public's right to know about the environment, and plays an important role in ensuring the healthy development of nuclear energy and nuclear technology utilization.

Nowadays, the hardware facilities of the radiation monitoring network are complete, with some functions of remote data transmission and storage [2]. However, there are still some issues to be addressed in the data utilization and processing, mainly due to the significant shortcomings in the analysis and utilization of monitoring data from automatic stations on the existing data platform. There is a lack of analysis and evaluation standards for NaI environmental spectra, and the monitoring and alarm functions of the spectrum data have not been effectively utilized. Due to the scattered distribution of automatic stations, they are affected by natural environmental and meteorological factors such as rainfall, and the types of nuclides in the environmental spectrum are complex and have low activity. It is difficult to achieve good results, whether based on spectrum decomposition or full spectrum analysis methods. Therefore, there is an urgent to establish a detection model that is not sensitive to the fluctuation of complex environmental radiation of NaI spectrometers.

* Corresponding author, e-mail: lxg1232022@126.com

In traditional gamma spectrum qualitative analysis, the identification of nuclide species is typically achieved by comparing the characteristic peaks in the energy spectrum with a standard nuclide library. This process involves spectral smoothing, peak finding, peak area determination, background deduction, full energy peak fitting, peak net area calculation, nuclide identification, and activity calculation [3]. However, during actual measurements, environmental noise can interfere with the analysis, especially when using a NaI gamma spectrometer with low energy resolution. If the substance to be measured contains many types of nuclides, low content, and complex media, and in complex environments such as high background, the measured gamma line will be quite complicated. Therefore, more advanced spectral analysis algorithms are required to obtain accurate results.

The energy spectrum analysis method based on peak analysis has undergone advancements over time. It started with the peak area method [4] and has since progressed to include the spectral stripping method [5], inverse matrix method [6], trace by trace least squares method [7], and function fitting peak area method [4]. In the case of similar types of nuclides in the mixed sample, there can be a significant overlap of energy peaks in the spectral lines. In such cases, the least squares method is often effective. This method aims to find the best function that matches the data by minimizing the sum of squared errors. For spectral lines, it essentially involves applying a low-pass filter to remove high-frequency noise from the spectral lines. The function fitting peak area method involves dividing the energy spectrum into several regions, fitting each region's full energy peak with a function, and then integrating the function to obtain the peak area. This approach can obtain the nuclide category and activity of the energy spectrum.

In their research, Liu *et al*. [8] applied the entropy averaging method to eliminate random interference in gamma spectrum measurements. They selected multiple points on the rock profile and obtained certain results. Fu [9] utilized the least squares fitting and five-point fitting smoothing methods to fit spectral lines and enhanced the original SNIP algorithm by implementing a dynamic window through peak boundary counting. Additionally, they replaced the original second-order filtering function with a fourth-order filtering function, resulting in a background deduction rate of over 95 %. Yang Kui employed the 3-point center of gravity method, the least squares moving smoothing method, and Gaussian low-pass filtering to smooth gamma spectra. They compared the effectiveness of various smoothing methods. The Gaussian fitting was used to fit the feature peaks, while the least squares method was employed to decompose the heavy peaks, achieving good results [5]. Zhao *et al*. [10] applied back propagation (BP) and OLAM neural networks to develop portable HPGe gamma spectral nuclide recognition systems. Li [11] combined the Monte Carlo method with a double-layer BP neural network algorithm. They used the double-layer BP neural network to integrate simulated data with measured data, successfully analyzing gamma spectrum data. Wang C. J., *et al*. [12], and Wang Y. *et al*. [13], utilized the fuzzy recognition mechanism of gamma spectroscopy fingerprint to perform energy spectra analysis and nuclide identification. Ren *et al*. [14] proposed a method based on singular value decomposition to extract feature vectors and utilized them as input to support vector machines for constructing classifiers. This method addresses challenges such as insignificant spectral features and low recognition accuracy in complex gamma-ray spectra. It reduces the requirements for detector accuracy, minimizes the impact of parameter settings, and enhances the recognition capability for mixed nuclides. Zhang *et al*. [15] presented a method for extracting gamma spectrum features using sparse representation. This method tackles the challenge of feature extraction in situations where heavy peaks and strong noise backgrounds coincide with weak peaks in gamma spectrum analysis. Liu [16] conducted a study on the application of fuzzy decision trees in gamma spectroscopy, effectively capitalizing on their advantages, including clear model structure and soft decision-making. A dynamic division of the sample space was employed to accurately identify nuclide types in various scenarios, including those involving small samples, limited attributes, and both single and mixed nuclides.

While current full spectrum analysis methods widely employ complete spectrum information, they primarily focus on HPGe gamma spectrometer spectra under controlled laboratory conditions. In the case of environmental NaI spectra analysis addressed in this article, these methods are not fully applicable, resulting in high rates of false positives and false negatives.

This paper addresses the significant shortcomings in the analysis and utilization of monitoring data. The *t*-SNE (stochastic neighbor embedding) method, which is effective for high-dimensional data visualization, is used to perform spectrum dimensional reduction analysis under different meteorological conditions. Additionally, the PCA-BIRCH clustering model, capable of high-quality clustering of large datasets with limited memory resources, is utilized to classify historical spectrum data. Also, other clustering methods are selected for comparison to avoid interference caused by mixed types of monitoring data on subsequent anomaly data detection and applications. Based on the dataset classification, an local outlier factor (LOF) algorithm-based spectrum anomaly discrimination model is established using the local density features of the dataset. The model obtains the anomaly score of the spectrum, where an anomalous score much greater than 1 indicates that the data point may be an outlier. Comparing the discrimination re-

sults on the unclassified dataset demonstrates that clustering analysis of the dataset can effectively improve the discrimination of the anomaly detection model on abnormal data points. The clustering anomaly detection algorithm proves capable of discriminating the fluctuations of the spectrum caused by non-environmental factors under complex environmental conditions, providing strong guarantees for data quality in automatic radiation monitoring and nuclear emergency monitoring. It improves the timeliness, accuracy, and foresight of radiation monitoring work, provides data support for enhancing the government's credibility in nuclear emergencies, offers technical support for national nuclear and radiation supervision work, and provides strong guarantees for national environmental and public safety.

