Anomaly Detection of Dam Monitoring Data based on Improved Spectral Clustering

Lixia Ji, Xiao Zhang, Yao Zhao, Zongkun Li*

School of Cyber Science and Engineering, Zhengzhou University, China
jilixia@zzu.edu.cn, 17839227723@163.com, zhaoyao0419@qq.com, lizongkun@zzu.edu.cn

Abstract

In response to the abnormal data mining in dam safety monitoring, and based on the traditional spectral clustering, this paper presents an anomaly detection method based on improved spectral clustering. This method applies a distance and density adaptive similarity measure. The natural eigenvalue is introduced to adaptively select the neighbors of data points, and the similarity is redefined to be combined with the natural k-nearest neighbor. Furthermore, the shared neighbor is introduced to adjust the similarity between the monitoring data samples according to the regional density. Moreover, considering the distribution of dam monitoring data, the initialization of clustering centers is optimized according to both the density and distance feature. This method can prevent the algorithm from local optimum, better adapt to the density of non-convex dataset, reduce the number of iterations, and enhance the efficiencies of clustering and anomaly detection. Taking the dam slab monitoring data as the research object, experimental datasets are formed. Experiments on these datasets further verify that the method of this paper can effectively adapt to discrete distribution datasets and is superior to the classical spectral clustering method in both clustering and anomaly detection.

Keywords: Anomaly detection, Dam, Spectral clustering, Natural eigenvalue, Shared neighbor

1 Introduction

The primary target of dam safety monitoring is to master the operation characteristics of the dam and the changing trend of each monitoring measurement [1]. According to the *Technical Specification for Earth-rockfill Dam Safety Monitoring* (SL 551-2012), the main contents of dam safety monitoring include deformation monitoring, seepage monitoring, internal monitoring, hydraulic monitoring, and environmental monitoring. The concrete face rockfill dam takes the slab as the main anti-seepage system, and thus it is particularly important to monitor the deformation of the slab. Through the analysis of historical data, it is found that slab deformation is closely related to changes in the environment [2]. This means that the abnormal monitoring data can be selected through the statistical analysis of environmental variables such as water pressure, water flow rate, and temperature, which are collected from sensors arranged on the dam slab. These can be further used to study the trend of dam safety and stability. However, the wide distribution of dam monitoring sites and the diversity of environmental variable data lead to multidimensional, discrete, and uneven distributions of the data, which challenges anomaly detection.

Anomaly detection is the classification and recognition of unbalanced data. The goal is to efficiently and accurately identify the suspected abnormal value that deviate from normal distribution [3]. Anomaly detection methods are generally classified into statistical-based methods, nearest neighbor-based methods, and clustering-based methods [4-5]. Statistical-based methods need to make assumptions about the normality of the data, which is only effective when the statistical assumptions meet the actual constraints. In many practical applications, it is very difficult to detect the anomalies of multivariate and unknown distribution data. The most common nearest neighbor-based methods are distance-based methods [6-7], and density-based methods [8-10]. Compared with statistical-based methods, they are computationally effective, but for high-dimensional datasets, especially those with discrete attributes, their performance is significantly reduced [11]. Clustering-based methods find outliers by data grouping, which are efficient and practical for anomaly detection [12-13]. K-means [14-15] is the most common clustering method. Jiang et al. [16] improved K-means clustering-based anomaly detection by optimizing the initial clustering centers, however, this only applies to datasets with a specific data distribution. Clustering with a density measure is another classic approach. In general, traditional clustering methods are good for convex spherical data distribution sample space clustering, but they easily fall into the local optimum, and the clustering effect is not good for multidimensional and non-convex samples such as dam slab monitoring data.

The spectral clustering method provides a new idea for clustering. Compared with traditional methods, it can work on any spatial dataset and can converge to global optimal value [17]. Spectral clustering performs well in practice, but it still has many issues to be further studied. Ayed et al. [18] proposed an improved strategy on the adaptive fuzzy mean, but it has high computational complexity. Liu et al. [19] conducted spectral set clustering by weighted K-means. Meanwhile, Beauchemin et al. [20] used the density estimation method constructing similarity matrix, which improved the accuracy of spectral clustering; however, this method has too many parameters. Yuan et al. [21] proposed a spectral clustering algorithm based on fast search of natural neighborhood, and it can quickly determine natural eigenvalues and improve the clustering accuracy and

*Corresponding Author: Zongkun Li; E-mail: lizongkun@zzu.edu.cn

DOI: 10.53106/160792642022072304010

efficiency on some datasets. There are various methods that can improve the spectral clustering by improving the similarity matrix or the clustering center initialization, but there is no universal similarity measurement method. Due to the multidimensional, discrete, and non-uniform characteristics of dam monitoring data, the existing approaches cannot reflect the similarity between the points well. Furthermore, spectral method is new compared with other clustering methods, and the studies on spectral clustering are focused on clustering, with limited work on anomaly detection.

Therefore, in view of the multidimensional, discrete, and uneven density characteristics of dam monitoring data distributions, we redefine the similarity matrix in spectral clustering by combining global and local density on the distance basis and optimize the selection of initial clustering centers by using the principle of density first. Experiments indicate that our approach is superior to other advanced models in dam anomaly detection. The main contributions of this paper are as follows:

1) An anomaly detection method based on improved spectral clustering is proposed. It takes the dam slab environmental monitoring data as the object to verify its superiority and provides references to dam safety monitoring.

2) On the basis of Euclidean distance, we redefine the data similarity by combining natural neighbors and shared neighbors, and adaptive similarity according to local density.

3) According to the discrete and non-uniform characteristics of dam monitoring data, we optimize the initialization of clustering centers based on the high-density first and maximum distance principles.

This paper is organized as follows. Section 2 discusses related works. Section 3 focuses on the method of anomaly detection by the improved spectral clustering proposed in this paper. Section 4 presents experimental procedures, results, and analysis. Finally conclusions and discussion on future works are given in Section 5.

2 Related work

2.1 Anomaly Detection

Anomaly detection is very important in global research and application fields [22-23]. Intrusion detection in network security [24] is a typical application. Huang et al. [25] proposed an outlier detection framework named CoDetect for financial transaction networks. Boddy et al. [26] proposed a model to detect abnormal access activities in electronic medical record systems. The sensor network [27] is the physical basis of automatic monitoring. Zhang et al. [28] developed an artificial neural network to detect abnormal temperatures of WSNs (Wireless Sensors Networks) in intelligent buildings. Li et al. [29] proposed an improved defense strategy that emphasizes employing KPCA and K-means clustering to defend against data-poisoning attacks in federated-learning systems. Bettencourt et al. [30] identified fault nodes through the space-time structure of sensors and neighbor measurements. Meanwhile, Bhatti et al. [31] developed outlier detection technology for Wi-Fi indoor by analyzing RSSs (Received Signal Strengths). These methods can be regarded as effective solutions in different fields, but outlier detection is always faced with many challenges. First of all, there is no accurate and clear boundary between abnormal data and normal data. Although more and more achievements appeared in anomaly detection, it remains a broad research topic, and there are still many basic problems to be solved in the application domain.

There are many techniques for anomaly detection. Statistical-based methods were the earliest approaches [32]. In recent works, outliers are mostly detected through approximate statistical models of sensor data distribution or time-space series [5, 33-34]. These methods rely on the statistical assumptions made on the data, and it is impractical to establish an effective hypothetical statistical model for multivariate data. Nearest neighbor-based methods lie in the outlier factor measurement. KNN (K-nearest neighbor) is the most fundamental approach. On the basis of the KNN, RNN (reverse nearest neighbor), and SNN (shared nearest neighbor), Wahid et al. [35] proposed an approach using a measure of *k*-nearest neighbor kernel density to estimate data density. In addition, LOF (local outlier factor) [10] and NOF (natural outlier factor) [8] are common calculation methods. The nearest neighbor based technique is simple and intuitive. It only needs to define an appropriate measurement for the given samples. However, in multivariate datasets, the computation of proximity is expensive, and the model is not easy to scale. In general, the measurement techniques between data patterns of nearest neighbor-based method are valuable. Clustering-based methods partition data into groups and implicitly define the outliers as background noise. There are many developments on clustering technology. Distance-based clustering [14, 36] is adequate for finding spherical clusters in small and medium-sized datasets, but the performance is poor for non-convex datasets. In contrast, the density-based clustering method [16, 37] is effective for non-convex datasets, and it is also better for noisy data. Its disadvantage is that the clustering results are highly parameter-dependent. The hierarchical clustering method partitions a set of objects into groups of different levels and has good interpretability but high time complexity. Lastly, the grid-based clustering method has the advantage of fast speed, but the algorithm's efficiency is improved at the cost of accuracy. As a whole, clustering-based methods do not need prior models, but their performance highly depend on the partition ability of clustering algorithms. In addition, deep approaches [38-39] to anomaly detection have recently shown promising results over shallow methods on large and complex datasets.

2.2 Spectral Clustering

As shown in Figure 1, spectral clustering is a new clustering method based on spectral graph theory. Different from the traditional clustering method, it can obtain the optimal result by solving the optimal partition problem of the graph. It is more adaptable to data distribution, can be applied to datasets of any shape, and can converge the global structure to obtain the optimal solution [21]. The most commonly Laplacian matrices types [40] used for spectral clustering are shown in Table 1.

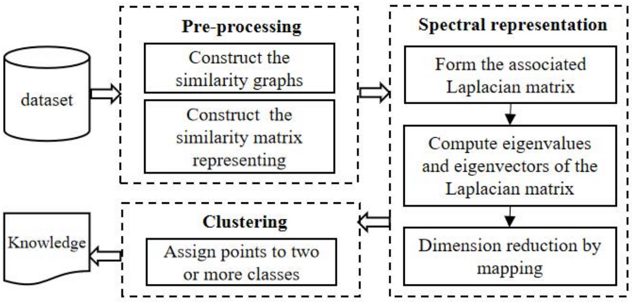
Pre-processing
Construct the similarity graphs
Construct the similarity matrix representing
Spectral representation
Form the associated Laplacian matrix
Compute eigenvalues and eigenvectors of the Laplacian matrix
Dimension reduction by mapping
Clustering
Assign points to two or more classes
dataset
Knowledge

Figure 1. Stages of spectral clustering method

Table 1. Table of Laplacian matrices types

Type	Formula
Unnormalized	$L = D - W$
Symmetric	$L_{Sy} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}WD^{-1/2}$
Asymmetric	$L_{As} = D^{-1}L = I - D^{-1}W$

Note: *L* is the Laplacian matrix, *D* is the degree matrix, and *W* is the adjacency matrix.

Considering the characteristics of the wide distribution of dam monitoring sites, multidimensional data, and uneven density, we propose an anomaly detection method using improved spectral clustering. Specifically, improvements are made in the pre-processing and clustering stages of the spectral clustering method. In the pre-processing stage, the proximity measurement approaches of nearest neighbor-based methods are imported into the spectral clustering-based methods. The natural eigenvalues generated by natural neighbors [8, 21], are combined with the shared neighbors [41], and the similarity matrix is redefined on global and local scales by integrating distance and density to make it more suitable for widely distributed and uneven datasets. In the clustering stage, the clustering centers are selected by the principles of maximum distance and the high-density first, which optimizes the centroid without significantly increasing the number of operations. The proposed method can effectively explore the cluster structure of dam monitoring datasets and can more accurately identify outliers.

3 Method

This section will introduce our improved spectral clustering and the method of dam data anomaly detection.

The environmental sensors monitoring data can be used to analyze the deformation of dam slabs, which may lead to serious water leakage. Engineering practice and research results show that under the action of high water head, peripheral joints of the slab will produce complex three-way displacement, which makes the peripheral joints become leakage channels [2, 34]. For this reason, monitoring data are usually multidimensional data objects that contain attributes such as pressure, temperature, and velocity. They are expressed as: $D = \{x_1, x_2, ..., x_n\}$, $x_i = (x_{i1}, x_{i2}, ..., x_{im})$, where n is the amount of data points, m is the dimension of data x_i, and x_{im} is the m-th attribute of the i-th data point.

As shown in Figure 2, the sample data of the monitoring sites located along the water line at a certain height of the dam usually have similarity, but the abnormal data do not. Anomaly detection in dam slab monitoring can be realized by clustering approach, which can be further abstracted into the problem of graph partition by using spectral method. The dataset of each slab monitoring station at a certain time can be abstracted as as vertices set V in the same spatial graph. According to the similarities of points, the edges set E are weighted, and thus an undirected weighted graph $G = (V, E)$ based on sample similarity is obtained. W_{ij} is defined as the weight between v_i and v_j. The basic rule of weights is that the edge weight between two points far away from each other is lower while the edge weight between two points close together is higher. The adjacency matrix W can be obtained by using the weights of all edges. It is an $n \times n$ matrix, and the j-th value of the i-th row corresponds to the weight W_{ij}.

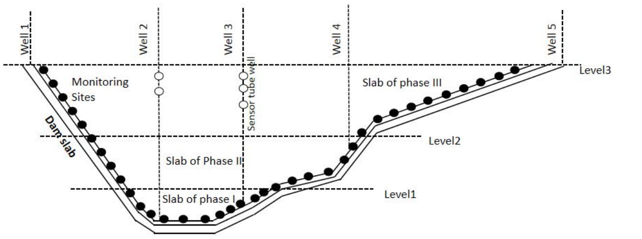

Figure 2. Example layout of monitoring sites for dam slab (The sensors arranged on the dam panel monitor the deformation, as well as the environmental variables such as water pressure, water velocity and temperature.)