PRINCIPLE OF RADIATION DETECTOR

Currently, the automatic stations in China use spectrometers with an energy range between 30 keV to 3 MeV and a working temperature range of –20 °C to +60 °C. The spectrometer consists of a NaI scintillator, a PMT photomultiplier tube, a GM detector (optional), an MCA multi-channel analyzer, and an embedded PC, as shown in fig. 1. The research object and data source of this article is the NaI spectrometer used in automatic stations [17], and the spectrum is shown in fig. 2.

DATA VISUALIZATION ALGORITHM

Principle of stochastic neighbor embedding algorithm

The SNE is a manifold learning method proposed by Hinton and others [18]. It replaces traditional Euclidean distance with conditional probability distance to measure the similarity between sample points and has better visualization effects in high-dimensional data visualization [19].

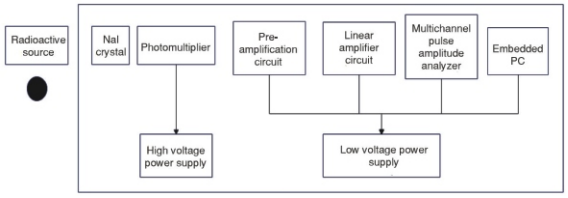

Figure 1. Hardware structure schematic diagram of NaI energy spectrometer

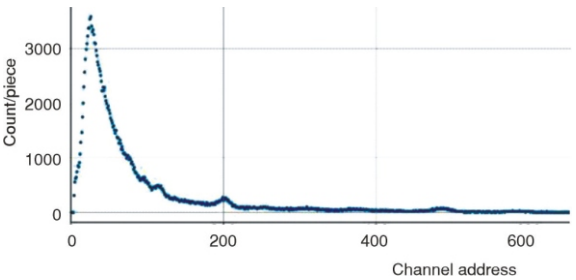

Figure 2. Environmental NaI energy spectrum

Assuming the input data is $\boldsymbol{X}$ $\boldsymbol{R^n}$, Output data is $\boldsymbol{Y}$ $\boldsymbol{R^t}$ ($\boldsymbol{t} \ll \boldsymbol{n}$), Assuming there are m sample data $\{\boldsymbol{x^{(1)}}, \boldsymbol{x^{(2)}}, \ldots, \boldsymbol{x^{m}}\}$, therein $\boldsymbol{x^{(i)}}$ $\boldsymbol{X}$, the data after dimensionality reduction is $\{\boldsymbol{y^{(1)}}, \boldsymbol{y^{(2)}}, \ldots, \boldsymbol{y^{(m)}}\}$, $\boldsymbol{y^{(i)}}$ $\boldsymbol{Y}$. The SNE converts the Euclidean distance between points into a conditional probability, which expresses their similarity. Specifically, SNE first calculates the conditional probability, which is proportional to the similarity between point and point. The formula for calculating this probability is

$$p_{j/i} = \frac{\exp\left(-\frac{\left\|x^{(i)} - x^{(j)}\right\|^2}{2\delta_i^2}\right)}{\sum_{k=1} \exp\left(-\frac{\left\|x^{(i)} - x^{(k)}\right\|^2}{2\delta_i^2}\right)} \quad (1)$$

where δ_i is the variance of the Gaussian distribution centered at $\boldsymbol{x^{(i)}}$.

For data points $\boldsymbol{y^{(i)}}$ in low dimensions, the conditional probability $\boldsymbol{q_{ji}}$ is used to describe the similarity between $\boldsymbol{y^{(i)}}$ and $\boldsymbol{y^{(j)}}$, and the formula for redefining $\boldsymbol{q_{ji}}$ using the t distribution is

$$q_{ij} = \frac{\left(1 + \left\|y^{(i)} - y^{(j)}\right\|^2\right)^{-1}}{\sum_{k=1} \sum_{l=1} \left(1 + \left\|y^{(k)} - y^{(l)}\right\|^2\right)^{-1}} \quad (2)$$

Algorithm principle of BIRCH

The BIRCH is a balanced iterative and clustering using hierarchical methods. Its main feature is the ability to achieve high-quality clustering on large datasets using limited memory resources while minimizing *I*/*O* costs by scanning the dataset in a single pass. The BIRCH is based on the concepts of clustering features (CF) and CF trees. The CF is essentially a summary of statistical information for a given cluster, which can be used to calculate other metrics. The metrics for a single cluster include

$$R = \sqrt{\frac{\sum_{i=1}^{n} (x^{(i)} - \bar{x})^2}{n}} \quad (3)$$

$$D = \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (x^{(i)} - x^{(j)})^2}{n(n-1)}} \quad (4)$$

Among them, the radius $\boldsymbol{R}$ and diameter $\boldsymbol{D}$ can reflect the tightness of the cluster around the centroid.

DATA CLUSTERING EFFECT ANALYSIS

Analysis of visualization results of energy spectrum data based on *t*-SNE

The historical energy spectrum data from a station in 2019 was selected and processed by daily averaging, resulting in a total of 298 effective data points, of which 138 are for rainy days and 160 are for non-rainy days. The energy spectra are 1024-dimensional. The energy spectrum is compressed from 1024- dimensions to 2-D using the *t*-SNE algorithm, and the relationship between the energy spectrum and rainfall is reflected by the type and size of the data points, as shown in fig. 3. By observing the 2-D distribution of energy spectrum data points and their relationship with rainfall, it can be determined that the characteristics of the energy spectrum on rainy days are significantly different from those on non-rainy days, resulting in the formation of separate clusters in low-dimensional visualization plots. However, affected by the rainfall intensity, the impact of energy spectrum data on rainy days is relatively small when the rainfall is low, resulting in some data points of rainy days being mixed in non-rainy energy spectrum clusters. As the influence of rainfall increases, the spectrum of rainy days gradually forms independent clusters.