The free partition criterion based on graph theory is to maximize the internal similarity of subgraphs and minimize the similarity between subgraphs. Therefore, the quality of clustering is directly affected by the partition criterion.

3.1 Outlier Definition for Dam Monitoring

Referring to the 12 different definitions of outliers given by Ayadi et al. [27], an anomaly or outlier can be described as a data point that manifests itself as a behavior that does not meet expectation or a well-defined abnormal behavior. In the clustering method for dam anomaly detection, let the dataset D contain n samples, each sample has m attributes, and D is divided into K clusters. A partial projection example of the dataset containing outliers in two-dimensional space is shown in Figure 3. There are three types of samples that are usually considered as abnormal data.

Proposition 1: Points that do not pertain to any cluster are outliers (as shown in Figure 1, the points in the rectangular regions are far away from all cluster centers).

Proposition 2: The point far from its nearest cluster is an outlier (as shown in Figure 1, the points in triangle regions are outliers relative to some clusters).

The degree of anomaly can be measured by comparing the distance of the point to its nearest centroid and the average distance within the cluster. For example, Gaddam et al. [36] used the method of combining K-means and ID3 to iteratively calculate the distances, and defined the threshold to filter outliers.

Proposition 3: All of the points in sparse clusters and smaller clusters are outliers (as shown in Figure 1, clusters in the circular regions are abnormally small).

The clustering results can be divided into large clusters and small clusters heuristically, and the anomaly degree can be measured by different calculation methods.

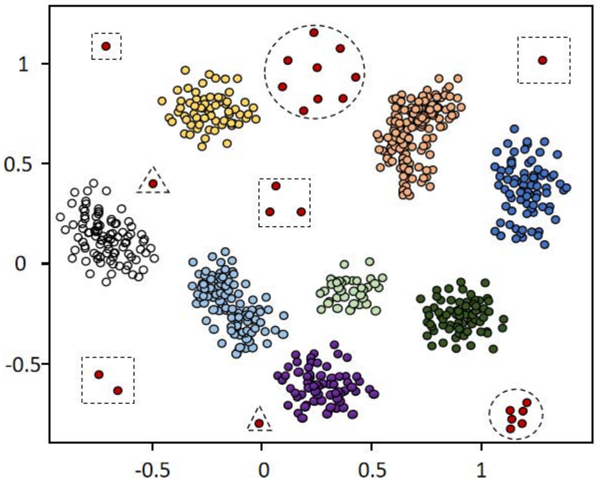

Figure 3. Sketch of outliers in dam anomaly detection
(The rectangular regions highlight the significant outliers far away from all clusters; the triangular regions encompass the outliers relative to a certain cluster; and the circular regions highlight the sparse clusters and smaller clusters, where the points they contain are all outliers.)

3.2 Construction of Similarity Measure

The similarity matrix directly affects the spectral separation. The data similarity in the spectral clustering algorithm is usually defined by a Gaussian kernel function:

$$S_{ij} = \exp(-\tfrac{1}{2\sigma^2} d^2(x_i, x_j)) \tag{1}$$

where S_{ij} is similarity of x_i and x_j, $d(x_i, x_j)$ is the Euclidean distance between x_i and x_j, and σ is the scale parameter, which controls the attenuation speed of the similarity coefficient with the Euclidean distance.

However, σ often needs to be determined by repeated experiments, which increases the calculation amount. Moreover, the Gaussian kernel function only uses the distance information between data points to construct similar functions, which is only suitable for datasets with uniform data distribution.

3.2.1 Adaptive Natural Nearest Neighbor Measurement

The natural neighbor is based on a phenomenon that point in high-density regions has more neighbors while point in low-density regions has fewer neighbors. This method increased the search in the neighborhood, detected the interaction of all data points, and determined the natural eigenvalue adaptively.

Definition 1 (Natural Stable States): Given dataset D, for each data object $x_i \in D$, its k-nearest neighbor is searched, and the variable k is 1, 2, 3, …, n in turn. With the growth of k, when any point in D has at least one inverse nearest neighbor, or when the data points without an inverse nearest neighbor in the dataset remain unchanged, the state is a natural stable state.

Definition 2 (Natural Eigenvalue): If natural neighbor search achieves natural stable, the number of searches is called the natural eigenvalue, which is denoted as $\sup k$.

The similarity measure based on natural neighbors can be defined as follows:

$$S_{ij} = \exp(-\tfrac{1}{\sup k^2} d^2(x_i, x_j)) \tag{2}$$

3.2.2 Adjust Similarity According to Local Density

Further study on the deformation history data of dam slabs showed that if two points belong to the same cluster, they should be located in the same area with relatively high density, and there will be many neighbors overlapping of the two points. The shared neighbors can adjust the similarity between data points according to the local density [41]. Therefore, for the shared performance of data the shared natural neighbor (SNN) is used to represent the neighbor relationship by combining the natural neighbor and the shared neighbor. $Snn(x_i, x_j)$ represents the shared neighbors of x_i and x_j in $\sup k$ natural neighborhood.

Definition 3 (Shared Natural Neighbors): For each point x_i in dataset D, its $\sup k$ nearest neighbor set is $NL(i)$; for two points x_i and x_j in dataset D, their shared natural neighbor set $Snn(x_i, x_j)$ is the intersection of $NL(i)$ and $NL(j)$.

The similarity measure based on shared natural neighbors can be defined as

$$S_{ij} = |Snn(x_i, x_j)| \times \frac{|Snn(x_i, x_j)|}{\sum_{o \in Snn(x_i, x_j)} (d(x_i, o) + d(o, x_j))} \tag{3}$$

where o is a shared natural neighbor of point x_i and point x_j, the similarity is only calculated when x_i and x_j are $\sup k$ nearest neighbors to each other. The right side of the multiplier is the reciprocal of the average distance between two points and their shared neighbors, which represents their local density to some extent.

3.2.3 Comprehensive Measurement based on Sample Features

By obtaining the local density and the shared neighborhood of two points at the same time, the shared natural nearest neighbor similarity can better adapt to various transformed datasets. However, in the case of an incomplete connection, the similarity between non-nearest neighbors is generally recorded as 0, which cannot detect the sparse outliers in dam monitoring datasets. By combining the Gaussian distance and SNN similarity, the similarity of dam monitoring data is defined as follows:

$$S_{ij}=\begin{cases}\dfrac{|Snn(x_i,x_j)|^2}{\sum\limits_{o\in Snn(x_i,x_j)}(d(x_i,o)+d(o,x_j))} & i,j\in Snn(x_i,x_j)\\ d^2(x_i,x_j)\ \text{otherwise}\end{cases}\tag{4}$$

When x_i and x_j are $\sup k$ nearest neighbors to each other, the nearest neighbor density weight is calculated; otherwise, the Euclidean distance is used for similarity.