Effect analysis of energy spectrum clustering based on PCA-BIRCH

Selecting the daily average energy spectrum data of a site in 2019, and using the BIRCH clustering algorithm to divide the energy spectrum data into two clusters: rainy-day spectrum and sunny-day spectrum. The

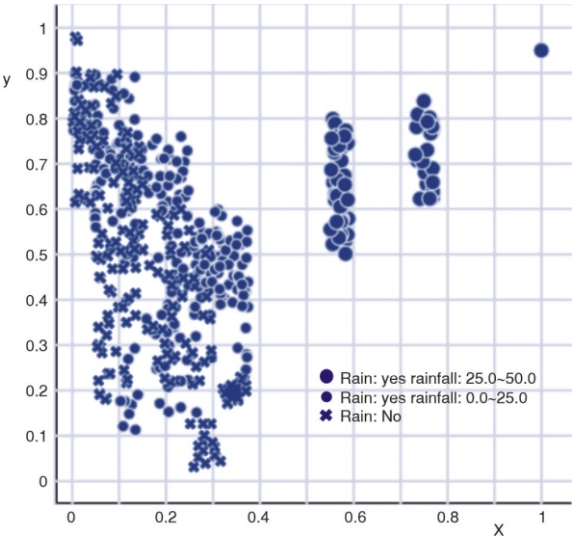

Figure 3. The 2-D visualization of spectral data

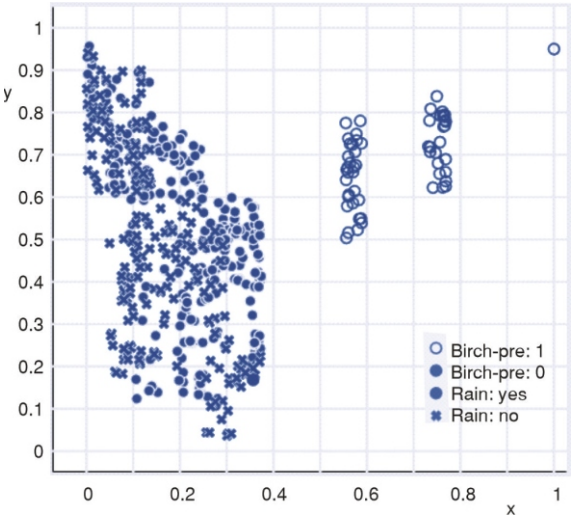

Figure 4. Spectral clustering effect based on BIRCH

resulting clustering was visualized using the *t*-SNE algorithm. To evaluate the performance of the BIRCH algorithm, we compared the clustering results with the actual data point distribution and also compared it with other clustering algorithms to verify its effectiveness in this problem.

The results of energy spectrum clustering based on the BIRCH algorithm are shown in fig. 4. The circular dots represent data points with rainfall greater than zero on that day, while the crosses represent data points with zero rainfall. The solid and hollow dots represent the clustering results, where hollow dots represent rainy-day spectra and solid dots represent sunny-day spectra. In the 2-D spectral distribution visualized by the *t*-SNE algorithm, it can be seen that the solid and hollow dots are clustered differently. In the cluster of solid dots, there are mixed non-rainfall data points and rainfall data points, indicating that the BIRCH algorithm also classified the part with rainfall but had little impact on spectral features into sunny-day spectra. In the cluster of hollow dots, all the data points have rainfall greater than zero, and this part of the spectrum is significantly affected by rainfall, resulting in a significant difference in spectral features compared to sunny-day spectra.

In addition, other clustering methods are also selected for comparison, as shown in fig. 4. From top left to bottom right, GaussianMixture, KMeans, SpectralClustering, and MeanShift algorithms were used for spectral clustering and their visualization results were shown. Compared with fig. 5, GaussianMixture, KMeans, and MeanShift all have similar problems, the transition state between the rainy spectrum and the sunny spectrum cannot be well distinguished. For example, in the upper-left figure, some square data points in the mixed cluster of rainfall spectra and non-rainfall spectra were identified as rainy-day spectra. The clustering effect of SpectralClustering in the lower left is ineffective due to the high dimensionality of the data. By comparing the effect with other clustering algorithms, the

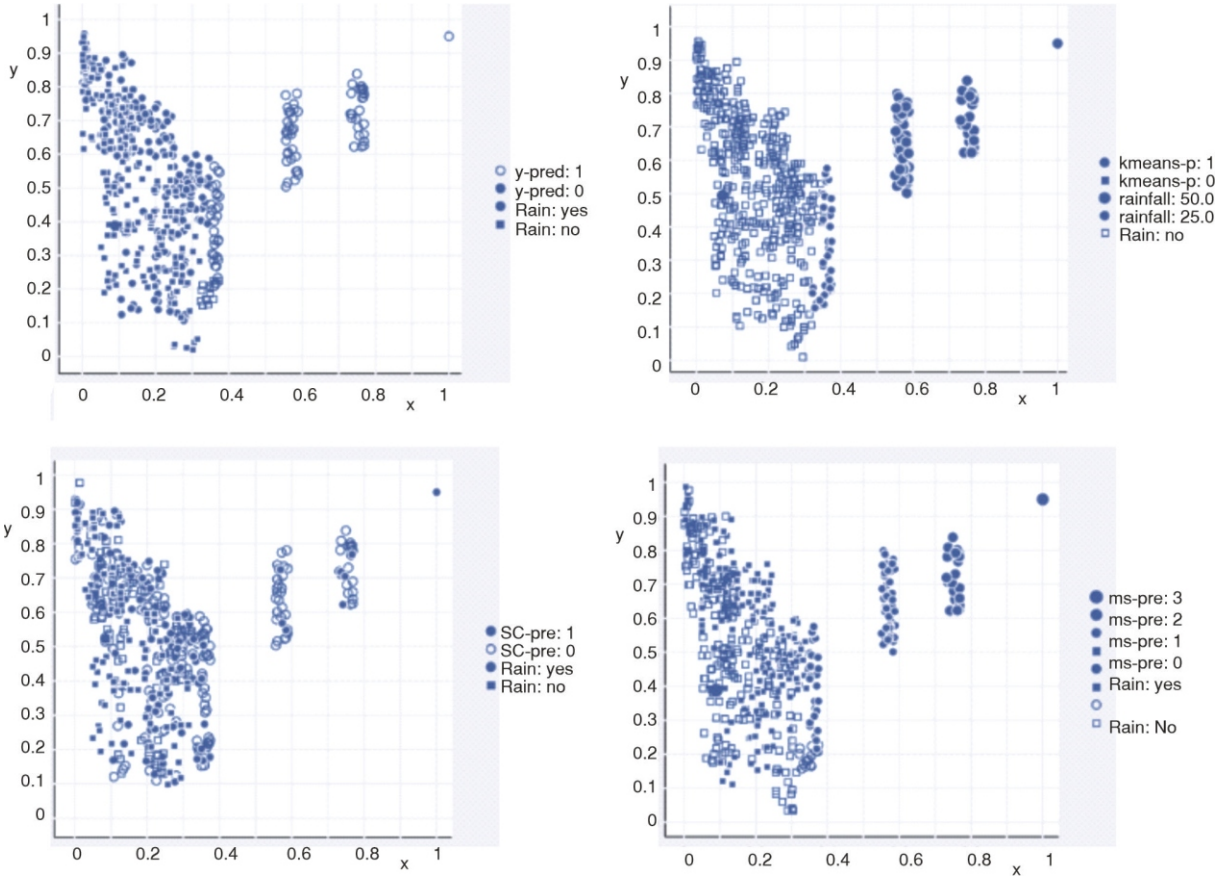

Figure 5. Comparison of energy spectrum clustering effects of different algorithms