3.2.4 Construction of Adjacency Matrix

There are three approaches used to construct adjacency matrix W from data similarity: the ε-neighboring approach, K-neighboring approach, and full connection approach. The ε-neighboring approach sets a threshold ε, and defines the adjacency matrix according to the relationship between similarity S_{ij} and ε:

$$W_{ij}=\begin{cases}\varepsilon\ S_{ij}>\varepsilon\\ 0\ S_{ij}<\varepsilon\end{cases}\tag{5}$$

In this case, the weight between two points can only be ε or 0, and the similarity measurement is very rough; therefore, this method is rarely used in practical applications. The K-neighboring approach generally uses the KNN algorithm to search the nearest neighbors of samples; only the W_{ij} values between the point and its k nearest neighbors are greater than 0. At present, the natural neighbor and shared neighbor can also be used to measure the neighbor relationship. The full connection approach defines the weight value between all points as greater than 0. In practical applications, the full connection approach is the most commonly used method to establish the adjacency matrix, and the Gaussian radial kernel function is often employed. [40] is formulated as follows:

$$W_{ij}=S_{ij}=\exp(-\tfrac{1}{2\sigma^2}d^2(x_i,x_j))\tag{6}$$

For dam slab monitoring data points, we use the full connection approach to construct the adjacency matrix but do not simply set the adjacency matrix to be the same as the similarity matrix. Further, we increase the similarity weight between the nearest neighbor samples in the shared natural neighbor search algorithm while retaining Euclidean algorithm to measure similarity. The adjacency matrix is thus formulated as follows:

$$W_{ij}=\begin{cases}\exp(-\dfrac{d^2(x_i,x_j)}{\sigma_i\sigma_j(Snn(x_i,x_j)+1)})\ i\neq j,SNN>5\\ \exp(-\dfrac{d^2(x_i,x_j)}{2\sigma\sigma})\ i\neq j,SNN\leq 5\\ 0\ i=j\end{cases}\tag{7}$$

where σ_i and σ_j are the Euclidean distances from x_i and x_j to their $\sup k$ nearest neighbors, respectively. They can adjust themselves automatically and timely according to the sparse or dense distribution between the two points in the specified neighborhood.

3.3 Improvement of Initialization Method for Clustering Centers

Clustering is an important step in spectral method, and we implement it by improving the K-means. The initialization of centroid in traditional K-means involves randomly selecting k data objects. However, this easily causes randomness in the clustering results, which may lead to low iteration efficiency. Therefore, a more reasonable initialization method of clustering centers will improve the clustering effect.

The basic rule of clustering is the data similarity. Generally, similarity and Euclidean distance are inverse relationship in K-means algorithm. For the task of anomaly detection in dam data, we first try to improve the selection of clustering centers based on distance.

Principle 1: The principle of maximum distance between clustering centers.

The distance between clustering centers should be as large as possible to ensure the uniform distribution of centroids. The K-means algorithm is optimized according to *Principle 1*, and then we get K-means+. This randomly selects the first clustering center, and then continuously selects new clustering centers according to *Principle 1* until k clustering centers are selected.

The initialization of clustering centers based on the maximum distance principle obviously speeds up the iteration speed of the algorithm, but this method cannot rule out the influence of discrete data points. It is possible to select outliers as clustering centers. Since the clustering center is characterized by higher density than other regions, the principle of high-density first is further considered.

Principle 2: The principle of high-density first.

Instead of randomly selecting the first clustering center, we pick out the sample having the highest local density, since that avoids the possibility of selecting outliers to a certain extent. K-means+ is further optimized according to *Principle 2* to yield K-means++. The initialization method of clustering centers in K-means++ is as follows.

Algorithm 1. Clustering center initialization on the basis of maximum distance and high-density priority principles

```
Input: dataset $X=\{x_1,x_2,...,x_n\}$, number of clusters $k$, natural eigenvalue $\sup k$.
Output: clustering centers $c_1,c_2,...,c_k$.
pick out the sample having the most neighbors under $\sup k$ as the first clustering center $c_1$;
for each point $x_i$ in $X$
    calculate the distance $d(x_i,c_1)$ between $x_i$ and $c_1$;
if $d(x_j,c_1)==MAX(d(x_i,c_1))$
    $c_2=x_j$
for $c\leftarrow 3$ to $k$
    for each point $x_i$ in $X$ and $x_i$ that is not a clustering center
```

```
        calculate $d(x_i,c_a)$ and $d(x_i,c_b)$, where $c_a$ and $c_b$ are the arbitrary two clustering centers;
        if $x_j$ makes $(d(x_j,c_a)+d(x_j,c_b))^2$ the largest one
            $c_c = x_j$;
```

3.4 Algorithm

Anomaly detection based on improved spectral clustering method is as follows.

Algorithm 2. Anomaly detection algorithm based on the improved spectral clustering

```
Input: dataset $X = \{x_1,x_2,...,x_n\}$, number of clusters $c$, anomaly threshold value $TVAL$.
Output: tags for each $x_i$.
construct the KNN matrix;
perform the natural neighbors search algorithm in the KNN matrix and obtain the natural eigenvalue $\sup k$ and neighborhood of each point;
calculate the degree matrix $D$;
calculate the adjacency matrix $W$ according to formula (7);
calculate and standardize the Laplacian matrix $L_{Sy} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}WD^{1-/2}$;
calculate eigenvector $v_1,v_2,...,v_c$ corresponding to the first $c$ minimum eigenvalues;
do
    get $c$ clustering centers $c_1,c_2,...,c_c$ according to Algorithm 1;
    use K-means++ to obtain clusters $C_1,C_2,...,C_c$, and the sizes of clusters $s_1,s_2,...,s_c$;
    for $i \leftarrow 1$ to $n$ do
        for $ck \leftarrow 1$ to $c$ do
            if $x_i \in C_{ck}$ and $Dist(x_i,c_{ck}) \geq \frac{1}{s_{ck}}\sum_{j=1}^{j=s_{ck}} Dist(x_j,c_{ck})$
                the abnormal value of $x_i$ $Abn_{x_i}++$;
        if $Abn_{x_i} > 3$
            mark $x_i$ as an outlier;
until no more new outliers are marked;
for $ck \leftarrow 1$ to $c$ do
    if $s_{ck} <= TVAL$
        for each $x_j$ in $C_{ck}$
            mark $x_j$ as an outlier;
```

In the classical spectral clustering algorithm, the Gaussian kernel function is used to construct the similarity matrix. Its sensitivity to scale will lead to unstable clustering results, and the single scale will lead to unrecognized, intertwined clusters on spiral datasets. In this paper, the improved similarity measure can dynamically judge the similarity relationship between two points by whether there are shared natural neighbors and the number of neighbors, which makes up for the defect of single scale scale. In addition, considering the influence of outliers on data partitioning, the maximum distance and high-density first principle are introduced to improve the clustering centers initialization so that the partitioning process contains more accurate prior information.