BIRCH algorithm has better performance in spectral data clustering in this paper. It can effectively separate sunny-day spectra and rainy-day spectra according to spectral features, and due to the characteristics of the BIRCH algorithm, it can complete high-quality clustering of a large amount of spectral data using limited memory resources. Also, the single-pass scanning method can minimize *I/O* costs and reduce the time consumption of each clustering.

Site history spectral cluster

We have selected historical spectral data from the site in 2019, which were recorded every 5 minutes and included 1024 channels of particle counts. There is a total of 83882 valid data records. To reduce the data fluctuation of the energy spectrometer, we averaged six consecutive sets of energy spectrum data. This extended the particle counting period of the instrument from 5 minutes to 30 minutes, resulting in a total of 16475 sets of valid data. The PCA was used to reduce the dimensionality of the spectral data from 1024 to 50, and then the BIRCH algorithm was used for clustering analysis. The data was divided into a sunny-day spectral dataset and a rainy-day spectral dataset according to the clustering labels. The clustering results are visualized in fig. 6. Through the dimensionality reduction cluster analysis of the energy

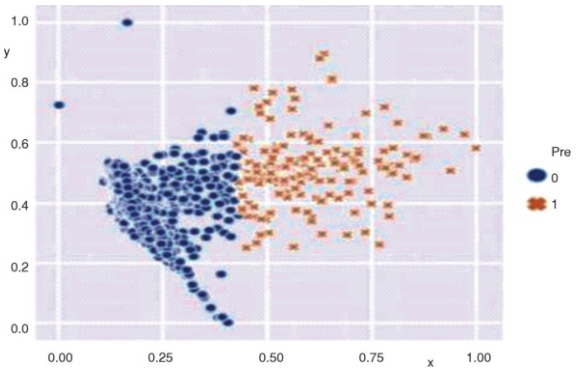

Figure 6. Spectral clustering performance of a site's historical data

spectrum dataset, it is divided into rainy-day spectra and sunny-day spectra. The circular dots indicate sunny-day spectral data, while the crosses indicate rainy-day spectral data.

Anomaly detection algorithm for local outlier factors

The key to density-based outlier detection methods is to assign a density value to each data point. The main idea is that, for any given data set, if the points in

its neighborhood are dense, it is considered a normal data point, while if the points in its neighborhood are far apart, it is considered an outlier. The threshold is used to define whether it is an outlier. In density-based anomaly detection methods, the local outlier factor (LOF) [20-22] algorithm is one of the most representative methods. The relevant definitions of the LOF algorithm are as follows:

– $d(O, P)$: the distance between two points O and P;
– the k^{th} distance $d_k(O)$ for point O is defined as follows: $d_k(O) = d(O, P)$, and satisfy:
  (a) there are at least k points not including O in the set P' $C\{x$ $0\}$, satisfy $d(O, P')$ $d(O, P)$
  (b) there are at most $k-1$ points not including O in the set P' $C\{x$ $0\}$, satisfy $d(O, P')$ $d(O, P)$; the k^{th} distance of O, that is, the distance from the k^{th} farthest point of O, as shown in fig. 7.
– The k^{th} distance neighborhood, that is, the neighborhood $N_k(O)$ within the point O to $d_k(O)$, contains all points within the k^{th} distance of O, including those on the k^{th} distance. Therefore, the number of k^{th} neighbor points of O is $|N_k(O)|$ k.
– According to fig. 8 the k^{th} reachable distance from point P to point O is defined as

$$d_k(O,P) \quad \max\{d_k(O), d(O,P)\} \tag{5}$$

The k^{th} reachable distance from point P to point O is at least the k^{th} distance of point O. The k points closest to point O, have the same reachable distance to point O, and they are all equal to $d_k(O)$

– The local reachable density is defined as

$$lrd_k(O) \quad \frac{1}{\frac{P \ N_k(O) d_k(O,P)}{|N_k(O)|}} \quad \frac{|N_k(O)|}{P \ N_k(O) d_k(O,P)} \tag{6}$$

Indicates the average reachable distance from all points in the k^{th} neighborhood of point O to O. If point O and the surrounding neighborhood points are in the same cluster, then the reachable distance is more likely to be a smaller $d_k(O)$, resulting in a smaller sum of reachable distances, the greater the local reachable density. If O is far away from the surrounding neighbor points, then the reachable distance may take a larger value $d(O, P)$, resulting in a larger sum of reachable distances and a smaller local reachable density.

Local outlier factor

According to the definition of local reachable density, when a data point is far away from other points, its local reachable density is likely to be smaller. The LOF algorithm determines whether a data point is an outlier based on its relative density to its neighboring data points, rather than its absolute local density. Therefore, the algorithm performs well in situations where data density and distribution are different or uneven. The local outlier factor is defined by the local relative density. The local relative density is defined as the ratio of the average reachable density of the neighborhood points of point O to the local reachable density of point O [23-26] Right now

$$LOF_k(O) \quad \frac{\frac{lrd(P)}{P \ N_k(O) \ lrd(O)}}{|N_k(O)|} \tag{7}$$

According to the definition of the local outlier factor, the anomaly score of a data point fluctuates around 1. If the anomaly score is less than 1, it indicates that the data density near the data point is high, and the probability of being an outlier is relatively low. If the anomaly score is far greater than 1, it indicates that point P is far away from other data points and is likely to be an outlier.