4 Experiment

The goal of this section is to verify the performance of the proposed method in clustering and anomaly detection. Experiments are carried out on synthetic datasets and real datasets, respectively. Comparative experiments are carried out with the other two clustering-based methods: K-means++ developed from the traditional K-means [14], and classical spectral clustering [42]. In addition, the method is compared with two deep anomaly detection methods on real datasets.

4.1 Datasets

4.1.1 Synthetic Datasets

In this paper, datasets with different characteristics are created for experiments, and the test abnormal data points are added to form three synthetic datasets, namely TwoCircles, FiveClusters, and TwoMoons, for clustering and anomaly detection. Details about the datasets are shown in Table 2.

Table 2. Details of synthetic datasets

Dataset	Number of instances	Number of clusters	Number of outliers
TwoCircles	323	2	9
FiveClusters	2000	5	95
TwoMoons	1525	2	26

4.1.2 Real Datasets

The real datasets are obtained from a dam data center. The data is profile data from 2019 and the data format is .dat. The data file contains the basic information of equipments such as instrument model, data format, and identification number, as well as environmental data such as flow rate, hydrostatic pressure, hydrodynamic pressure, and temperature.

The raw sensor data contains many redundant attributes that have nothing to do with anomaly detection, and thus it is not possible to import the unprocessed data file directly. So, we preprocess raw data first. According to the spatial distribution and seasonal characteristics of the environmental sensors data, 567 observation data near the top area of a specific slab and at a specific time (12:00AM) were selected for the experiment. The data point is recorded as $x_i =$ (hydrostatic pressure, hydrodynamic pressure, flow rate, temperature). The real datasets after pre-processing can be divided into Phase 1 and Phase 2 according to the location of the measuring points; the details are shown in Table 3.

Table 3. Details of real dam slab monitoring datasets

Dataset	Number of dimensions	Number of instances
Phase I	4	95
Phase II	4	472

4.2 Effectiveness of Anomaly Detection

We experiment on three synthetic datasets and real dam monitoring datasets. In order to evaluate the effectiveness of the proposed anomaly detection method, we select accuracy, precision, false negative ratio (FNR), and false positive ratio (FPR) as evaluation indices. They are defined as follows:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad \textbf{(8)}$$

$$\Pr ecision = TP/(TP + FP) \quad \textbf{(9)}$$

$$FNR = FN/(FN + TP) \quad \textbf{(10)}$$

$$FPR = FP/(FP + TN) \quad \textbf{(11)}$$

In anomaly detection, we pay more attention to outliers, and thus we express outliers as positive classes and normal data as negative classes. TP (true positive) denotes an outlier is correctly recognized, FN (false negative) denotes an outlier mistaken as normal data, TN (true negative) denotes normal data correctly identified, and FP (false positive) denotes normal data mistaken as an outlier.

Table 4 shows the experimental results of anomaly detection on synthetic and real datasets.

Table 4. Effectiveness of anomaly detection

Dataset	Accuracy	Precision	FNR	FPR
TwoCircles	**1**	**1**	**0**	**0**
FiveClusters	0.984	0.9024	0.0042	0.2449
TwoMoons	0.9987	0.9231	**0**	0.0769
DamMonitoring	**1**	**1**	**0**	**0**

4.3 Clustering Efficiency and Anomaly Detection Efficiency

Furthermore, we conduct experiments to assess the clustering efficiency and anomaly detection efficiency.

In the experiment, we use four indices: the running time of the clustering algorithm, the RAND index (RI), the adjusted RAND index (ARI), and the normalized mutual information (NMI). Th improved spectral clustering (SC+) algorithm is compared with K-means++ and classical spectral clustering (SC) method. Table 5 compares the running time of each method during clustering. Table 6 compares the clustering RI, ARI, and NMI scores.

Table 5. Clustering execution time comparison, the unit is second (s)

Dataset	K-means++	SC	SC+
TwoCircles	**0.045**	0.185	0.075
FiveClusters	0.39	4.82	**0.21**
TwoMoons	0.21	1.365	**0.195**
DamMonitoring	0.635	0.44	**0.12**

Table 6. Comparison of RI, ARI, and NMI

Dataset	K-means++	SC	SC+
TwoCircles	0.5535	**1**	**1**
	0.4197	**1**	**1**
	0.2796	**1**	**1**
FiveClusters	**1**	0.7455	0.8146
	1	0.5265	0.7228
	1	0.6412	0.7566
TwoMoons	0.6654	0.6854	**1**
	0.2980	0.4876	**1**
	0.2867	0.4215	**1**
DamMonitoring	0.4886	0.7422	**1**
	0.0397	0.6478	**1**
	0.0277	0.7256	**1**

We next evaluate the performance on anomaly detection using four evaluation indices: run time, accuracy, false negative ratio (FNR), and false positive ratio (FPR), and compared with other representative clustering-based methods. The method based on the improved spectral clustering is recorded as SC+AD, the method based on K-means++ is recorded as K++AD, and the method based on classical spectral clustering is recorded as SC-AD. Table 7 shows the experimental results and comparison. Figure 4 shows the original datasets, which contains three synthetic datasets sets and the dam monitoring dataset.

Table 7. Comparison of anomaly detection efficiency

Dataset	Evaluation index	K++-AD	SC-AD	SC+-AD
TwoCircles	run time (s)	**0.075**	0.215	0.126
	Accuracy	0.9907	0.9907	**1**
	FPR	0.0003	0.0003	**0**
	FNR	0.2222	0.2222	**0**
FiveClusters	run time (s)	0.545	5.035	**0.52**
	Accuracy	0.969	**0.991**	0.984
	FPR	0.0016	**0.0005**	0.0042
	FNR	0.602	**0.1735**	0.2449
TwoMoons	run time (s)	0.37	1.525	**0.45**
	Accuracy	0.9889	0.9887	**0.9987**
	FPR	0.002	**0**	**0**
	FNR	0.5385	**0.0769**	**0.0769**
DamMonitoring	run time (s)	0.675	0.62	**0.23**
	Accuracy	0.9841	0.9912	**1**
	FPR	0.0071	0.0089	**0**
	FNR	0.8333	**0**	**0**

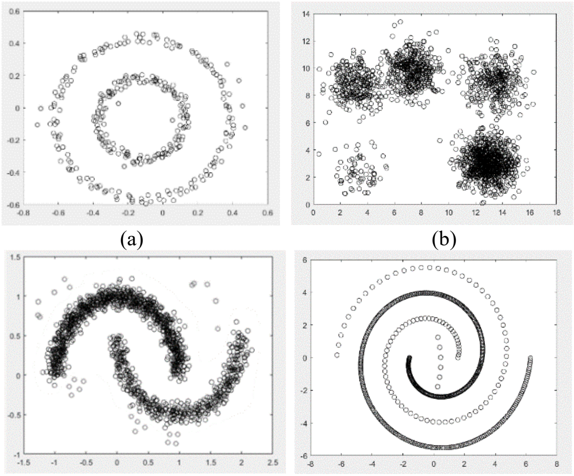
(a) (b) (c) (d)

Figure 4. Original datasets

(a) The original synthetic dataset named TwoCircles
(b) The original synthetic dataset named FiveClusters
(c) The original synthetic dataset named TwoMoons
(d) The dam monitoring dataset named DamMonitoring

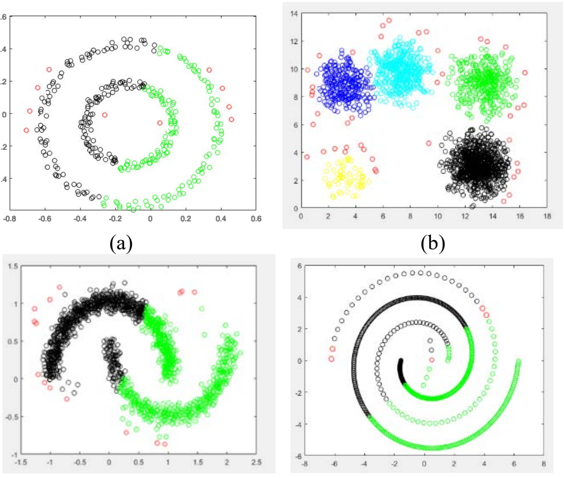
(a) (b) (c) (d)

Figure 5. Results based on K-means++

(a) Results on the dataset TwoCircles
(b) Results on the dataset FiveClusters
(c) Results on the dataset TwoMoons
(d) Results on the dataset DamMonitoring

The clustering and anomaly detection results based on K-means++ are shown in Figure 5. K-means++ improves the centroid initialization method according to the maximum distance and high-density priority principles. The algorithm works well on simple datasets such as FiveClusters, as it improves the efficiency of clustering centers iteration, still has the problem of a poor effect on non-convex datasets. In the experiment of the convex sample dataset FiveClusters, its clustering score is higher than others, but its clustering effects in the other, non-convex datasets, are poor (as shown in Table 4). In addition, it can be seen from Figure 5, the manifold spiral datasets cannot be recognized. This method simply measures the data similarity by distance, and the false negative is too high.

Figure 6 shows the results of clustering and anomaly detection based on classical spectral clustering. Since the scale parameters need to be selected by many experiments, the run time of this method is generally high. The clustering effect of the spectral method has good performance on sparse datasets and can also be applied to some non-convex datasets, but it is still insufficient for the spiral winding datasets such as (*c*) and (*d*) shown in Figure 6. In regards to outlier recognition, this approach is evidently better than K-means++. In addition, because the similarity measure of classical spectral clustering is still completely based on distance and cannot consider the density, the efficiency of classical spectral clustering on non-uniform datasets declines sharply.

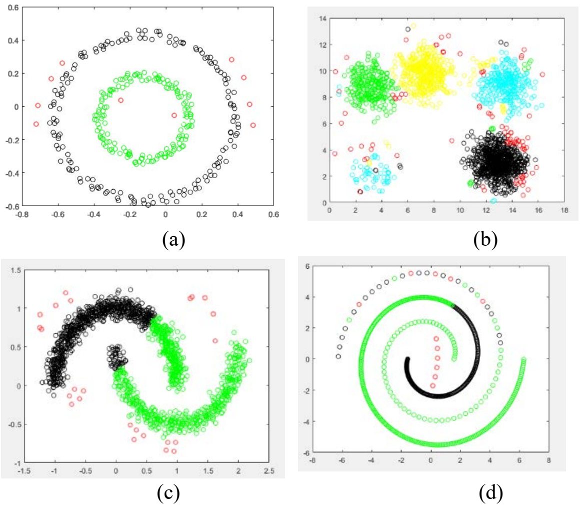

Figure 6. Results based on classical spectral clustering
(a) Results on the dataset TwoCircles
(b) Results on the dataset FiveClusters
(c) Results on the dataset TwoMoons
(d) Results on the dataset DamMonitoring

Figure 7 shows the results of clustering and anomaly detection by improved spectral clustering. Adaptively setting the neighborhood greatly reduces the cost of the neighbor search. It can be seen from table 5 and table 7 that the improved method has obvious advantages in both clustering run time and anomaly detection execution time. Because it takes distance and local density into account when measuring similarity, the clustering results reflect the datasets structure more accurately, and the clustering performance and anomaly detection performance are significantly better than classical spectral clustering. Especially on the real dataset, the effect is obviously better than the classical N-cut algorithm, which only takes the distance as the unique standard and constructs the similarity matrix by the Gaussian function.

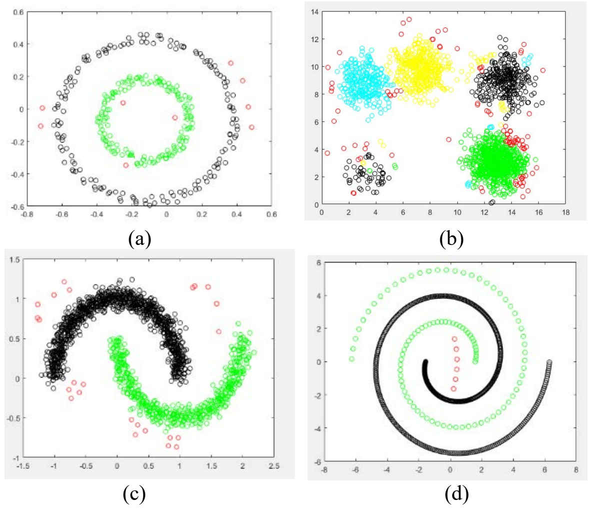

Figure 7. Results based on improved spectral clustering
(a) Results on the dataset TwoCircles
(b) Results on the dataset FiveClusters
(c) Results on the dataset TwoMoons
(d) Results on the dataset DamMonitoring

We further test the anomaly detection efficiency and consider to compare with deep learning methods DAGMM [38] and DUAD [39]. At the same time, we expand datasets to include more observation data from May 2019 to April 2021. Detailed information about the datasets is shown in Table 8. Table 9 shows the experimental results and comparison.