Anomaly detection model of energy spectrum based on the LOF method

The aim of anomaly detection for the station spectrum data in this paper is to distinguish the spectrum anomaly fluctuations caused by non-environmental factors under complex conditions. To achieve

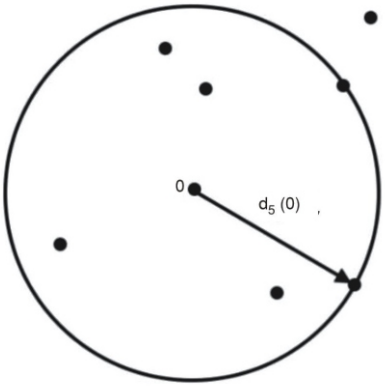

Figure 7. The 5th distance of point *O*

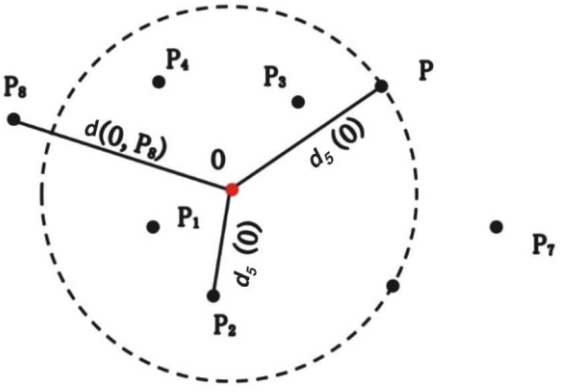

Figure 8. Schematic diagram of reachable distance

this, an anomaly detection model has been established to detect and discriminate spectra with different cluster distributions based on the prior discussion of clustering of historical energy spectrum data.

Due to the particularity of the environmental spectra of automatic stations, the probability of nuclear anomaly events is extremely low. For the historical data set of the current station, it can be considered that there is no obvious nuclide anomaly event. In the process of establishing the anomaly detection model for the spectra, the lack of labeled abnormal spectra makes it impossible to directly assess the effectiveness of the detection model. To address this issue, Gaussian peaks are added to the actual spectrum to simulate the spectrum anomaly fluctuations. The fitted Gaussian peak shown in fig. 9, where the *x*-axis represents channel number and the *y*-axis represents particle counts, has a count of 1000. The fitted spectrum data serves as the abnormal data points, and the detection of abnormal data points by the detection model is used as the evaluation index of the model's effectiveness.

Due to the high dimensionality of energy spectrum data, which has 1024 dimensions, many traditional outlier detection methods are ineffective and are affected by the curse of dimensionality. To overcome this challenge, we chose the LOF algorithm to establish a spectrum anomaly discrimination model. Unlike distance-based or density-based algorithms, the LOF algorithm relies on the local outlier factor score, which is based on the local relative density of data points and is less limited by dimensionality. The LOF method provides a quantitative measure of the degree to which an object is an outlier and can handle well even if the data set has regions of different densities [27-32]. The process of energy spectrum anomaly discrimination is shown in tab. 1.

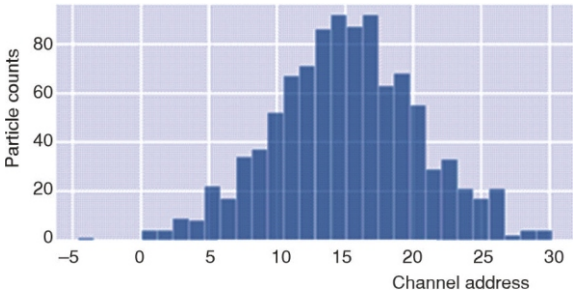

Figure 9. Fitting gaussian peak

Analysis of anomaly detection model results

The test datasets include a single weak peak test dataset, a single strong peak test dataset, and a multiple weak peak test dataset (with a single weak peak added to the spectrum as shown in fig. 10, a single strong peak added to the spectrum as shown in fig. 11, and multiple weak peaks added to the spectrum as shown in fig. 12). The effectiveness of the anomaly detection model based on the LOF algorithm was tested using these datasets. The LOF algorithm model was also tested using the uncategorized dataset to verify if spectrum clustering enhances the ability of the model to detect spectrum anomalies.

For each trial, 20 sets of spectra data were randomly selected from the anomaly dataset and added to the normal dataset. Each anomaly detection model was utilized to identify the anomalous data points in the test dataset, and the number of identified anomalous points was used as the model evaluation metric. This process was repeated five times The model directly outputs the outlier score of this set of data. Therefore, by sorting the outlier scores of the data set, the top 30 outlier scores were used as predicted abnormal data points. The number of real abnormal data points served as the model detection result. The results are as follows in tab. 2 shows.

The LOF-based detection model successfully identified over 80 anomalous data points in the single weak peak, single strong peak, and multiple weak peak datasets. Notably, it performed exceptionally well in the single strong peak dataset, correctly identifying 99 anomalous data points. Additionally, the performance of the model was improved after clustering

Table 1. The algorithm steps of LOF

The detailed process of the algorithm is as follows:
Data: Spectrum data of the test set x_1, ..., x_n
Calculation parameters: Positive integer K (Used to calculate the k-th distance)
Output: Local outlier factors for each spectral data
Start:
Computes the Euclidean distance between each data point and other objects.
Sort the distances, computing the k-th distance and k-th distance neighborhood for each data point.
Calculate the local reachable density of each data point according to formula (4.2)
Calculate the local outlier factor for each data point according to formula (4.3)
Finish
The local outlier factor is used as the abnormal score of the data point, and sorted, and the data point with a high abnormal score is the abnormal point