Table 8. Details of more real dam slab monitoring datasets

Dataset	Number of dimensions	Number of instances	Number of outliers
Dam 1	4	43,282	137
Dam 2	8	6,183	113
Dam 3	23	796	21

Table 9. Comparison of anomaly detection efficiency

Dataset	Evaluation index	DAGMM	DUAD	SC+-AD
Dam 1	Precision	0.9010	0.9226	**0.9227**
	FPR	0.00033	**0.000257**	0.000259
	FNR	0.0511	0.0336	**0.0248**
Dam 2	Precision	0.8814	0.8914	**0.8916**
	FPR	0.00242	**0.00221**	0.00222
	FNR	0.0336	0.0257	**0.0207**
Dam 3	Precision	0.7568	0.8042	**0.8083**
	FPR	0.00684	**0.00606**	0.00619
	FNR	**0.0667**	0.081	0.0762

From the experimental results, compared with deep anomaly detection methods, SC+AD still performs well, especially in the FNR index, which is exactly what anomaly detection needs. With the increase of dimensions, the performance of our method decreases, but there is a small gap with the best results.

In short, experiments show that the clustering and anomaly detection efficiency of the proposed method are greatly improved. It has better efficiency and stability in clustering, and can reflect the characteristics of datasets more efficiently and accurately. The significant advantages of this clustering method are its speed and accuracy on complex non-uniform datasets. In the anomaly detection further, the efficiency is significantly improved. Moreover, its outliers identification ratio is higher, and false detection ratio is lower, especially on the real dataset.

5 Conclusion

Following research on the spectral clustering algorithm and anomaly detection, an improved method for anomaly detection in dam monitoring data is proposed. The natural eigenvalue is introduced to determine the neighborhood adaptively, and the similarity calculation method is redefined by using the shared neighborhood and distance information between data. The similarity can effectively describe the actual distribution and internal relationship between data. Then, anomaly detection is carried out according to anomaly criteria and assumptions, which can fully reflect the overall situation of the data. Experimental results show that this method has stronger adaptive ability than other advanced methods and has higher clustering and anomaly detection efficiency.

Data is growing exponentially in big data era. In the field of dam anomaly detection, processing higher dimensional and large datasets will be the focus of our future studies. In addition, we will consider further improvement of this method to make it applicable to more domains and help accomplish more tasks, such as domain entity classification [43].

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant No. 51679222, in part by the Major Science and Technology Project in Henan Province under Grant No. 201300210500, and in part by the Zhengzhou Major Science and Technology Innovation Project under Grant No. 2020CXZX0053.

References

[1] C. Liao, D. Cai, M. Li, J. Zhang, Y. Wang, Application of fuzzy clustering method in dam monitoring data analysis, *Yangtze River*, Vol. 46, No. 13, pp. 86-89, July, 2015.

[2] Z. Li, J. Zhou, J. Zheng, Multi-pole fuzzy pattern recognition method for behavior evaluation of earth-rock dam, *Journal of Hydraulic Engineering*, Vol. 34, No. 9, pp. 83-87, September, 2003.

[3] A. Taha, A. S. Hadi, Anomaly detection methods for categorical data: A review, *ACM Computing Surveys (CSUR)*, Vol. 52, No. 2, pp. 1-35, March, 2020.

[4] M. Z. Zaheer, A. Mahmood, M. Astrid, S. I. Lee, CLAWS: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection, *European Conference on Computer Vision*, Glasgow, UK, 2020, pp. 358-376.

[5] H. K. Verma, K. V. Samparthi, Outlier detection of data in wireless sensor networks using kernel density estimation, *International Journal of Computer Applications*, Vol. 5, No.7, pp. 28-32, August, 2010.

[6] E. M. Knorr, R. T. Ng, Algorithms for Mining distance-based outliers in large datasets, *VLDB*'98 *International Conference on Very Large Databases*, New York, USA, 1998, pp. 392-403.

[7] J. Ha, S. Seok, J. S. Lee, A precise ranking method for outlier detection, *Information Sciences*, Vol. 324, pp. 88-107, December, 2015.

[8] J. Huang, Q. Zhu, L. Yang, J. Feng, A non-parameter outlier detection algorithm based on natural neighbor, *Knowledge-Based Systems*, Vol. 92, pp. 71-77, January, 2016.

[9] M. Bai, X. Wang, J. Xin, G. Wang, An efficient algorithm for distributed density-based outlier detection on big data, *Neurocomputing*, Vol. 181, pp. 19-28, March, 2016.

[10] E. Schubert, A. Zimek, H. P. Kriegel, Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection, *Data mining and knowledge discovery*, Vol. 28, No. 1, pp. 190-237, January, 2014.

[11] H. Wang, M. J. Bah, M. Hammad, Progress in outlier detection techniques: A survey, *IEEE Access*, Vol. 7, pp. 107964-108000, August, 2019.

[12] P. Berkhin, *A survey of clustering data mining techniques*, in: J. Kogan, C. Nicholas, M. Teboulle (Eds.), Springer, Berlin, Heidelberg, 2006, pp. 25-71.

[13] M. Li, Y. Deng, A Machine learning-based building operational pattern identification, *International Journal of Performability Engineering*, Vol. 16, No. 11, pp. 1835-1844, November, 2020.

[14] A. K. Jain, Data clustering: 50 years beyond K-means, *Pattern Recognition Letters*, Vol. 31, No. 8, pp. 651-666, June, 2010.

[15] C. Zhang, H. Cui, Y. Wang, T. Zhao, Y. Zhou, LDKM: an improved K-means algorithm with linear fitting density peak, *International Journal of Performability Engineering*, Vol. 16, No. 3, pp. 454-461, March, 2020.

[16] H. Jiang, Y. Wu, X. Wang, H. J. Wang, Study on Ocean Data Anomaly Detection Algorithm Based on Improved K-means Clustering, *Computer Science*, Vol. 46, No. 7, pp. 211-216, July, 2019.

[17] S. Ding, H. Jia, L. Zhang, F. Jin, Research of semi-supervised spectral clustering algorithm based on pairwise constraints, *Neural Computing and Applications*, Vol. 24, No. 1, pp. 211-219, January, 2014.

[18] A. B. Ayed, M. B. Halima, A. M. Alimi, Adaptive fuzzy exponent cluster ensemble system based feature selection and spectral clustering, *IEEE International Conference on Fuzzy Systems*, Naples, Italy, 2017, pp. 1-6.