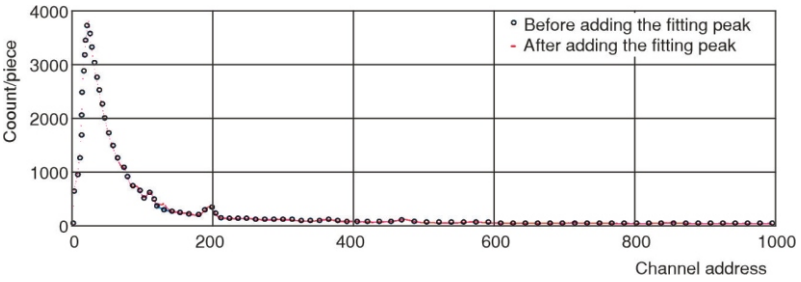

Figure 10. Comparison of spectra before and after adding a single weak peak

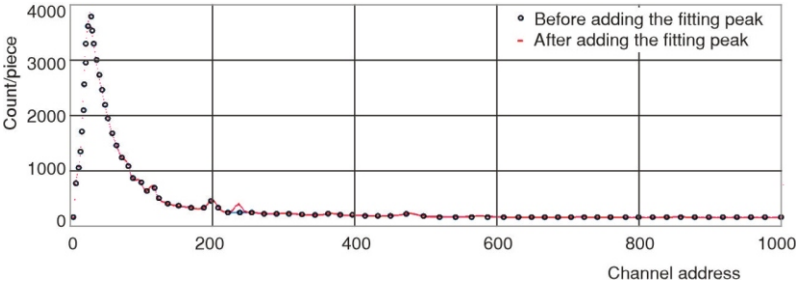

Figure 11. Comparison of spectra before and after adding a single strong peak

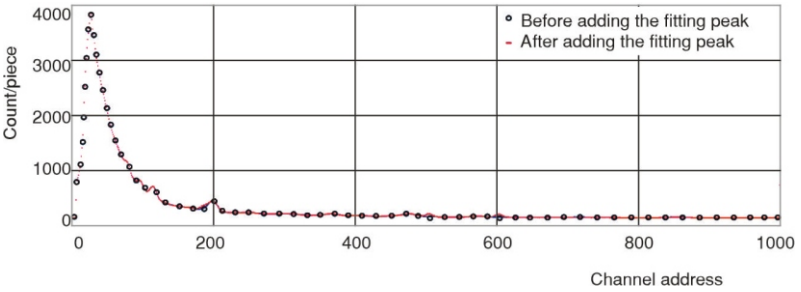

Figure 12. Comparison of spectra before and after adding multiple weak peaks

Table 2. Dataset test results

Dataset	1st round	2nd round	3rd round	4th round	5th round	Total count	Detection rate [%]
Single weak peak	18	18	18	16	16	86	86
Single weak peak (unclassified)	17	17	16	15	16	81	81
Single strong peak	20	19	20	20	20	99	99
Single strong peak (unclassified)	20	19	20	20	20	99	99
Multiple weak peaks	18	20	17	18	17	90	90
Multiple weak peaks (unclassified)	17	16	17	17	84	84	84

the spectra data. By adding a fitted Gaussian peak to the spectra data, the reachability distance between neighboring data points increased, resulting in decreased local reachability density. However, the LOF algorithm evaluates data points based on their relative density to neighboring data points, rather than their absolute local density. This allows the algorithm to perform well under conditions of imbalanced data distribution and varying densities. Furthermore, the local relative density is less affected by dimensionality constraints. Thus, even in high-dimensional data with sparse distribution and minimal distance differences, the local relative density can still accurately reflect the degree of data outliers.

Energy peak resolution

The automatic station utilizes the SARA gamma spectrometer to measure energy spectrums. Since the relationship between particle energy levels and channel addresses can be approximately considered linear, a multi-point linear regression can be used to obtain this functional relationship. In laboratory experi-

ments, the measured relationship between channel addresses and particle energy levels is as follows

$$y \quad 0.4611x \quad 4.2626 \tag{8}$$

Due to the differences in performance between energy spectrometers and the disparities in testing environments, the energy calibration of the station's energy spectrometer may deviate slightly from eq. (8). However, the linear relationship between channel addresses and energy levels remains the same. The environmental spectrum contains complex nuclide species with low concentrations, and the resolution of the NaI energy spectrometer is low. As a consequence, direct spectral analysis of the spectrum proves ineffective, and there are many similar data points in the measured spectrum. Performing spectral analysis on each daily spectrum data would result in a large number of repetitive calculations. Therefore, spectral analysis is only performed when the abnormality score of the spectrum exceeds a certain threshold. Since the peak width calibration curve of the energy spectrometer has not been calibrated, only qualitative analysis of nuclides can be performed. Considering the low instrument resolution, the results of spectral analysis supplement the energy anomaly detection and provide additional information to the operators. The process of spectral decomposition is shown in fig. 13.

The original energy spectrum is smoothed through the average of multiple data sets. Then, according to the reference energy spectrum determined based on the BIRCH method, the energy spectrum shown in fig. 14 is obtained after deducting the background. In the process of photoelectric spectral effect, Compton effect, and forming electron pair effect, especially the Compton scattering effect, gamma ray energy tends to drift towards the low energy region of the track, resulting in an increase in gamma energy in the low energy region. Consequently, the spectral lines overlap seriously. Therefore, there are higher *energy peaks* in the low-energy region in the figure, but these are the result of a large number of overlapping spectral lines. However, there are obvious characteristic peaks in the interval from 100 to 220 tracks, The number of peak center sites is determined by peak searching. The energy level of the nuclide can be determined according to the energy scale curve, and the approximate type of nuclide can be identified using the energy peak database. The energy level of the region marked by the red line in the figure is roughly 0.5 MeV, which corresponds to ^{222}Rn. The main daughter of radon, ^{214}Pb, is less obvious on the left side. While spectral analysis provides information about the nuclide type, its accuracy is limited by resolution rate, fuzzy peak area boundaries, high background leading to low recognition rates, or even misjudgment. Therefore, it should be used as supplementary information for detecting environmental energy spectrum anomalies to help nuclear emergency personnel make better decisions.

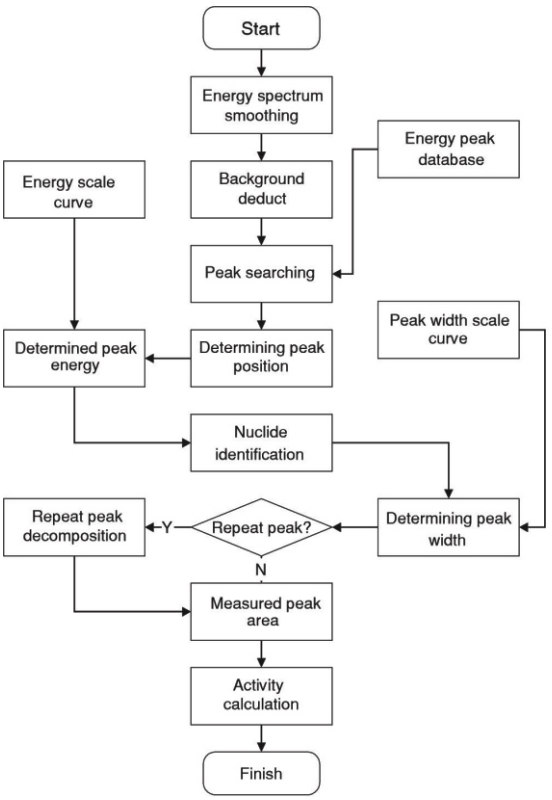

Figure 13. Flow chart of energy spectrum peak analysis

CONCLUSIONS

This article addresses the shortcomings of the current analysis and utilization of monitoring data from automatic station detectors. To address this issue, the *t*-SNE method is innovatively adopted for energy spectrum dimensionality reduction analysis under different meteo-

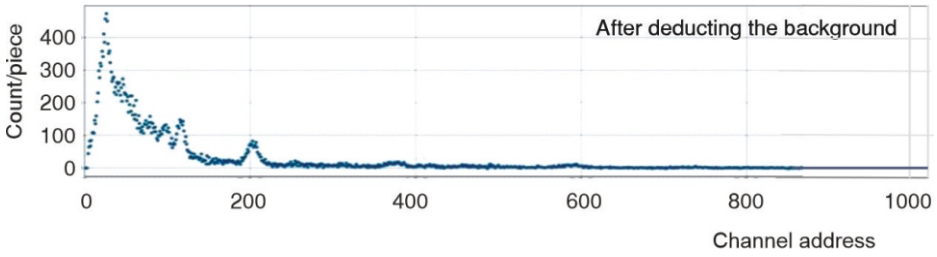

Figure 14. Environmental energy spectrum after subtracting background

rological conditions. The PCA-BIRCH algorithm clustering model, which can perform high-quality clustering on large datasets with limited memory resources, is then used to conduct clustering analysis research on historical energy spectrum data. The comparison results show that the clustering characteristics of the energy spectrum dataset can be displayed through the dimensionality reduction visualization method. The BIRCH algorithm has a better clustering effect on the radiation monitoring energy spectrum data and can effectively separate the sunny spectrum and the rainy spectrum according to the energy spectrum characteristics, with minimal time consumption.

Since existing energy spectrum analysis methods cannot identify abnormal fluctuations of the energy spectrum in complex environmental backgrounds, or energy spectrum fluctuations due to meteorological influences. This paper proposes an environmental energy spectrum anomaly detection method based on local anomaly factors. Based on the classification of the energy spectrum dataset, the local density feature of the dataset is used to establish the LOF algorithm and obtain the abnormal score of the energy spectrum. The simulation results show that the recognition accuracy of the LOF-based detection model can exceed 80 % in the single weak peak, single strong peak, and multiple weak peak datasets. This confirms that the cluster analysis of the dataset can effectively improve the discrimination of abnormal data points by the anomaly detection model. Furthermore, it is verified that the clustering anomaly detection algorithm can discriminate energy spectrum fluctuations caused by non-environmental factors under complex environmental background conditions.

The clustering-based anomaly detection algorithm proposed in this paper could help to address the interference of mixed monitored data types on subsequent anomaly detection and nuclear accident warnings, providing technical support for nuclear emergency decision-making.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (12105246, 62073288, 11975207,12075212), the National Key RD Program of China (Grant No. 2021YFC2101100).

AUTHORS' CONTRIBUTIONS

N. Yunlong: Writing – Original Draft; Conceptualization; Formal analysis. L. Yunfei; Writing – Review & Editing; Methodology; Formal analysis. C. Feiyun: Validation; Data Curation. Z. Huidi: Formal analysis; Investigation. L. Xinggao: Resources; Funding acquisition; Supervision.

ORCID NO

Niu Yunlong: 0000-0002-1179-9095
Cong FEIYUN: 0000-0003-1727-7164
Zheng HUIDI: 0009-0007-3315-0186
Liu XINGGAO: 0000-0002-0948-1942

REFERENCES

[1] Niu, Y. L., *et al*., Operation Maintenance and Work Planning of Radiation Environment Automatic Monitoring Stations in China (in Chinese), *Radiation Protection Bulletin, 36* (2016), 06, pp. 21-25+29
[2] Gostilo, V. V., *et al*., Development of Nuclear Radiation Monitors for Radiation Early Warning Systems, *Nucl Technol Radiat, 37* (2022), 3, pp. 193-200
[3] Liu, Y. G., Study on γ Spectral Data Analysis Methods (in Chinese), M. Sc. thesis, China University of Geosciences, Beijing, China, 2011
[4] Pang, J., Simplification of the Calculation Formula of Wasson-Sterlinski Peak Area, *Atomic Energy Science and Technology, 1* (1983), 10, pp. 29-31
[5] Wang, J., Jiang, J., Determination of Net Area for 92.6 keV peak of ^{238}U by the Spectrum-Stripping Method, *Nuclear Techniques, 4* (1992), 3, pp. 205-207
[6] Liu, J., *et al*., Study on Parsing γ Radial sPectrum of NaI (TI) Spectrum Instrument by Peak Area and Perverse Matrix Method, *Arid Environmental Monitoring, 1* (2004), 5, pp. 12-15
[7] Liu, P., *et al*., The Use of One by One Channel Least Square Method for the Analysis of natural Radioactivity NaI (TI)-Spectrum in Soil, *Nuclar Techniques, 9* (1986), 19, pp. 49-50+65
[8] Liu, K., *et al*., Application of Entropy Averaging Method in Eliminating Random Interference of Ground Gamma Spectroscopy Measurement, *World Nonferrous Metals, 15* (2019), Sept., pp. 225-226
[9] Fu, R. X., The Design and Implementation of Nuclide Identification Algorithm for Portable Gamma Spectrometer, (in Chinese), M. Sc. thesis, Chengdu University of Technology, Chengdu, China, 2018
[10] Zhao, L., *et al*., Neural Network Algorithm Analysis HPGe Germanium Gamma Spectrum, *Journal of East China University of Technology(Natural Science), 36* (2013), 15, pp. 79-83
[11] Li, F., Gamma Ray Energy Spectrum Data Processing-Technology Based on Monte Carlo Method and the Neural Network Algorithm (in Chinese), Ph. D. Thesis, Chengdu University of Technology, Chengdu, China, 2016
[12] Wang, C. J., *et al*., Investigation on the fuzzy Recognition Mechanism for γ-Ray Fingerprints of Nuclear Materials, *Wuli Xuebao/Acta Physica Sinica, 57* (2008), 9, pp. 5361-5365
[13] Wang, Y., Wei, Y., Fuzzy Logic Based Nuclide Identification for γ Ray Spectra, *Qinghua Daxue Xuebao/Journal of Tsinghua University, 52* (2012), 12, pp. 1736-1740
[14] Ren, J., *et al*., Radioactive Nuclide Identification Method Based on SVD and SVM, *Ordnance Industry Automation, 36* (2017), 05, pp. 50-53
[15] Zhang, J., *et al.*, Spectrum Feature Extraction by Sparse Representation, *Nuclear Electronics & Detection Technology, 37* (2017), 10, pp. 974-979
[16] Liu, H. L., The Study of Radioactive Nuclide Identification Method Based on Fuzzy Decision Tree (in Chinese), M. Sc. thesis, Southwest University of Science and Technology, Mianyang, China, 2018