[19] H. Liu, J. Wu, T. Liu, D. Tao, Y. Fu, Spectral ensemble clustering via weighted k-means: Theoretical and practical evidence, *IEEE transactions on knowledge and data engineering*, Vol. 29, No. 5, pp. 1129-1143, May, 2017.

[20] M. Beauchemin, A density-based similarity matrix construction for spectral clustering, *Neurocomputing*, Vol. 151, pp. 835-844, March, 2015.

[21] M. Yuan, Q. Zhu, Spectral Clustering Algorithm Based on Fast Search of Natural Neighbors, *IEEE Access*, Vol. 8, pp. 67277-67288, April, 2020.

[22] Z. K. Alkhateeb, A. T. Maolood, Machine Learning-Based Detection of Credit Card Fraud: A Comparative Study, *American Journal of Engineering and Applied Sciences*, Vol. 12, No. 4, pp. 535-542, 2019.

[23] Y. Djenouri, A. Belhadi, J. C. W. Lin, D. Djenouri, A. Cano, A survey on urban traffic anomalies detection algorithms, *IEEE Access*, Vol. 7, pp. 12192-12205, January, 2019.

[24] G. Singh, F. Masseglia, C. Fiot, A. Marascu, P. Poncelet, Mining common outliers for intrusion detection, *Advances in Knowledge Discovery and Management*, Strasbourg, France, 2009, pp. 217-234.

[25] D. Huang, D. Mu, L. Yang, X. Cai, CoDetect: financial fraud detection with anomaly feature detection, *IEEE Access*, Vol. 6, pp. 19161-19174, March, 2018.

[26] A. J. Boddy, W. Hurst, M. Mackay, A. el Rhalibi, Density-based outlier detection for safeguarding electronic patient record systems, *IEEE Access*, Vol. 7, pp. 40285-40294, March, 2019.

[27] A. Ayadi, O. Ghorbel, A. M. Obeid, M. Abid, Outlier detection approaches for wireless sensor networks: A survey, *Computer Networks*, Vol. 129, pp. 319-333, December, 2017.

[28] K. Zhang, K. Yang, S. Li, D. Jing, H. B. Chen, ANN-based outlier detection for wireless sensor networks in smart buildings, *IEEE Access*, Vol. 7, pp. 95987-95997, July, 2019.

[29] D. Li, W. E. Wong, W. Wang, Y. Yao, M. Chau, Detection and mitigation of label-flipping attacks in federated learning systems with KPCA and K-means, *2021 8th International Conference on Dependable Systems and Their Applications (DSA)*, Yinchuan, China, 2021, pp. 551-559.

[30] L. M. Bettencourt, A. A. Hagberg, L. B. Larkey, Separating the wheat from the chaff: Practical anomaly detection schemes in ecological applications of distributed sensor networks, *International Conference on Distributed Computing in Sensor Systems*, Santa Fe, NM, USA, 2007, pp. 223-239.

[31] M. A. Bhatti, R. Riaz, S. S. Rizvi, S. Shokat, F. Riaz, S. J. Kwon, Outlier detection in indoor localization and Internet of Things (IoT) using machine learning, *Journal of Communications and Networks*, Vol. 22, No. 3, pp. 236-243, June, 2020.

[32] R. Zhang, F. Nie, M. Guo, X. Wei, X. Li, Joint learning of fuzzy k-means and nonnegative spectral clustering with side information, *IEEE Transactions on Image Processing*, Vol. 28, No. 5, pp. 2152-2162, May, 2019.

[33] Y. Zhang, N. A. S. Hamm, N. Meratnia, A. Stein, M. van de Voort, P. J. M. Havinga, Statistics-based outlier detection for wireless sensor networks, *International Journal of Geographical Information Science*, Vol. 26, No. 8, pp. 1373-1392, February, 2012.

[34] H. Zhang, Z. Fan, M. Chen, Application of Isolated Forest in Abnormal Identification of Dam Monitoring Data, *Yellow River*, Vol. 42, No. 8, pp. 154-157+168, September, 2020.

[35] A. Wahid, A. C. S. Rao, An Outlier Detection Algorithm based on KNN-kernel Density Estimation, *International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, 2020, pp. 1-8.

[36] S. R. Gaddam, V. V. Phoha, K. S. Balagani, K-Means+ ID3: A novel method for supervised anomaly detection by cascading K-Means clustering and ID3 decision tree learning methods, *IEEE transactions on knowledge and data engineering*, Vol. 19, No. 3, pp. 345-354, March, 2007.

[37] M. Ester, H. P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, USA, 1996, pp. 226-231.

[38] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, H. Chen, Deep autoencoding gaussian mixture model for unsupervised anomaly detection, *International Conference on Learning Representations*, Vancouver, Canada, 2018, pp. 1-19.

[39] T. Li, Z. Wang, S. Liu, W. Lin, Deep unsupervised anomaly detection, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, virtual, 2021, pp. 3636-3645.

[40] D. Hamad, P. Biela, Introduction to spectral clustering, *International Conference on Information and Communication Technologies: From Theory to Applications*, Damascus, Syria, 2008, pp. 1-6.

[41] R. Liu, H. Wang, X. Yu, Shared-nearest-neighbor-based clustering by fast search and find of density peaks, *Information Sciences*, Vol. 450, pp. 200-226, June, 2018.

[42] U. V. Luxburg, A tutorial on spectral clustering, *Statistics and Computing*, Vol. 17, No. 4, pp. 395-416, December, 2007.

[43] H. Zhang, Y. Guo, T. Li, Multifeature named entity recognition in information security based on adversarial learning, *Security and Communication Networks*, Vol. 2019, pp. 1-9, February, 2019.

Biographies

Lixia Ji, born in 1979, she is currently an Associate Professor of computer science and applications with Zhengzhou University, and the Deputy Director of Zhengzhou Key Laboratory of Blockchain and Data Intelligence. Her current research interests include knowledge mining, multimodal learning and data intelligence.

Xiao Zhang received the bachelor's degree from Henan University in 2018, the M.ENG degree from School of Cyber Science and Engineering, Zhengzhou University in 2021. His research interest include data mining and data intelligence.

Yao Zhao received the bachelor's degree from Henan University of Economics and Law in 2019. Now, he is currently a graduate student at School of Cyber Science and Engineering, Zhengzhou University. His research interests include database system and data intelligence.

Zongkun Li received the Ph.D. degree in Hydraulic Structure Engineering from Dalian University of Technology, China in 2003. Currently, he is a professor at School of College of Hydraulic Science and Engineering, Zhengzhou University. His research interests include risk evaluation and control of hydraulic projects and intelligent water conservancy.