[17] Yang, K., Application of NaI EDS in the Measurement of Activity of Radioactive Material in Nuclear Accident (in Chinese), M. Sc. thesis, East China University of Technology, Nanchang, China, 2016
[18] Hinton, G., Roweis, S., Stochastic Neighbor Embedding, in NIPS 2002: *Proceedings*, 15th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2002, Sept., pp. 833-840
[19] Xu, W. W., Exploration of Dimensionality Reduction for High-dimensional Data Visualization and Its Application in Biomedicine (in Chinese), Ph. D. thesis, Wuhan University, Wuhan, China, 2017
[20] Li, H., Research on Detection Method of Outlier Value of Power Data based on Fast density Peak Clustering and LOF (in Chinese), M. Sc. thesis, Lanzhou University of Technology, Lanzhou, China, 2019
[21] Yang, H., *et al*., K-Means Algorithm Based on LOF (in Chinese), *Communications Technology, 52* (2019), 08, pp. 1884-1888
[22] Zhang, S., *et al*., Outlier Detection Algorithm Based on Grid LOF and Adaptive K-Means (in Chinese), *Command Information System and Technology, 10* (2019), 01, pp. 90-94
[23] Angiulli, F., Pizzuti, C., Fast Outlier Detection in High Dimensional Spaces, in Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2431 (2002), pp. 15-27
[24] Bellman, R. E., Dreyfus, S. E., Applied Dynamic Programming (Applied Dynamic Programming), 2015, pp. 1-363
[25] Shyu, M. L., *et al*., Principal Component-Based Anomaly Detection Scheme, *Studies in Computational Intelligence, 9* (2006), pp. 311-329
[26] Yang, X. F., Research of Text Clustering Based on Improved BIRCH (in Chinese), M. Sc. thesis, Beijing Forestry University, Beijing, China, 2013
[27] Al-Labadi, L., Zarepour, M., Two-Sample Kolmogorov -Smirnov Test Using a Bayesian Nonparametric Approach, *Mathematical Methods of Statistics, 26* (2017), 3, pp. 212-225
[28] Cohen, J., *et al*., *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed. (Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, Third Edition), 2013, pp. 1-704
[29] Hassani, H., Silva, E. S., A Kolmogorov-Smirnov Based Test for Comparing the Predictive Accuracy of Two Sets of Forecasts, *Econometrics, 3* (2015), 3, pp. 590-609
[30] Liu, C., *et al*., Development of a National Cosmic-Ray Dose Monitoring System with Health Canada's Fixed Point Surveillance Network, *Journal of Environmental Radioactivity, 190-191* (2018), pp. 31-38
[31] Zhang, H., Anomaly Detection of Bolt Tightening Quality Based on Big Data Analysis in Car Production, (in Chinese), M. Sc. thesis, Shenyang University of Technology, Shenyang, China, 2019
[32] Zhao, Y., Zhang, X., Environmental Radiation Monitoring System in the United States, *Foreign Nuclear News, 9* (2018), pp. 27-31

Received on August 1, 2023
Accepted on April 4, 2024

Њу ЈУЕНЛУНГ, Луо ЈУЕНФЕЈ, Цунг ФЕИЈУЕН, Ценг ХУЕЈДИ, Љу СИНГАО

ИСТРАЖИВАЊЕ МЕТОДОМ ДЕТЕКЦИЈЕ АНОМАЛИЈА ЕНЕРГЕТСКОГ СПЕКТРА ЗА ДРЖАВНУ КОНТРОЛУ СТАЊА РАДИЈАЦИОНОГ МОНИТОРИНГА ЖИВОТНЕ СРЕДИНЕ НА ОСНОВУ ЛОФ АЛГОРИТМА

Државно контролисане аутоматске станице за праћење радијационог окружења су важни објекти за праћење квалитета и трендова радијационе средине. Међутим, анализа и коришћење података мониторинга са ових станица имају ограничења. Да би се решио овај проблем, овај рад предлаже приступ који користи високодимензионални метод визуелизације података заснован на моделу кластера спектра t-SNE и PCA-BIRCH алгоритма за класификацију историјских података спектра. Након тога, успостављен је модел детекције абнормалних флуктуација спектра околине заснован на фактору локалног ванредног фактора како би се побољшала дискриминација фактора који нису из окружења који узрокују аномалије спектра. Резултати указују на значајне разлике у агрегацији спектра под различитим метеоролошким условима. Алгоритам груписања ефикасно одваја јасне и кишне временске спектре на основу спектралних карактеристика, уз минималну потрошњу времена и смањене сметње од података мешовитог спектра. Користећи карактеристике локалне густине, успостављен је модел детекције аномалија, додељујући абнормални резултат сваком спектру. Компаративна анализа показује побољшану раздвојивост модела детекције аномалија за некласификоване скупове података. Наш алгоритам тачно идентификује флуктуације спектра узроковане факторима који нису у окружењу усред сложене позадине, нудећи вредну техничку подршку за осигурање квалитета животне средине и доношење одлука у случају нуклеарних опасности.

Кључне речи: радијациона средина, визуализација података,PCA-BIRCH алгоритам, анализа кластера